aprendiendo estadística con jamovi

DANIELLE J NAVARRO
DAVID R FOXCROFT
ELENA GERVILLA
FEDERICO LEGUIZAMO

un tutorial para estudiantes de psicología y las ciencias de la salud y el comportamiento

aprendiendo estadística con jamovi

Danielle J. Navarro, David R. Foxcroft, Elena Gervilla, Federico Leguizamo

Table of contents

Este libro de texto cubre el contenido de una clase de introducción a la estadística, tal como se enseña típicamente a estudiantes de pregrado en psicología, salud o ciencias sociales. El libro cubre cómo comenzar en jamovi y brinda una introducción a la manipulación de datos. Desde una perspectiva estadística, el libro analiza primero la estadística descriptiva y los gráficos, seguidos de capítulos sobre teoría de probabilidad, muestreo y estimación, y prueba de hipótesis nula. Después de presentar la teoría, el libro cubre el análisis de tablas de contingencia, correlación, pruebas *t*, regresión, ANOVA y análisis factorial. Las estadísticas bayesianas se abordan al final del libro.

Citation: Navarro DJ, Foxcroft DR, Gervilla E, Leguizamo F (2022). aprendiendo estadística con jamovi: un tutorial para estudiantes de psicología y las ciencias de la salud y el comportamiento. (Version 0.75).

Prefacio

Historial y Licencia

Este libro es una adaptación de DJ Navarro (2018). Aprendiendo estadísticas con R: un tutorial para estudiantes de psicología y otros principiantes. (Versión 0.6). https://learningstatisticswithr.com/.

[Foto de portada de Edward Howell en Unsplash]

Prefacio a la versión 0.75

En esta versión hemos actualizado las figuras, imágenes y texto para mantener compatibilidad con las últimas versiones de jamovi (2.2); muchas gracias a pedro Fisk por su ayuda con esto. También se modificaron y corrigieron algunos secciones donde los lectores han sugerido mejoras. Esto tiene incluía principalmente la corrección de errores tipográficos, pero también en algunos lugares la corrección conceptual detalle, por ejemplo hemos actualizado la información sobre la curtosis para reflejan que no se trata realmente de la "puntualidad" de la distribución y, en cambio, la curtosis se trata de si las distribuciones de datos tienen colas delgadas o gruesas. Gracias a todos los lectores que hicieron sugerencias, ya sea a través de contactarme por correo electrónico o plantear un problema en github.

David Foxcroft 9 de febrero de 2022

Prefacio a la versión 0.70

Esta actualización de la versión 0.65 introduce algunos análisis nuevos. En el ANOVA capítulos hemos agregado secciones sobre medidas repetidas ANOVA y análisis de covarianza (ANCOVA). En un nuevo capítulo hemos introducido Factor Análisis y técnicas relacionadas. Esperemos que el estilo de este nuevo El material es consistente con el resto del libro, aunque con ojos de águila. los lectores pueden notar un poco más de

énfasis en lo conceptual y lo práctico. explicaciones y un poco menos de álgebra. No estoy seguro de que esto sea algo bueno, y podría agregar el álgebra un poco más tarde. Pero refleja tanto mi enfoque para entender y enseñar estadística, y también algunos retroalimentación que he recibido de los estudiantes en un curso que enseño. En línea con esto, también he repasado el resto del libro y he tratado de separe algo del álgebra colocándolo en una caja o marco. Es no es que estas cosas no sean importantes o útiles, pero para algunos estudiantes es posible que deseen omitirlo y, por lo tanto, el boxeo de estas partes debería ayudar a algunos lectores.

Con esta versión agradezco mucho los comentarios y la retroalimentación recibida. de mis estudiantes y colegas, en particular Wakefield Morys-Carter, y también a numerosas personas en todo el mundo que han enviado en pequeños sugerencias y correcciones - muy apreciadas, ¡y sigan viniendo! Una característica nueva bastante interesante es que los archivos de datos de ejemplo para el libro ahora se puede cargar en jamovi como un módulo adicional, gracias a Jonathon Me encanta ayudar con eso.

David Foxcroft 1 de febrero de 2019

Prefacio a la versión 0.65

En esta adaptación del excelente 'Aprender estadística con R', de Danielle Navarro, hemos reemplazado el software estadístico utilizado para el análisis y ejemplos con jamovi. Aunque R es un potente estadístico lenguaje de programación, no es la primera opción para todos los instructores y estudiante al inicio de su aprendizaje estadístico. Alguno los instructores y los estudiantes tienden a preferir el estilo de apuntar y hacer clic de software, y ahí es donde entra jamovi. jamovi es un software que tiene como objetivo para simplificar dos aspectos del uso de R. Ofrece un apuntar y hacer clic interfaz gráfica de usuario (GUI), y también proporciona funciones que combinar las capacidades de muchos otros, trayendo un SPSS- o más Método de programación similar a SAS para R. Es importante destacar que jamovi siempre será libre y abierto - ese es uno de sus valores centrales - porque jamovi está hecho por la comunidad científica, para la comunidad científica.

Con esta versión estoy muy agradecido por la ayuda de otros que han leyó los borradores y proporcionó excelentes sugerencias y correcciones, particularmente el Dr. David Emery y Kirsty Walter.

David Foxcroft 1 de julio de 2018

Prefacio a la versión 0.6

El libro no ha cambiado mucho desde 2015 cuando lancé la versión 0.5. probablemente sea justo decir que he cambiado más de lo que ha cambiado. Me mudé de Adelaide a Sydney en 2016 y mi perfil docente en la UNSW es diferente a lo que era en Adelaide, y realmente no he tenido la oportunidad a trabajar en ello desde que llegué aquí! Es un poco extraño mirar hacia atrás esto en realidad Unos comentarios rápidos…

- Extrañamente, el libro constantemente me malinterpreta, pero supongo que tengo solo yo tengo la culpa de eso :-) Ahora hay una breve nota al pie en la página 12 que menciona este tema; en la vida real he estado trabajando a través de un proceso de afirmación de género durante los últimos dos años y en su mayoría van

por ella / sus pronombres. Sin embargo, soy tan perezoso como siempre. fue así que no me he molestado en actualizar el texto en el libro.
- Para la versión 0.6 no he cambiado mucho, he hecho algunos cambios menores cuando las personas han señalado errores tipográficos u otros errores. En particular vale la pena señalar el problema asociado con la función etaSquared en el paquete lsr (que en realidad ya no se mantiene) en Sección 14.4. La función funciona bien para los ejemplos simples en el libro, pero definitivamente hay errores allí que no he encontrado tiempo para comprobar! Así que por favor ten cuidado con eso.
- ¡El cambio más grande es realmente la licencia! Lo he lanzado bajo un Licencia Creative Commons (CC BY-SA 4.0, en concreto), y colocado todos los archivos fuente al repositorio de GitHub asociado, si alguien quiere adaptarlo.

Tal vez a alguien le gustaría escribir una versión que haga uso de la tidyverse… Escuché que eso se ha vuelto bastante importante para R en estos días :-)

Mejor, *Danielle Navarro*

Prefacio a la versión 0.5

Otro año, otra actualización. Esta vez, la actualización se ha centrado casi enteramente en las secciones teóricas del libro. Capítulos 9, 10 y 11 han sido reescritos, con suerte para mejor. En la misma linea, El capítulo 17 es completamente nuevo y se centra en las estadísticas bayesianas. pienso los cambios han mejorado mucho el libro. siempre me he sentido incómodo por el hecho de que todas las estadísticas inferenciales en el libro se presentan desde una perspectiva ortodoxa, aunque casi Siempre presento análisis de datos bayesianos en mi propio trabajo. Ahora que tengo me las arreglé para incluir métodos bayesianos en el libro en alguna parte, estoy empezando a sentirme mejor con el libro en su conjunto. queria conseguir unos cuantos otras cosas hechas en esta actualización, pero como de costumbre, me encuentro enseñando plazos, ¡así que la actualización tiene que salir como está!

Danielle Navarro 16 de febrero de 2015

Prefacio a la versión 0.4

Ha pasado un año desde que escribí el último prefacio. el libro ha cambiado en algunas formas importantes: los capítulos 3 y 4 hacen un mejor trabajo al documentar algunas de las características de ahorro de tiempo de Rstudio, los Capítulos 12 y 13 ahora hacen uso de nuevas funciones en el paquete lsr para ejecutar pruebas de chi-cuadrado y pruebas t, y la discusión de las correlaciones se ha adaptado para referirse a las nuevas funciones en el paquete lsr. La copia blanda de 0.4 ahora tiene mejores referencias internas (es decir, hipervínculos reales entre secciones), aunque eso se introdujo en 0.3.1. Hay algunos ajustes aquí y allí, y muchas correcciones de errores tipográficos (gracias a todos los que señalaron errores tipográficos!), pero en general 0.4 no es muy diferente de 0.3.

Desearía haber tenido más tiempo en los últimos 12 meses para agregar más contenido. La ausencia de cualquier discusión de ANOVA de medidas repetidas y mixto los modelos en general realmente me molestan. Mi excusa para esta falta de El progreso es que mi

segundo hijo nació a principios de 2013, por lo que Pasé la mayor parte del año pasado tratando de mantener mi cabeza fuera del agua. Como un En consecuencia, los proyectos paralelos no remunerados como este libro quedaron relegados a favor de las cosas que realmente pagan mi salario! Las cosas están un poco más tranquilas ahora, así que, con un poco de suerte, la versión 0.5 será un gran paso adelante.

Una cosa que me ha sorprendido es la cantidad de descargas del libro. obtiene. Finalmente obtuve información básica de seguimiento del sitio web un hace un par de meses, y (después de excluir a los robots obvios) el libro tiene estado promediando alrededor de 90 descargas por día. Eso es alentador: hay al menos algunas personas que encuentran útil el libro!

Danielle Navarro 4 de febrero de 2014

Prefacio a la versión 0.3

Hay una parte de mí que realmente no quiere publicar este libro. Es sin terminar.

Y cuando digo eso, lo digo en serio. La referencia es irregular en el mejor de los casos, el los resúmenes de los capítulos son solo listas de títulos de secciones, no hay índice, no hay ejercicios para el lector, la organización es subóptima, y la cobertura de los temas no es lo suficientemente completa para mi gusto. Además, hay secciones con contenido que no me gusta. con, figuras que realmente necesitan ser redibujadas, y casi no he tenido tiempo para buscar inconsistencias, errores tipográficos o errores. En otras palabras, este libro no está terminado. Si no tuviera una fecha límite de enseñanza que se avecina y un bebé que nacerá en unas pocas semanas, realmente no estaría haciendo esto disponible en absoluto.

Lo que esto significa es que si eres un académico que busca enseñar materiales, un Ph.D. estudiante que busca aprender R, o simplemente un miembro de la público en general interesado en las estadísticas, le aconsejo que sea precavido. Lo que está viendo es un primer borrador, y puede que no sirva tus propósitos Si viviéramos en los días en que la publicación era caro y no había Internet, nunca consideraría publicar un libro de esta forma. La idea de que alguien pague $80 por esto (que es lo que un editor comercial me dijo que vendería al por menor para cuando se ofrecieron a distribuirlo) me hace sentir más que un poco incómodo. Sin embargo, es el siglo XXI, así que puedo publicar el pdf en mi sitio web de forma gratuita, y puedo distribuir copias impresas a través de un servicio de impresión bajo demanda por menos de la mitad de lo que cuesta un editor de libros de texto cobraría. ¡Y así mi culpa se alivia, y estoy dispuesto a compartir! Con eso en mente, puede obtener copias impresas gratuitas y copias impresas baratas. en línea, desde las siguientes páginas web:

Copia electrónica: www.compcogscisydney.com/learning-statistics-with-r.html Copia impresa: www.lulu.com/content/13570633 [***Ed: estos enlaces están obsoletos, intente esto en su lugar: estadísticasdeaprendizajeconr.com***]

Aun así, la advertencia sigue en pie: lo que estás viendo es Versión 0.3 de un trabajo en progreso. Si llega a la versión 1.0 y cuando llegue, estaría dispuesto a respaldar el trabajo y decir, sí, este es un libro de texto que alentaría a otras personas a usar. En ese momento, probablemente empezaré azotar descaradamente la cosa en Internet y, en general, actuar como una herramienta. Pero hasta que llegue ese día, me gustaría que quedara claro que estoy realmente ambivalente sobre el trabajo tal como está.

Habiendo dicho todo lo anterior, hay un grupo de personas que puedo Respaldo con entusiasmo este libro a: los estudiantes de psicología que toman nuestras clases de métodos de investigación de pregrado (DRIP y DRIP:A) en 2013. Para ti, este libro es ideal, porque fue escrito para acompañar tu conferencias de estadísticas. Si surge un problema debido a una deficiencia de estas notas, Puedo adaptar y adaptaré el contenido sobre la marcha para solucionar ese problema. Efectivamente, tienes un libro de texto escrito específicamente para tu clases, distribuidas de forma gratuita (copia electrónica) o a precios cercanos al costo (copia impresa). Mejor aún, las notas han sido probadas: Versión 0.1 de estas notas se usaron en la clase de 2011, la versión 0.2 se usó en la clase de 2012 clase, y ahora está viendo la nueva y mejorada Versión 0.3. estoy No digo que estas notas sean una genialidad chapada en titanio en un palo. aunque si quisiera decirlo en los formularios de evaluación de los estudiantes, entonces eres totalmente bienvenido a hacerlo, porque no lo son. pero estoy diciendo que se han probado en años anteriores y parecen funcionar bien. Además, hay un grupo de nosotros para solucionar cualquier problema surgir, y puede garantizar que al menos uno de sus disertantes ha ¡Lee todo de cabo a rabo!

De acuerdo, con todo eso fuera del camino, debería decir algo sobre lo que el libro pretende ser. En esencia, es una estadística introductoria libro de texto dirigido principalmente a estudiantes de psicología. Como tal, cubre los temas estándar que esperaría de un libro de este tipo: diseño del estudio, estadística descriptiva, la teoría de la prueba de hipótesis, pruebas t, 2 pruebas, ANOVA y regresión. Sin embargo, también hay varios capítulos. dedicado al paquete estadístico R, incluido un capítulo sobre datos manipulación y otra sobre guiones y programación. Además, cuando miras el contenido presentado en el libro, notarás muchas temas que tradicionalmente se barren debajo de la alfombra cuando se enseña estadística a los estudiantes de psicología. La división bayesiana/frecuentista es discutido abiertamente en el capítulo de probabilidad, y el desacuerdo entre Neyman y Fisher sobre la prueba de hipótesis hace su aparición. los se discute la diferencia entre probabilidad y densidad. Un detallado tratamiento de sumas de cuadrados Tipo I, II y III para factorial desbalanceado Se proporciona ANOVA. Y si echas un vistazo en el Epílogo, debería ser Claro que mi intención es agregar mucho más contenido avanzado.

Mis razones para seguir este enfoque son bastante simples: los estudiantes pueden manejarlo, e incluso parecen disfrutarlo. En los últimos años Me ha sorprendido gratamente la poca dificultad que he tenido para hacer que los estudiantes de psicología aprendan R. Ciertamente no es fácil para ellos, y he descubierto que necesito ser un poco caritativo al establecer marcando estándares, pero eventualmente lo logran. Del mismo modo, ellos no parece tener muchos problemas para tolerar la ambigüedad y la complejidad en la presentación de ideas estadísticas, siempre que se les asegure que los estándares de evaluación se establecerán de manera apropiada para ellos. Entonces, si los estudiantes pueden manejarlo, ¿por qué no enseñarlo? los las ganancias potenciales son bastante tentadoras. Si aprenden R, los estudiantes obtienen acceso a CRAN, que es quizás el más grande y completo library(de herramientas estadísticas existentes. Y si aprenden sobre teoría de la probabilidad en detalle, es más fácil para ellos cambiar de prueba de hipótesis nula ortodoxa a métodos bayesianos si así lo desean. Mejor aún, aprenden habilidades de análisis de datos que pueden llevar a un empleador sin depender de software costoso y propietario.

Lamentablemente, este libro no es la panacea que hace posible todo esto. Es un trabajo en progreso, y tal vez cuando esté terminado será un Herramienta útil. Uno entre mu-

chos, diría yo. Hay una serie de otros libros que tratan de proporcionar una introducción básica a la estadística usando R, y no soy tan arrogante como para creer que la mía es mejor. Aún así, yo gusta bastante el libro, y tal vez a otras personas les resulte útil, aunque sea incompleto.

Danielle Navarro 13 de enero de 2013

Part I

Comienzo

Chapter 1

¿Por qué aprendemos estadística?

"No responderás cuestionarios
O cuestionarios sobre Asuntos Mundiales,
Tampoco con el cumplimiento
Realizaras ninguna prueba. No te sentarás
Con estadísticos ni cometerás
Una ciencia social"
– WH Auden[1]

1.1 Sobre la psicología de la estadística

Para sorpresa de muchas y muchos estudiantes, la estadística es una parte bastante significativa de la educación psicológica. Para sorpresa de nadie, la estadística rara vez es la parte *favorita* de la educación psicológica. Después de todo, si realmente te encantara la idea de hacer estadística, probablemente te habrías inscrito en una clase de estadística en este momento, no en una clase de psicología. Así que, como es lógico, hay una proporción bastante grande del estudiantado que no está contenta con el hecho de que la psicología tenga tanta estadística. Por todo ello, pensé que el lugar correcto para comenzar podría ser responder algunas de las preguntas más comunes que la gente tiene sobre la estadística.

Una gran parte de este tema en cuestión se relaciona con la idea misma de la estadística. ¿Qué es? ¿Para qué está ahí? ¿Y por qué los científicos están tan obsesionados con eso? Todas son buenas preguntas, si lo pensamos. Así que empecemos con la última. Como grupo, los científicos parecen estar extrañamente obsesionados con realizar pruebas estadísticas en todo. De hecho, usamos la estadística con tanta frecuencia que a veces nos olvidamos de explicarle a la gente por qué lo hacemos. Es una especie de artículo

[1] la cita proviene del poema de Auden de 1946 Under Which Lyre: A Reactionary Tract for the Times, pronunciado como parte de un discurso de graduación en la Universidad de Harvard. La historia del poema es bastante interesante: https://www.harvardmagazine.com/2007/11/a-poets-warning.html

de fe entre los científicos, y especialmente entre los científicos sociales, que no se puede confiar en los hallazgos hasta que hayamos utilizado la estadística. Se puede perdonar al estudiantado universitario por pensar que todos estamos completamente locos, porque nadie se toma la molestia de responder una pregunta muy sencilla:

¿Por qué haces estadística? ¿Por qué los científicos no usan el sentido común?

Es una pregunta ingenua en algunos aspectos, pero la mayoría de las buenas preguntas lo son. Hay muchas buenas respuestas, [2] pero, en mi opinión, la mejor respuesta es realmente sencilla: no confiamos lo suficiente en nosotras mismas. Nos preocupa que seamos humanos y susceptibles a todos los prejuicios, tentaciones y debilidades que sufren los humanos. Gran parte de la estadística es básicamente una salvaguarda. Usar el "sentido común" para evaluar la evidencia significa confiar en los instintos, confiar en argumentos verbales y usar el poder puro de la razón humana para llegar a la respuesta correcta. La mayoría de los científicos no cree que este enfoque funcione.

De hecho, ahora que lo pienso, esto me suena mucho a una pregunta psicológica, y dado que trabajo en un departamento de psicología, parece una buena idea profundizar un poco más aquí. ¿Es realmente plausible pensar que este enfoque de "sentido común" es muy fiable? Los argumentos verbales tienen que construirse con lenguaje, y todos los lenguajes tienen sesgos: algunas cosas son más difíciles de decir que otras, y no necesariamente porque sean falsas (p. ej., la electrodinámica cuántica es una buena teoría, pero difícil de explicar con palabras). Los instintos de nuestro "intestino" no están diseñados para resolver problemas científicos, están diseñados para manejar inferencias cotidianas, y dado que la evolución biológica es más lenta que el cambio cultural, deberíamos decir que están diseñados para resolver los problemas cotidianos para un *mundo diferente* al que vivimos. Fundamentalmente, el razonamiento sensato requiere que las personas participen en la "inducción", haciendo conjeturas sabias y llevando el razonamiento más allá de la evidencia inmediata de los sentidos para hacer generalizaciones sobre el mundo. Si crees que puedes hacer eso sin dejarte influir por diversos factores, bueno, no hace falta que continuemos discutiendo. Incluso, como muestra la siguiente sección, ni siquiera podemos resolver problemas "deductivos" (aquellos en los que no se requiere adivinar) sin que nuestros sesgos preexistentes nos influyan.

1.1.1 La maldición del sesgo de creencia

La gente es en su mayoría bastante inteligente. Ciertamente somos más inteligentes que las otras especies con las que compartimos el planeta (aunque muchas personas podrían estar en desacuerdo). Nuestras mentes son bastante sorprendentes, y parecemos ser capaces de las hazañas más increíbles del pensamiento y la razón. Aunque eso no nos hace perfectos. Y entre las muchas cosas que la Psicología ha demostrado a lo largo de los años es que realmente nos resulta difícil ser neutrales, evaluar la evidencia de manera imparcial y sin dejarnos influir por sesgos preexistentes. Un buen ejemplo de esto es el **efecto del sesgo de creencia** en el razonamiento lógico: si le pides a la gente que decida si un argumento en particular es lógicamente válido (es decir, la conclusión sería verdadera si las premisas fueran verdaderas), tendemos a estar influenciadas por la credibilidad de la conclusión, incluso cuando no deberíamos. Por ejemplo, aquí hay un argumento válido donde la conclusión es creíble:

> Todos los cigarrillos son caros (Premisa 1)

[2] incluye la sugerencia de que el sentido común escasea entre los científicos.

Algunas cosas adictivas son baratas (Premisa 2)
Por lo tanto, algunas cosas adictivas no son cigarrillos (Conclusión)

Y aquí hay un argumento válido donde la conclusión no es creíble:

Todas las cosas adictivas son caras (Premisa 1)
Algunos cigarrillos son baratos (Premisa 2)
Por lo tanto, algunos cigarrillos no son adictivos (Conclusión)

La *estructura* lógica del segundo argumento es idéntica a la estructura del primer argumento y ambos son válidos. Sin embargo, en el segundo argumento, hay buenas razones para pensar que la premisa 1 es incorrecta y, como resultado, es probable que la conclusión también sea incorrecta. Pero eso es completamente irrelevante para el tema en cuestión; un argumento es deductivamente válido si la conclusión es una consecuencia lógica de las premisas. Es decir, un argumento válido no tiene que involucrar declaraciones verdaderas.

Por otro lado, aquí hay un argumento inválido que tiene una conclusión creíble:

Todas las cosas adictivas son caras (Premisa 1)
Algunos cigarrillos son baratos (Premisa 2)
Por lo tanto, algunas cosas adictivas no son cigarrillos (Conclusión)

Y finalmente, un argumento inválido con una conclusión increíble:

Todos los cigarrillos son caros (Premisa 1)
Algunas cosas adictivas son baratas (Premisa 2)
Por lo tanto, algunos cigarrillos no son adictivos (Conclusión)

Ahora, supongamos que la gente es perfectamente capaz de dejar de lado sus prejuicios sobre lo que es cierto y lo que no, y evaluar un argumento únicamente por sus méritos lógicos. Esperaríamos que el 100 % de la gente dijera que los argumentos válidos son válidos y que el 0 % de la gente dijera que los argumentos inválidos son válidos. Por tanto, si hiciéramos un experimento sobre este tema, esperaríamos ver datos como los de @ tbl-tab1-1.

Table 1.1: Validez de los argumentos

	conclusion feels true	**conclusion feels false**
argument is valid	100% say "valid"	100% say "valid"
argument is invalid	0% say "valid"	0% say "valid"

Si los datos psicológicos fueran así (o incluso una buena aproximación), podríamos confiar en nuestro instinto. Es decir, estaría perfectamente bien dejar que los científicos evaluaran los datos basándose en su sentido común, y no molestarse con todo este turbio asunto de la estadística. Sin embargo, habéis hecho clases de psicología, y ahora probablemente ya sabéis a dónde va esto.

En un estudio clásico, J. St. B. T. Evans et al. (1983) realizó un experimento que analizaba exactamente esto. Lo que descubrieron es que cuando los sesgos preexistentes (es decir, las creencias) coincidían con la estructura de los datos, todo salía como se esperaba (Table 1.2).

Table 1.2: sesgos preexistentes y validez del argumento

	conclusion feels true	**conclusion feels false**
argument is valid	92% say "valid"	
argument is invalid		8% say "valid"

No es perfecto, pero es bastante bueno. Pero mira lo que sucede cuando nuestros sentimientos intuitivos sobre la verdad de la conclusión van en contra de la estructura lógica del argumento (Table 1.3):

Table 1.3: Intuición y validez de los argumentos

	conclusion feels true	**conclusion feels false**
argument is valid	92% say "valid"	46% say "valid"
argument is invalid	92% say "valid"	8% say "valid"

Vaya, eso no es tan bueno. Aparentemente, cuando a las personas se nos presenta un argumento sólido que contradice nuestras creencias preexistentes, nos resulta bastante difícil incluso percibirlo como un argumento sólido (la gente solo lo hizo el 46 % de las veces). Peor aún, cuando a las personas se nos presenta un argumento débil que está de acuerdo con nuestros prejuicios preexistentes, casi nadie puede ver que el argumento es débil (¡la gente se equivocó el 92 % de las veces!). [3]

Si lo piensas bien, no es que estos datos sean extremadamente incriminatorios. En general, a las personas les fue mejor que al azar para compensar sus sesgos anteriores, ya que alrededor del 60 % de los juicios de las personas fueron correctos (se esperaría que el 50 % fuera por casualidad). Aun así, si fueras un "evaluador de evidencia" profesional, y alguien viniera y te ofreciera una herramienta mágica que mejora tus posibilidades de tomar la decisión correcta del 60% al (digamos) 95%, probablemente lo aceptarías, ¿verdad? Por supuesto que lo harías. Afortunadamente, tenemos una herramienta que puede hacer esto. Pero no es magia, es la estadística. Esa es la razón número 1 por la que a los científicos les encanta la estadística. Es demasiado fácil para nosotros "creer lo que queremos creer". Entonces, si queremos "creer en los datos", vamos a necesitar un poco de ayuda para mantener bajo control nuestros sesgos personales. Eso es lo que hace la estadística, nos ayuda a mantenernos honestos.

1.2 La historia con moraleja de la paradoja de Simpson

La siguiente es una historia real (¡creo!). En 1973, la Universidad de California, Berkeley, tenía algunas preocupaciones sobre la admisión de estudiantes en sus cursos de posgrado. Específicamente, lo que causó el problema fue el desglose por género de sus admisiones (Table 1.4).

[3] En mis momentos más cínicos siento que este hecho por sí solo explica el 95% de lo que leo en Internet.

Table 1.4: Estudiantes de Berkeley por género

	Number of applicants	**Percent admitted**
Males	8442	44%
Females	4321	35%

Dado esto, ¡les preocupaba que los demandaran![4] Dado que había casi 13 000 solicitantes, una diferencia del 9 % en las tasas de admisión entre hombres y mujeres es demasiado grande para ser una coincidencia. Datos bastante convincentes, ¿verdad? Y si te dijera que estos datos *en realidad* reflejan un sesgo débil a favor de las mujeres (¡más o menos!), probablemente pensarías que estoy loca o soy sexista.

Muchas gracias a Wilfried Van Hirtum por señalarme esto.

Curiosamente, en realidad es cierto. Cuando las personas comenzaron a observar con más atención los datos de admisión, contaron una historia bastante diferente (Bickel et al., 1975). Específicamente, cuando lo analizaron departamento por departamento, resultó que la mayoría de los departamentos en realidad tenían una tasa de éxito ligeramente *más alta* para las mujeres solicitantes que para los hombres. Table 1.5 muestra las cifras de admisión para los seis departamentos más grandes (con los nombres de los departamentos eliminados por razones de privacidad):

Table 1.5: Estudiantes de Berkeley por género para los seis departamentos más grandes

	Males		**Females**	
Department	Applicants	Percent admitted	Applicants	Percent admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

Sorprendentemente, ¡la mayoría de los departamentos tenían una tasa *más alta* de admisiones para mujeres que para hombres! Sin embargo, la tasa general de admisión en la universidad para mujeres fue menor que para hombres. ¿Cómo puede ser esto? ¿Cómo pueden ambas afirmaciones ser verdaderas al mismo tiempo?

Esto es lo que está pasando. En primer lugar, fíjate que los departamentos no son iguales entre sí en términos de sus porcentajes de admisión: algunos departamentos (por ejemplo, A, B) tendían a admitir un alto porcentaje de los solicitantes calificados, mientras que otros (por ejemplo, F) tendían a rechazar la mayoría de los candidatos,

[4] Las versiones anteriores de estas notas sugirieron incorrectamente que en realidad fueran demandadas. Pero eso no es cierto. Hay un buen comentario sobre esto aquí: https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html

aunque fueran de alta calidad. Entonces, entre los seis departamentos que se muestran arriba, fíjate que el departamento A es el más generoso, seguido por B, C, D, E y F en ese orden. A continuación, observa que los hombres y las mujeres tendían a postularse a diferentes departamentos. Si clasificamos los departamentos en términos del número total de candidatos masculinos, obtenemos **A**>**B**>D>C>F>E (los departamentos "fáciles" están en negrita). En general, los hombres tendían a postularse a los departamentos que tenían altas tasas de admisión. Ahora compara esto con cómo se distribuyeron las candidatas. Clasificar los departamentos en términos del número total de candidatas produce un orden bastante diferente C>E>D>F>**A**>**B**. En otras palabras, lo que estos datos parecen sugerir es que las candidatas tendían a postularse a departamentos "más difíciles". Y de hecho, si observamos la Figura Figure 1.1 vemos que esta tendencia es sistemática y bastante llamativa. Este efecto se conoce como **paradoja de Simpson**. No es común, pero sucede en la vida real, y la mayoría de las personas se sorprenden mucho cuando lo encuentran por primera vez, y muchas personas se niegan incluso a creer que es real. Es muy real y si bien hay muchas lecciones estadísticas muy sutiles enterradas allí, quiero usarlas para resaltar un punto mucho más importante: investigar es difícil, y hay muchas trampas sutiles que contradicen la intuición al acecho de los desprevenidos. Esa es la razón número 2 por la que a los científicos les encantan las estadísticas y por la que enseñamos métodos de investigación. Porque la ciencia es difícil, y la verdad a veces se oculta astutamente en los rincones y grietas de datos complicados.

Antes de dejar este tema por completo, quiero señalar algo más realmente crítico que a menudo se pasa por alto en una clase de métodos de investigación. La estadística solo resuelve *parte* del problema. Recuerda que comenzamos todo esto con la preocupación de que los procesos de admisión de Berkeley pudieran estar sesgados injustamente en contra de las mujeres solicitantes. Cuando miramos los datos "agregados", parecía que la universidad estaba discriminando a las mujeres, pero cuando "desagregamos" y miramos el comportamiento individual de todos los departamentos, resultó que los departamentos estaban, en todo caso, ligeramente sesgados a favor de las mujeres. El sesgo de género en el total de admisiones se debió al hecho de que las mujeres tendían a autoseleccionarse para los departamentos más difíciles. Desde un punto de vista legal, eso probablemente eximiría a la universidad. Las admisiones de posgrado se determinan a nivel del departamento individual, y hay buenas razones para hacerlo. A nivel de departamentos individuales, las decisiones son más o menos imparciales (el débil sesgo a favor de las mujeres en ese nivel es pequeño y no es consistente en todos los departamentos). Dado que la universidad no puede dictar a qué departamentos eligen postularse las personas, y la toma de decisiones se lleva a cabo a nivel del departamento, difícilmente se le puede responsabilizar por cualquier sesgo que produzcan esas elecciones.

Esa fue la base de mis comentarios algo simplistas anteriores, pero esa no es exactamente toda la historia, ¿verdad? Después de todo, si estamos interesadas en esto desde una perspectiva más sociológica y psicológica, podríamos preguntarnos *por qué* hay diferencias de género tan marcadas en las solicitudes. ¿Por qué los hombres tienden a postularse a la ingeniería con más frecuencia que las mujeres, y por qué esto se invierte en el departamento de inglés? ¿Y por qué los departamentos que tienden a tener un sesgo de solicitud de mujeres tienden a tener tasas de admisión generales más bajas que aquellos departamentos que tienen un sesgo de solicitud de hombres? ¿No podría esto seguir reflejando un sesgo de género, a pesar de que cada departamento es imparcial en sí mismo? Que podría. Supongamos, hipotéticamente, que los hombres prefieren aplicar

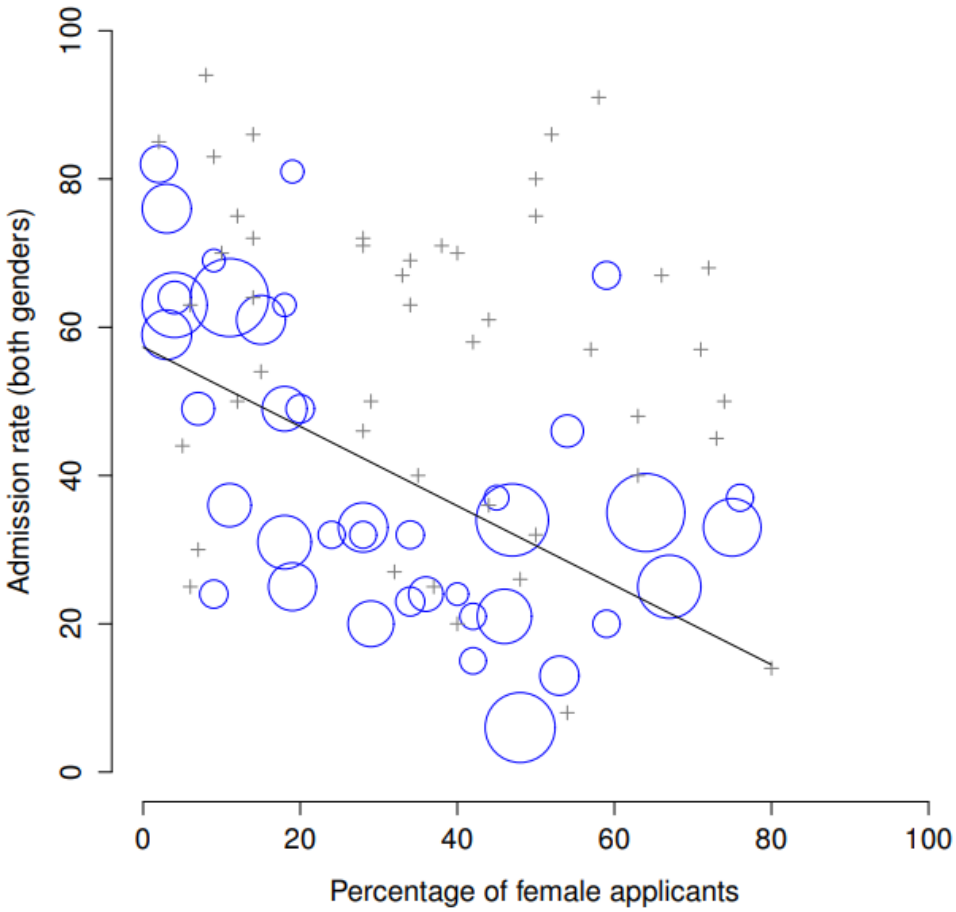


Figure 1.1: Los datos de admisión a la universidad de Berkeley de 1973. Esta cifra traza la tasa de admisión para los 85 departamentos que tenían al menos una aspirante mujer, en función del porcentaje de aspirantes que eran mujeres. La trama es un nuevo dibujo de la Figura 1 de Bickel et al. (1975). Los círculos parcelan departamentos con más de 40 aspirantes; el área del círculo es proporcional al número total de aspirantes. Los cruces parcelan los departamentos con menos de 40 aspirantes

a "ciencias duras" y las mujeres prefieren "humanidades". Y supongamos además que la razón por la cual los departamentos de humanidades tienen bajas tasas de admisión es porque el gobierno no quiere financiar las humanidades (los lugares de doctorado, por ejemplo, a menudo están vinculados a proyectos de investigación financiados por el gobierno). ¿Eso constituye un sesgo de género? ¿O simplemente una visión poco ilustrada del valor de las humanidades? ¿Qué pasaría si alguien de alto nivel en el gobierno recortara los fondos de humanidades porque sintiera que las humanidades son "cosas de chicas inútiles"? Eso parece bastante descaradamente sesgado por género. Nada de esto cae dentro del ámbito de la estadística, pero es importante para el proyecto de investigación. Si estás interesada en los efectos estructurales generales de los sutiles sesgos de género, entonces probablemente quieras ver los datos agregados y desagregados. Si estás interesada en el proceso de toma de decisiones en Berkeley, entonces probablemente solo estés interesada en los datos desagregados.

En resumen, hay muchas preguntas críticas que no puedes responder con estadísticas, pero las respuestas a esas preguntas tendrán un gran impacto en la forma en que analizas e interpretas los datos. Y esta es la razón por la que siempre debes pensar en la estadística como una herramienta para ayudarte a conocer tus datos. Nada mas y nada menos. Es una herramienta poderosa para ese fin, pero no hay sustituto para una reflexión cuidadosa.

1.3 Estadística en psicología

Espero que la discusión anterior haya ayudado a explicar por qué la ciencia en general está tan enfocada en la estadística. Pero supongo que tienes muchas más preguntas sobre el papel que juega la estadística en la psicología, y específicamente por qué las clases de psicología siempre dedican tantas horas a la estadística. Así que aquí está mi intento de responder a algunas de ellas…

¿Por qué la psicología tiene tanta estadística?

Para ser totalmente honesta, hay algunas razones diferentes, algunas de las cuales son mejores que otras. La razón más importante es que la psicología es una ciencia estadística. Lo que quiero decir con eso es que las "cosas" que estudiamos son *personas*. Gente real, complicada, gloriosamente desordenada, exasperantemente perversa. Las "cosas" de la física incluyen objetos como los electrones, y aunque surgen todo tipo de complejidades en la física, los electrones no tienen mente propia. No tienen opiniones, no difieren entre sí de forma extraña y arbitraria, no se aburren en medio de un experimento y no se enfadan con el experimentador y luego intentan sabotear deliberadamente el conjunto de datos (¡no es que lo haya hecho alguna vez!). En un nivel básico, la psicología es más difícil que la física.[5] Básicamente, os enseñamos estadística a vosotras como psicólogas porque necesitáis ser mejores en estadística que los físicos. En realidad, hay un dicho que se usa a veces en física, en el sentido de que "si tu experimento necesita estadística, deberías haber hecho un experimento mejor". Tienen el lujo de poder decir eso porque sus objetos de estudio son tristemente simples en comparación con el enorme lío al que se enfrentan los científicos sociales. Y no es solo psicología. La mayoría de las ciencias sociales dependen desesperadamente de la estadística. No porque seamos malos experimentadores, sino porque hemos elegido un problema más difícil de resolver. Te enseñamos estadística porque realmente la necesitas.

[5] lo que podría explicar por qué la física está un poco más avanzada como ciencia que nosotros.

¿Puede cualquiera hacer estadística?

Hasta cierto punto, pero no completamente. Es cierto que no necesitas convertirte en un estadístico completamente capacitado solo para hacer psicología, pero sí necesitas alcanzar un cierto nivel de competencia estadística. En mi opinión, hay tres razones por las que todo investigador psicológico debería poder hacer estadística básica:

- En primer lugar, está la razón fundamental: la estadística está profundamente entrelazada con el diseño de la investigación. Si quieres ser buena en el diseño de estudios psicológicos, al menos debes comprender los conceptos básicos de la estadística.
- En segundo lugar, si quieres ser buena en el aspecto psicológico de la investigación, entonces debes ser capaz de comprender la literatura psicológica, ¿verdad? Pero casi todos los artículos de la literatura psicológica informan los resultados de los análisis estadísticos. Entonces, si realmente quieres comprender la psicología, debes poder comprender lo que otras personas hicieron con sus datos. Y eso significa comprender una cierta cantidad de estadísticas.
- En tercer lugar, existe un gran problema práctico al depender de otras personas para realizar todas las estadísticas: el análisis estadístico es *caro*. Si alguna vez te aburres y quieres averiguar cuánto cobra el gobierno australiano por las tasas universitarias, notarás algo interesante: las estadísticas están designadas como una categoría de "prioridad nacional", por lo que las tasas son mucho, mucho más bajas que para cualquier otra área de estudio. Esto se debe a que existe una escasez masiva de estadísticos. Entonces, desde tu perspectiva como investigador psicológico, ¡las leyes de la oferta y la demanda no están exactamente de tu lado aquí! Como resultado, en casi cualquier situación de la vida real en la que desees realizar una investigación psicológica, la cruda realidad será que no tendrás suficiente dinero para pagar a un estadístico. Entonces, la economía de la situación significa que tienes que ser bastante autosuficiente.

Ten en cuenta que muchas de estas razones se generalizan más allá de los investigadores. Si quieres ser una psicóloga en ejercicio y mantenerte al día en el campo, es útil poder leer la literatura científica, que se basa en gran medida en la estadística.

No me importan los trabajos, la investigación o el trabajo clínico. ¿Necesito la estadística?

Está bien, ahora solo estás jugando conmigo. Aún así, creo que debería importarte a ti también. La estadística debe ser importante para ti de la misma manera que la estadística debe ser importante para *todas*. Vivimos en el siglo XXI y los datos están *en todas partes*. Francamente, dado el mundo en el que vivimos en estos días, ¡un conocimiento básico de estadística es bastante parecido a una herramienta de supervivencia! Lo cual es el tema de la siguiente sección.

1.4 La Estadística en la vida cotidiana

> *"Nos estamos ahogando en información,*
> *pero estamos hambrientos de conocimiento"*
> - Varios autores, original probablemente John Naisbitt

Cuando empecé a redactar mis apuntes de clase, cogí los 20 artículos más recientes publicados en la web de noticias de ABC. De esos 20 artículos, resultó que en 8 de ellos

se discutía algo que yo llamaría un tema estadístico y en 6 de ellos se cometía un error. El error más común, si tienes curiosidad, fue no informar de los datos de referencia (p. ej., el artículo menciona que el 5% de las personas en la situación X tienen alguna característica Y, pero no dice lo común que es la característica para todos los demás). Lo que quiero decir con esto no es que los periodistas sean malos en estadística (aunque casi siempre lo son), sino que un conocimiento básico de estadística es muy útil para intentar averiguar cuándo alguien está cometiendo un error o incluso mintiéndote. De hecho, una de las cosas más importantes que te aporta el conocimiento de la estadística es que te enfadas con el periódico o con Internet con mucha más frecuencia. Puedes encontrar un buen ejemplo de esto en Section 4.1.5 en el Chapter 4. En versiones posteriores de este libro intentaré incluir más anécdotas en ese sentido.

1.5 Los métodos de investigación van más allá de las estadísticas

Hasta ahora, la mayor parte de lo que he hablado es de estadística, por lo que se te perdonaría que pensaras que la estadística es lo único que me interesa en la vida. Para ser justos, no estarías muy equivocada, pero la metodología de investigación es un concepto más amplio que la estadística. Por eso, la mayoría de los cursos de métodos de investigación abarcan muchos temas relacionados con la pragmática del diseño de la investigación y, en particular, con los problemas que surgen al investigar con seres humanos. Sin embargo, alrededor del 99% de los temores del alumnado se relacionan con la parte de estadística del curso, por lo que me he centrado en la estadística en esta discusión y espero haberte convencido de que la estadística es importante y, lo que es más importante, que no hay que temerla. Dicho esto, es bastante típico que las clases de introducción a los métodos de investigación estén muy cargadas de estadística. Esto no se debe (normalmente) a que los profesores y profesoras sean malas personas. Todo lo contrario. Las clases introductorias se centran mucho en la estadística porque casi siempre se necesita la estadística antes que el resto de la formación sobre métodos de investigación. ¿Por qué? Porque casi todas las tareas de las demás clases se basarán en la formación estadística, mucho más que en otras herramientas metodológicas. No es habitual que los trabajos de grado requieran que diseñes tu propio estudio desde cero (en cuyo caso necesitarías saber mucho sobre diseño de investigación), pero *es* habitual que los trabajos te pidan que analices e interpretes los datos que se recogieron en un estudio que diseñó otra persona (en cuyo caso necesitas estadística). En ese sentido, desde la perspectiva de que te permita rendir bien en el resto de tus clases, la estadística es más urgente.

Pero ten en cuenta que "urgente" no es lo mismo que "importante": ambos son importantes. Quiero insistir en que el diseño de la investigación es tan importante como el análisis de datos, y este libro le dedica bastante tiempo. Sin embargo, mientras que la estadística tiene una especie de universalidad y proporciona un conjunto de herramientas básicas que son útiles para la mayoría de los tipos de investigación psicológica, los métodos de investigación no son tan universales. Hay algunos principios generales que todo el mundo debería tener en cuenta, pero gran parte del diseño de la investigación es muy idiosincrásico y específico del área de investigación a la que quieras dedicarte. En la medida en que son los detalles lo que importan, esos detalles no suelen aparecen en una clase introductoria de estadística y métodos de investigación.

Chapter 2

Una breve introducción al diseño de investigación

> *"A menudo consultar al estadístico una vez finalizado un experimento no es más que pedirle que realice un examen post mortem. Quizás pueda decir de qué murió el experimento".*
> – Sir Ronald Fisher[1]

En este capítulo, vamos a empezar a reflexionar sobre las ideas básicas que intervienen en el diseño de un estudio, la recogida de datos, la comprobación de si la recogida de datos funciona, etc. No te dará suficiente información para permitirte diseñar tus propios estudios, pero sí muchas de las herramientas básicas que necesitas para evaluar los estudios realizados por otras personas. Sin embargo, como este libro se centra mucho más en el análisis de datos que en su recogida, solo voy a dar una breve visión general. Ten en cuenta que este capítulo es "especial" en dos sentidos. En primer lugar, es mucho más específico de psicología que los capítulos posteriores. En segundo lugar, se centra mucho más en el problema científico de la metodología de la investigación y mucho menos en el problema estadístico del análisis de datos. Sin embargo, los dos problemas están relacionados entre sí, por lo que es tradicional que los libros de texto de estadística analicen el problema con un poco de detalle. Este capítulo se basa en gran medida en Campbell & Stanley (1963) y Stevens (1946) para la discusión de las escalas de medida.

2.1 Introducción a la medición psicológica

Lo primero que hay que entender es que la recogida de datos se puede considerar un tipo de **medida**. Es decir, lo que estamos tratando de hacer aquí es medir algo sobre el comportamiento humano o la mente humana. ¿Qué quiero decir con "medida"?

2.1.1 Algunas reflexiones sobre la medición psicológica

La medición en sí es un concepto sutil, pero básicamente se reduce a encontrar alguna forma de asignar números, etiquetas o algún otro tipo de descripciones bien definidas a

[1] Discurso presidencial ante el Primer Congreso de Estadística de la India, 1938. Fuente: http://en.wikiquote.org/wiki/Ronald Fisher

las "cosas". Por tanto, cualquiera de los siguientes elementos podría considerarse una medida psicológica:

- Mi **edad** es *33* años.
- *No* **me gustan las anchoas.**
- Mi **género cromosómico** es *masculino*.
- Mi **género autoidentificado** es *femenino*.

En la breve lista anterior, la **parte en negrita** es "lo que se va a medir", y la *parte en cursiva* es "la medida en sí". De hecho, podemos ampliarlo un poco más, pensando en el conjunto de posibles medidas que podrían haber surgido en cada caso:

- Mi **edad** (en años) podría haber sido *0, 1, 2, 3* ..., etc. El límite superior de lo que podría ser mi edad es un poco difuso, pero en la práctica se puede decir que la mayor edad posible es *150*, ya que ningún ser humano ha vivido tanto tiempo.
- A la pregunta de si **me gustan las anchoas**, podría haber respondido que *me gustan, o no, o no tengo opinión, o a veces me gustan.*
- Es casi seguro que mi **género cromosómico** será *masculino* (XY) o *femenino* (XX), pero existen otras posibilidades. También podría tener *síndrome de Klinfelter* (XXY), que es más parecido al masculino que al femenino. E imagino que también hay otras posibilidades.
- También es muy probable que mi género **autoidentificado** sea masculino o femenino, pero no tiene por qué coincidir con mi género cromosómico. También puedo elegir identificarme con *ninguno*, o llamarme explícitamente *transgénero*.

Como puedes ver, para algunas cosas (como la edad) parece bastante obvio cuál debería ser el conjunto de medidas posibles, mientras que para otras cosas la cosa se complica un poco. Pero quiero señalar que incluso en el caso de la edad de alguien es mucho más sutil que esto. Por ejemplo, en el ejemplo anterior asumí que estaba bien medir la edad en años. Pero si eres un psicólogo del desarrollo, eso es demasiado burdo, por lo que a menudo se mide la edad en *años y meses* (si un niño tiene 2 años y 11 meses, se suele escribir como "2;11"). Si te interesan los recién nacidos, quizás prefieras medir la edad en *días desde el nacimiento*, o incluso en *horas desde el nacimiento*. En otras palabras, la forma de especificar los valores de medición permitidos es importante.

Si lo analizamos un poco más detenidamente, nos daremos cuenta de que el concepto de "edad" no es tan preciso. En general, cuando decimos "edad" implícitamente queremos decir "el tiempo transcurrido desde el nacimiento". Pero no siempre es así. Supongamos que nos interesa saber cómo los bebés recién nacidos controlan sus movimientos oculares. Si te interesan los niños tan pequeños, es posible que también empieces a preocuparte de que el "nacimiento" no sea el único momento significativo del que preocuparse. Si Alice nace 3 semanas prematura y Bianca nace 1 semana tarde, ¿tendría sentido decir que tienen "la misma edad" si las encontramos "2 horas después de nacer"? En cierto sentido, sí. Por convención social, usamos el nacimiento como punto de referencia para hablar de la edad en la vida cotidiana, ya que define el tiempo que la persona lleva funcionando como una entidad independiente en el mundo. Pero desde una perspectiva científica no es lo único que nos importa. Cuando pensamos en la biología de los seres humanos, suele ser útil considerarnos organismos que han estado creciendo y madurando desde su concepción, y desde esa perspectiva, Alice y Bianca no tienen la misma edad en absoluto. Por lo tanto, es posible que queramos definir el concepto de "edad" de dos maneras diferentes: el tiempo transcurrido desde la concepción y el tiempo transcurrido desde el nacimiento. Cuando se trata de adultos no hay mucha diferencia, pero cuando

se trata de recién nacidos sí.

Más allá de estas cuestiones, está la cuestión de la metodología. ¿Qué "método de medición" específico vas a usar para averiguar la edad de alguien? Como antes, hay muchas posibilidades:

- Podrías preguntarle a la gente "¿cuántos años tienes?" El método de autoinforme es rápido, barato y fácil. Pero solo funciona con personas de edad suficiente para entender la pregunta, y algunas mienten sobre su edad.
- Podrías preguntarle a una autoridad (p. ej., un padre) "¿cuántos años tiene su hijo?" Este método es rápido y cuando se trata de niños no es tan difícil ya que los padres casi siempre están presentes. No funciona tan bien si quieres saber la "edad desde la concepción", ya que muchos padres no pueden decir con certeza cuándo tuvo lugar la concepción. Para eso, es posible que necesites una autoridad diferente (por ejemplo, un obstetra).
- Puedes buscar registros oficiales, por ejemplo, certificados de nacimiento o defunción. Es una tarea larga y frustrante, pero tiene su utilidad (por ejemplo, si la persona ya ha fallecido).

2.1.2 Operativización: definiendo la medida

Todas las ideas expuestas en la sección anterior se relacionan con el concepto de **operativización**. Para precisar un poco más la idea, la operativización es el proceso mediante el cual tomamos un concepto significativo pero algo vago y lo convertimos en una medida precisa. El proceso de operativización puede implicar varias cosas diferentes:

- Ser preciso sobre lo que se intenta medir. Por ejemplo, ¿"edad" significa "tiempo desde el nacimiento" o "tiempo desde la concepción" en el contexto de tu investigación?
- Determinar qué método usarás para medirlo. ¿Utilizarás el autoinforme para medir la edad, preguntarás a uno de los padres o buscarás un registro oficial? Si utilizas autoinforme, ¿cómo formularás la pregunta?
- Definir el conjunto de valores admisibles que puede tomar la medida. Ten en cuenta que estos valores no siempre tienen que ser numéricos, aunque a menudo lo son. Cuando se mide la edad, los valores son numéricos, pero aún así debemos pensar cuidadosamente qué números están permitidos. ¿Queremos la edad en años, años y meses, días u horas? Para otros tipos de medidas (p. ej., sexo), los valores no son numéricos. Pero, al igual que antes, debemos pensar qué valores están permitidos. Si pedimos a los encuestados que indiquen su sexo, ¿entre qué opciones les permitimos elegir? ¿Es suficiente permitir solo "hombre" o "mujer"? ¿Es necesaria la opción "otro"? ¿O no deberíamos dar a la gente opciones específicas y dejar que respondan con sus propias palabras? Y si abrimos el conjunto de valores posibles para incluir todas las respuestas verbales, ¿cómo interpretamos sus respuestas?

La operativización es un asunto complicado, y no hay una "única y verdadera manera" de hacerlo. La forma de operativizar el concepto informal de "edad" o "sexo" para convertirlo en una medida formal depende del uso que se le quiera dar. A menudo la comunidad científica que trabaja en tu área tiene ideas bastante consolidadas sobre cómo hacerlo. En otras palabras, la operativización debe estudiarse caso por caso.

Sin embargo, aunque hay muchas cuestiones que son específicas de cada proyecto de investigación, hay algunos aspectos que son bastante generales.

Antes de continuar, quiero aclarar la terminología y, de paso, introducir un término más. He aquí cuatro cosas diferentes que están estrechamente relacionadas entre sí:

- **Un constructo teórico.** Es aquello que se intenta medir, como "edad", "sexo" o una "opinión". Un constructo teórico no se puede observar directamente y, a menudo, son un poco vagos.
- **Una medida.** La medida se refiere al método o la herramienta que se utiliza para realizar las observaciones. Una pregunta en una encuesta, una observación del comportamiento o un escáner cerebral pueden considerarse medidas.
- **Una operativización.** El término "operativización" se refiere a la conexión lógica entre la medida y el constructo teórico, o al proceso mediante el cual intentamos derivar una medida a partir de un constructo teórico.
- **Una variable.** Finalmente, un nuevo término. Una variable es lo que obtenemos cuando aplicamos nuestra medida a algo del mundo. Es decir, las variables son los "datos" reales con los que terminamos en nuestros conjuntos de datos.

En la práctica, incluso los científicos tienden a difuminar la distinción entre estas cosas, pero es muy útil intentar comprender las diferencias.

2.2 Escalas de medida

Como se indica en el apartado anterior, el resultado de una medición psicológica se denomina variable. Pero no todas las variables son del mismo tipo, por lo que es útil entender qué tipos hay. Un concepto muy útil para distinguir entre diferentes tipos de variables es lo que se conoce como **escalas de medida**.

2.2.1 Escala nominal

Una variable de **escala nominal** (también conocida como variable **categórica**) es aquella en la que no existe una relación particular entre las diferentes posibilidades. Para este tipo de variables no tiene ningún sentido decir que una de ellas es "mayor" o "mejor" que otra, y no tiene absolutamente ningún sentido promediarlas. El ejemplo clásico es el "color de ojos". Los ojos pueden ser azules, verdes o marrones, entre otras posibilidades, pero ninguna de ellas es "más grande" que otra. Por tanto, sería muy extraño hablar de un "color de ojos promedio". Del mismo modo, el género también es nominal: el hombre no es mejor ni peor que la mujer. Tampoco tiene sentido hablar de un "género medio". En resumen, las variables de escala nominal son aquellas para las que lo único que se puede decir sobre las diferentes posibilidades es que son diferentes. Eso es todo.

Veámoslo con un poco más de detalle. Supongamos que estoy investigando cómo se desplaza la gente hacia y desde el trabajo. Una variable que tendría que medir sería qué tipo de transporte usa la gente para ir a trabajar. Esta variable "tipo de transporte" podría tener bastantes valores posibles, entre ellos: "tren", "autobús", "coche", "bicicleta". De momento, supongamos que estas cuatro son las únicas posibilidades. Imaginemos entonces que le pregunto a 100 personas cómo han llegado hoy al trabajo, con este resultado (Table 2.1).

Table 2.1: ¿Cómo llegaron 100 personas al trabajo hoy?

Transportation	Number of people
(1) Train	12
(2) Bus	30
(3) Car	48
(4) Bicycle	10

Entonces, ¿cuál es el tipo de transporte promedio? Obviamente, la respuesta es que no hay ninguno. Es una pregunta tonta. Se puede decir que viajar en coche es el método más popular y viajar en tren es el menos popular, pero eso es todo. Del mismo modo, fíjate que el orden en que enumero las opciones no es muy interesante. Podría haber elegido mostrar los datos como en Table 2.2.

Table 2.2: Cómo llegaron 100 personas al trabajo hoy, una vista diferente

Transportation	Number of people
(3) Car	48
(1) Train	12
(4) Bicycle	10
(2) Bus	30

… y nada cambia realmente.

2.2.2 Escala ordinal

Las variables de **escala ordinal** tienen un poco más de estructura que las variables de escala nominal, pero no mucho. Una variable de escala ordinal es aquella en la que existe una forma natural y significativa de ordenar las diferentes posibilidades, pero no se puede hacer nada más. El ejemplo habitual de una variable ordinal es "posición final en una carrera". *Puedes* decir que la persona que terminó primera fue más rápida que la que terminó segunda, pero *no* sabes cuánto más rápida. En consecuencia, sabemos que $1^o > 2^o$, y sabemos que $2^o > 3^o$, pero la diferencia entre el 1^o y el 2^o podría ser mucho mayor que la diferencia entre el 2^o y el 3^o.

He aquí un ejemplo psicológicamente más interesante. Supongamos que me interesan las actitudes de las personas hacia el cambio climático. Entonces pido a algunas personas que elijan la afirmación (de las cuatro enumeradas) que más se acerque a sus creencias:

1. Las temperaturas están aumentando debido a la actividad humana
2. Las temperaturas están aumentando pero no sabemos por qué
3. Las temperaturas están aumentando pero no a causa de la actividad humana
4. Las temperaturas no están aumentando

Observa que estas cuatro afirmaciones tienen un orden natural, en términos de "hasta qué punto coinciden con la ciencia actual". La afirmación 1 es muy parecida, la afirmación 2 es razonable, la afirmación 3 no es muy parecida y la afirmación 4 se opone rotundamente a la ciencia actual. Así que, en términos de lo que me interesa (hasta qué

punto la gente está de acuerdo con la ciencia), puedo ordenar las opciones como $1 > 2 > 3 > 4$. Dado que existe este orden, sería muy raro enumerar las opciones así…

1. Las temperaturas están aumentando pero no a causa de la actividad humana
2. Las temperaturas están aumentando debido a la actividad humana
3. Las temperaturas no están aumentando
4. Las temperaturas están aumentando pero no sabemos por qué

… porque parece violar la "estructura" natural de la pregunta.

Entonces, supongamos que hago estas preguntas a 100 personas y obtengo las respuestas que se muestran en Table 2.3.

Table 2.3: Actitudes ante el cambio climático

Response	**Number**
(1) Temperatures are rising because of human activity	51
(2) Temperatures are rising but we donâ t know why	20
(3) Temperatures are rising but not because of humans	10
(4) Temperatures are not rising	19

Al analizar estos datos, parece bastante razonable tratar de agrupar (1), (2) y (3) y decir que 81 de cada 100 personas estaban dispuestas a respaldar *al menos parcialmente* la ciencia. Y también es bastante razonable agrupar (2), (3) y (4) y decir que 49 de cada 100 personas marcaron *al menos algún desacuerdo* con la opinión científica dominante. Sin embargo, sería totalmente extraño intentar agrupar (1), (2) y (4) y decir que 90 de cada 100 personas dijeron… ¿qué? No hay nada sensato que permita agrupar esas respuestas.

Dicho esto, observa que si bien *podemos* usar el orden natural de estos elementos para construir agrupaciones razonables, lo que no podemos hacer es promediarlos. Por ejemplo, en mi sencillo ejemplo, la respuesta "promedio" a la pregunta es 1.97. Si me puedes decir qué significa eso, me encantaría saberlo, ¡porque me parece un galimatías!

2.2.3 Escala de intervalo

A diferencia de las variables de escala nominal y ordinal, las variables de **escala de intervalo** y de escala de razón son variables para las que el valor numérico es realmente significativo. En el caso de las variables de escala de intervalo, las *diferencias* entre los números son interpretables, pero la variable no tiene un valor cero "natural". Un buen ejemplo de una variable de escala de intervalo es medir la temperatura en grados centígrados. Por ejemplo, si ayer hacía 15° y hoy 18°, la diferencia de 3° entre ambas es realmente significativa. Además, esa diferencia de 3° es *exactamente la misma* que la diferencia de 3° entre 7° y 10°. En resumen, la suma y la resta tienen sentido para las variables de escala de intervalo.[2]

[2]En realidad, lectores con más conocimientos de física que yo me han informado de que la temperatura no es estrictamente una escala de intervalo, en el sentido de que la cantidad de energía necesaria

Sin embargo, fíjate que 0° no significa "ninguna temperatura". En realidad significa "la temperatura a la que se congela el agua", lo cual es bastante arbitrario. En consecuencia, no tiene sentido intentar multiplicar y dividir las temperaturas. Es incorrecto decir que 20° es el doble de caliente que 10°, del mismo modo que es extraño y carece de sentido intentar afirmar que 20° es dos veces más caliente que -10°.

Veamos de nuevo un ejemplo más psicológico. Supongamos que me interesa analizar cómo han cambiado las actitudes de los estudiantes universitarios de primer año con el tiempo. Obviamente, voy a querer registrar el año en el que empezó cada estudiante. Se trata de una variable de escala de intervalo. Un estudiante que empezó en 2003 llegó 5 años antes que un estudiante que empezó en 2008. Sin embargo, sería completamente absurdo dividir 2008 entre 2003 y decir que el segundo estudiante empezó "1,0024 veces más tarde" que el primero. Eso no tiene ningún sentido.

2.2.4 Escala de razón

El cuarto y último tipo de variable a considerar es una variable de **escala de razón**, en la que cero significa realmente cero, y está bien multiplicar y dividir. Un buen ejemplo psicológico de una variable de escala de razón es el tiempo de respuesta (TR). En muchas tareas es muy común registrar la cantidad de tiempo que alguien tarda en resolver un problema o responder una pregunta, porque es un indicador de lo difícil que es la tarea. Supongamos que Alan tarda 2,3 segundos en responder a una pregunta, mientras que Ben tarda 3,1 segundos. Al igual que con una variable de escala de intervalo, la suma y la resta tienen sentido en este caso. Ben realmente tardó 3,1 - 2,3 = 0,8 segundos más que Alan. Sin embargo, fíjate que la multiplicación y la división también tienen sentido aquí: Ben tardó 3,1/2,3 = 1,35 veces más que Alan en responder la pregunta. Y la razón por la que puedes hacer esto es que para una variable de escala de razón como TR, "cero segundos" realmente significa "nada de tiempo".

2.2.5 Variables continuas versus discretas

Hay un segundo tipo de distinción que debes conocer, con respecto a los tipos de variables con las que puedes encontrarte. Se trata de la distinción entre variables continuas y variables discretas (Table 2.4). La diferencia entre ellas es la siguiente:

- Una **variable continua** es aquella en la que, para dos valores cualesquiera que se te ocurran, siempre es lógicamente posible tener otro valor en medio.
- Una **variable discreta** es, en efecto, una variable que no es continua. En el caso de una variable discreta, a veces no hay nada en medio.

Probablemente estas definiciones parezcan un poco abstractas, pero son bastante sencillas si vemos algunos ejemplos. Por ejemplo, el tiempo de respuesta es continuo. Si Alan tarda 3,1 segundos y Ben tarda 2,3 segundos en responder a una pregunta, el tiempo de respuesta de Cameron estará en el medio si tarda 3,0 segundos. Y, por supuesto, también sería posible que David tardara 3,031 segundos en responder, lo que significa que su TR estaría entre el de Cameron y el de Alan. Y aunque en la práctica sea imposible medir TR con tanta precisión, en principio es posible. Dado que siempre podemos

para calentar algo 3° depende de su temperatura actual. Por tanto, en el sentido que interesa a los físicos, la temperatura no es en realidad una escala de intervalo. Pero sigue siendo un buen ejemplo, así que voy a ignorar esta pequeña verdad incómoda.

Table 2.4: La relación entre las escalas de medida y la distinción discreta/continua. Las celdas con una marca de verificación corresponden a cosas que son posibles

	continuous	**discrete**
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

encontrar un nuevo valor de TR entre dos valores cualesquiera, consideramos que el TR es una medida continua.

Las variables discretas ocurren cuando se infringe esta regla. Por ejemplo, las variables de escala nominal siempre son discretas. No hay un tipo de transporte que se encuentre "entre" los trenes y las bicicletas, no de la forma matemática estricta en que 2,3 se encuentra entre 2 y 3. Por lo tanto, el tipo de transporte es discreto. Del mismo modo, las variables de escala ordinal siempre son discretas. Aunque el "segundo lugar" se encuentra entre el "primer lugar" y el "tercer lugar", no hay nada que pueda estar lógicamente entre el "primer lugar" y el "segundo lugar". Las variables de escala de intervalo y escala de razón pueden ir en cualquier dirección. Como vimos anteriormente, el tiempo de respuesta (una variable de escala de razón) es continuo. La temperatura en grados centígrados (una variable de escala de intervalo) también es continua. Sin embargo, el año en que fuiste a la escuela (una variable de escala de intervalo) es discreto. No hay ningún año entre 2002 y 2003. El número de preguntas que aciertas en una prueba de verdadero o falso (una variable de escala de razón) también es discreto. Dado que una pregunta de verdadero o falso no permite ser "parcialmente correcta", no hay nada entre 5/10 y 6/10. Table 2.4 resume la relación entre las escalas de medida y la distinción discreta/continua. Las celdas con una marca de verificación corresponden a cosas que son posibles. Intento insistir en este punto porque (a) algunos libros de texto se equivocan y (b) la gente suele decir "variable discreta" cuando quiere decir "variable de escala nominal". Es una lástima.

2.2.6 Algunos aspectos complejos

Sé que te va a sorprender oír esto, pero el mundo real es mucho más complicado de lo que sugiere este pequeño esquema de clasificación. Muy pocas variables de la vida real encajan en estas bonitas categorías, por lo que hay que tener cuidado de no tratar las escalas de medida como si fueran reglas rígidas. No funcionan así. Son directrices que te ayudan a pensar en las situaciones en las que debes tratar diferentes variables de manera diferente. Nada mas.

Miremos un ejemplo clásico, tal vez *el* ejemplo clásico, de una herramienta de medición psicológica: la **escala Likert**. La humilde escala Likert es el pan de cada día en el diseño de encuestas. Tú misma has completado cientos, tal vez miles, de ellas y lo más probable es que incluso hayas usado una. Supongamos que tenemos una pregunta de encuesta parecida a esta:

¿Cuál de las siguientes opciones describe mejor su opinión sobre la afirmación de que "todos los piratas son increíbles"?

y luego las opciones que se le presentan al participante son estas:

1. Totalmente en desacuerdo
2. En desacuerdo
3. Ni de acuerdo ni en desacuerdo
4. De acuerdo
5. Totalmente de acuerdo

Este conjunto de ítems es un ejemplo de una escala Likert de 5 puntos, en la que se pide a las personas que elijan entre varias (en este caso 5) posibilidades claramente ordenadas, generalmente con un descriptor verbal dado en cada caso. Sin embargo, no es necesario que todos los elementos se describan explícitamente. Este es un buen ejemplo de una escala Likert de 5 puntos también:

1. Totalmente en desacuerdo
2.
3.
4.
5. Totalmente de acuerdo

Las escalas Likert son herramientas muy útiles, aunque algo limitadas. La pregunta es ¿qué tipo de variable son? Obviamente son discretas, ya que no se puede dar una respuesta de 2.5. Obviamente no son de escala nominal, ya que los ítems están ordenados; y tampoco son escalas de razón, ya que no hay un cero natural.

¿Pero son escala ordinal o escala de intervalo? Uno de los argumentos dice que no podemos demostrar que la diferencia entre "totalmente de acuerdo" y "de acuerdo" sea del mismo tamaño que la diferencia entre "de acuerdo" y "ni de acuerdo ni en desacuerdo". De hecho, en la vida cotidiana es bastante obvio que no son lo mismo. Esto sugiere que deberíamos tratar las escalas Likert como variables ordinales. Por otro lado, en la práctica, la mayoría de los participantes parecen tomarse bastante en serio la parte "en una escala del 1 al 5", y tienden a actuar como si las diferencias entre las cinco opciones de respuesta fueran bastante similares entre sí. Como consecuencia, muchos investigadores tratan los datos de la escala Likert como una escala de intervalo.[3] No es una escala de intervalo, pero en la práctica se acerca lo suficiente como para pensar en ella como si fuera una **escala de cuasi-intervalo**.

2.3 Evaluación de la fiabilidad de una medida

En este punto, hemos pensado un poco sobre cómo operativizar un constructo teórico y, por lo tanto, crear una medida psicológica. Y hemos visto que al aplicar medidas psicológicas terminamos con variables, que pueden ser de muchos tipos diferentes. En este punto, deberíamos comenzar a discutir la pregunta obvia: ¿es buena la medición? Haremos esto en términos de dos ideas relacionadas: *fiabilidad* y *validez*. En pocas palabras, la **fiabilidad** de una medida te dice con qué precisión está midiendo algo, mientras que la validez de una medida te dice qué tan precisa es la medida. En esta sección hablaré sobre fiabilidad; hablaremos sobre la validez en la sección [Evaluación de la validez de un estudio].

La fiabilidad es en realidad un concepto muy simple. Se refiere a la repetibilidad o

[3] Ah, la psicología... ¡nunca hay una respuesta fácil para nada!

consistencia de tu medición. La medida de mi peso por medio de una "balanza de baño" es muy fiable. Si subo y bajo de la balanza una y otra vez, me seguirá dando la misma respuesta. Medir mi inteligencia por medio de "preguntarle a mi mamá" es muy poco fiable. Algunos días me dice que soy un poco torpe y otros días me dice que soy un completo idiota. Ten en cuenta que este concepto de fiabilidad es diferente a la cuestión de si las medidas son correctas (la corrección de una medida se relaciona con su validez). Si estoy sosteniendo un saco de patatas cuando subo y bajo de la báscula del baño, la medición seguirá siendo fiable: siempre me dará la misma respuesta. Sin embargo, esta respuesta altamente fiable no coincide en absoluto con mi peso real, por lo tanto, es incorrecta. En términos técnicos, esta es una medida fiable pero inválida. Del mismo modo, aunque la estimación de mi madre sobre mi inteligencia es poco fiable, puede que tenga razón. Tal vez simplemente no soy demasiado brillante, y aunque su estimación de mi inteligencia fluctúa bastante de un día para otro, básicamente es correcta. Esa sería una medida poco fiable pero válida. Por supuesto, si las estimaciones de mi madre son demasiado poco fiables, será muy difícil averiguar cuál de sus muchas afirmaciones sobre mi inteligencia es realmente la correcta. En cierta medida, pues, una medida muy poco fiable tiende a resultar inválida a efectos prácticos; tanto es así que mucha gente diría que la fiabilidad es necesaria (pero no suficiente) para asegurar la validez.

Bien, ahora que tenemos clara la distinción entre fiabilidad y validez, pensemos en las diferentes formas en que podríamos medir la fiabilidad:

- **Fiabilidad test-retest**. Esto se relaciona con la consistencia en el tiempo. Si repetimos la medición en una fecha posterior, ¿obtenemos la misma respuesta?
- **Fiabilidad entre evaluadores**. Esto se relaciona con la consistencia entre las personas. Si alguien más repite la medición (p. ej., alguien más califica mi inteligencia), ¿producirá la misma respuesta?
- **Fiabilidad de formas paralelas**. Esto se relaciona con la consistencia entre mediciones teóricamente equivalentes. Si uso un juego diferente de básculas de baño para medir mi peso, ¿da la misma respuesta?
- **fiabilidad de consistencia interna**. Si una medida se construye a partir de muchas partes diferentes que realizan funciones similares (p. ej., el resultado de un cuestionario de personalidad se suma a varias preguntas), ¿las partes individuales tienden a dar respuestas similares? Veremos esta forma particular de fiabilidad más adelante en el libro, en la sección sobre [Análisis de fiabilidad de consistencia interna].

No todas las mediciones necesitan poseer todas las formas de fiabilidad. Por ejemplo, la evaluación educativa puede considerarse como una forma de medición. Una de las materias que enseño, *Ciencia Cognitiva Computacional*, tiene una estructura de evaluación que tiene un componente de investigación y un componente de examen (además de otras cosas). El componente del examen está *destinado* a medir algo diferente del componente de investigación, por lo que la evaluación en su conjunto tiene una consistencia interna baja. Sin embargo, dentro del examen hay varias preguntas que pretenden (aproximadamente) medir las mismas cosas, y tienden a producir resultados similares. Entonces, el examen por sí solo tiene una consistencia interna bastante alta. Lo que es como debería ser. ¡Solo debes exigir fiabilidad en aquellas situaciones en las que deseas medir lo mismo!

2.4 El "rol" de las variables: predictores y resultados

Tengo una última terminología que explicarte antes de pasar a las variables. Normalmente, cuando investigamos, acabamos teniendo muchas variables diferentes. Después, cuando analizamos los datos, solemos intentar explicar algunas de las variables en función de otras variables. Es importante distinguir entre "lo que explica" y "lo que se explica". Así que seamos claros al respecto. En primer lugar, es mejor que nos acostumbremos a la idea de usar símbolos matemáticos para describir variables, ya que sucederá una y otra vez. Denotemos la variable "a ser explicada" Y, y las variables "que explican" como X_1, X_2, etc.

Cuando realizamos un análisis, tenemos diferentes nombres para X y Y, ya que desempeñan diferentes roles en el análisis. Los nombres clásicos para estos roles son **variable independiente** (VI) y **variable dependiente** (VD). La VI es la variable que se utiliza para hacer la explicación (es decir, X) y la VD es la variable que se explica (es decir, Y). La lógica detrás de estos nombres es la siguiente: si realmente existe una relación entre X y Y, entonces podemos decir que Y depende de X, y si hemos diseñado nuestro estudio "adecuadamente", entonces $X $ no depende de nada más. Sin embargo, personalmente encuentro esos nombres horribles. Son difíciles de recordar y muy engañosos porque (a) la VI nunca es realmente "independiente de todo lo demás", y (b) si no hay relación, entonces la VD en realidad no depende de la VI. Y, de hecho, como no soy la única persona que piensa que VI y VD son nombres horribles, hay una serie de alternativas que me parecen más atractivas. Los términos que usaré en este libro son **predictores** y **resultados**. La idea es que lo que se intenta es usar X (los predictores) para hacer conjeturas sobre Y (los resultados).[4] Esto se resume en Table 2.5.

Table 2.5: Distinciones de variables

role of the variable	**classical name**	**modern name**
"to be explained"	dependent variable (DV)	outcome
"to do the explaining"	independent variable (IV)	predictor

2.5 Investigación experimental y no experimental

Una de las grandes distinciones que debes conocer es la que existe entre "investigación experimental" e "investigación no experimental". Cuando hacemos esta distinción, de lo que realmente estamos hablando es del grado de control que el investigador ejerce sobre las personas y los acontecimientos del estudio.

2.5.1 Investigación experimental

La característica clave de la **investigación experimental** es que el investigador controla todos los aspectos del estudio, especialmente lo que experimentan los participantes

[4] Sin embargo, hay muchos nombres diferentes que se utilizan. No voy a enumerarlos todos (no tendría sentido hacerlo), salvo señalar que a veces se usa "variable de respuesta" donde he usado "resultado". Este tipo de confusión terminológica es muy común, me temo.

durante el mismo. En particular, el investigador manipula o varía las variables predictoras (VI) pero deja que la variable de resultado (VD) varíe de forma natural. La idea es variar deliberadamente los predictores (VI) para ver si tienen algún efecto causal sobre los resultados. Además, para garantizar que no haya ninguna posibilidad de que algo distinto de las variables predictoras esté causando los resultados, todo lo demás se mantiene constante o se "equilibra" de alguna otra forma, para garantizar que no tengan ningún efecto en los resultados. En la práctica, es casi imposible *pensar* en todo lo demás que pueda influir en el resultado de un experimento, y mucho menos mantenerlo constante. La solución estándar es la **aleatorización**. Es decir, asignamos aleatoriamente a las personas a diferentes grupos y luego le damos a cada grupo un tratamiento diferente (es decir, les asignamos diferentes valores de las variables predictoras). Hablaremos más sobre la aleatorización más adelante, pero por ahora basta con decir que lo que hace la aleatorización es minimizar (pero no eliminar) la posibilidad de que haya diferencias sistemáticas entre los grupos.

Veamos un ejemplo muy sencillo, completamente irreal y muy poco ético. Supongamos que queremos averiguar si fumar provoca cáncer de pulmón. Una forma de hacerlo sería buscar personas que fumen y personas que no fumen y ver si los fumadores tienen una tasa más alta de cáncer de pulmón. Esto *no* es un experimento propiamente dicho, ya que el investigador no tiene mucho control sobre quién es fumador y quién no. Y esto es realmente importante. Por ejemplo, podría ser que las personas que eligen fumar cigarrillos también tiendan a tener una dieta pobre, o tal vez tiendan a trabajar en minas de amianto, o lo que sea. La cuestión es que los grupos (fumadores y no fumadores) difieren en muchas cosas, no solo en el hábito de fumar. Por lo tanto, es posible que la mayor incidencia de cáncer de pulmón entre los fumadores se deba a otra cosa, y no al tabaquismo per se. En términos técnicos, estos otros factores (por ejemplo, la dieta) se denominan "factores de confusión", y hablaremos de ellos en un momento.

Mientras tanto, veamos cómo sería un experimento adecuado. Recordemos que nuestra preocupación era que los fumadores y los no fumadores podrían diferir en muchos aspectos. La solución, siempre que no tengas ética, es controlar quién fuma y quién no. En concreto, si dividimos aleatoriamente a los jóvenes no fumadores en dos grupos y obligamos a la mitad de ellos a convertirse en fumadores, es muy poco probable que los grupos difieran en algún aspecto que no sea el hecho de que la mitad fuma. De esa manera, si nuestro grupo de fumadores contrae cáncer en mayor proporción que el grupo de no fumadores, podemos estar bastante seguras de que (a) fumar sí causa cáncer y (b) somos asesinos.

2.5.2 Investigación no experimental

Investigación no experimental es un término amplio que abarca "cualquier estudio en el que el investigador no tiene tanto control como en un experimento". Obviamente, el control es algo que a los científicos les gusta tener, pero como ilustra el ejemplo anterior, hay muchas situaciones en las que no se puede o no se debe intentar obtener ese control. Dado que es muy poco ético (y casi con toda seguridad criminal) obligar a la gente a fumar para averiguar si contraen cáncer, este es un buen ejemplo de una situación en la que realmente no se debería intentar obtener un control experimental. Pero también hay otras razones. Incluso dejando de lado las cuestiones éticas, nuestro "experimento de fumar" tiene algunos otros problemas. Por ejemplo, cuando sugerí que "obliguemos" a la mitad de las personas a convertirse en fumadores, me refería a *comenzar* con una muestra de no fumadores y luego obligarlos a convertirse en fumadores. Aunque esto suena como

el tipo de diseño experimental sólido y malvado que le encantaría a un científico loco, podría no ser una forma muy sólida de investigar el efecto en el mundo real. Por ejemplo, supongamos que fumar solo causa cáncer de pulmón cuando las personas tienen dietas deficientes, y supongamos también que las personas que normalmente fuman tienden a tener dietas deficientes. Sin embargo, dado que los "fumadores" en nuestro experimento no son fumadores "naturales" (es decir, obligamos a los no fumadores a convertirse en fumadores, pero no adoptaron todas las demás características normales de la vida real que los fumadores tienden a tener) probablemente tengan mejores dietas. Como tal, en este ejemplo tonto no tendrían cáncer de pulmón y nuestro experimento fallaría, porque viola la estructura del mundo "natural" (el nombre técnico para esto es un "artefacto").

Una distinción que vale la pena hacer entre dos tipos de investigación no experimental es la diferencia entre **investigación cuasi-experimental** y **estudios de casos**. El ejemplo que mencioné anteriormente, en el que queríamos examinar la incidencia de cáncer de pulmón entre fumadores y no fumadores sin intentar controlar quién fuma y quién no, es un diseño cuasi-experimental. Es decir, es lo mismo que un experimento pero no controlamos los predictores (VIs). Podemos seguir utilizando la estadística para analizar los resultados, pero tenemos que ser mucho más cuidadosos y circunspectos.

El enfoque alternativo, los estudios de casos, pretende ofrecer una descripción muy detallada de uno o unos pocos casos. En general, no se puede usar la estadística para analizar los resultados de los estudios de casos y suele ser muy difícil sacar conclusiones generales sobre "la gente en general" a partir de unos pocos ejemplos aislados. Sin embargo, los estudios de casos son muy útiles en algunas situaciones. En primer lugar, hay situaciones en las que no se tiene otra alternativa. La neuropsicología se enfrenta mucho a este problema. A veces, simplemente no se puede encontrar a mucha gente con daño cerebral en un área específica del cerebro, así que lo único que se puede hacer es describir los casos que sí se tienen con tanto detalle y cuidado como sea posible. Sin embargo, los estudios de casos también tienen sus ventajas. Al no tener que estudiar a tanta gente, se puede invertir mucho tiempo y esfuerzo en comprender los factores específicos de cada caso. Esto es algo muy valioso. En consecuencia, los estudios de casos pueden complementar los enfoques más orientados a la estadística que se ven en los diseños experimentales y cuasi-experimentales. En este libro no hablaremos mucho de los estudios de casos, pero sin embargo son herramientas muy valiosas.

2.6 Evaluar la validez de un estudio

Más que cualquier otra cosa, un científico quiere que su investigación sea "válida". La idea conceptual detrás de **validez** es muy simple. ¿Puedes confiar en los resultados de tu estudio? Si no, el estudio no es válido. Sin embargo, si bien es fácil de establecer, en la práctica es mucho más difícil verificar la validez que verificar la fiabilidad. Y con toda honestidad, no existe una noción precisa y claramente acordada de lo que realmente es la validez. De hecho, hay muchos tipos diferentes de validez, cada uno de los cuales plantea sus propios problemas. Y no todas las formas de validez son relevantes para todos los estudios. Voy a hablar de cinco tipos diferentes de validez:

- Validez interna
- Validez externa
- Validez de constructo
- Validez aparente

- Validez ecológica

Primero, una guía rápida sobre lo que importa aquí. (1) La validez interna y externa son las más importantes, ya que se relacionan directamente con la pregunta fundamental de si tu estudio realmente funciona. (2) La validez de constructo pregunta si estás midiendo lo que crees que estás midiendo. (3) La validez aparente no es demasiado importante, excepto en la medida en que te preocupes por las "apariencias". (4) La validez ecológica es un caso especial de validez aparente que corresponde a un tipo de apariencia que podría interesarte mucho.

2.6.1 Validez interna

Validez interna se refiere a la medida en que puedes sacar las conclusiones correctas sobre las relaciones causales entre las variables. Se llama "interna" porque se refiere a las relaciones entre las cosas "dentro" del estudio. Ilustremos el concepto con un ejemplo sencillo. Imagina que estás interesada en averiguar si una educación universitaria te permite escribir mejor. Para hacerlo, reúnes a un grupo de estudiantes de primer año, les pides que escriban un ensayo de 1000 palabras y cuentas la cantidad de errores ortográficos y gramaticales que cometen. Luego encuentras algunos estudiantes de tercer año, que obviamente han tenido más educación universitaria que los de primer año, y repites el ejercicio. Y supongamos que resulta que los estudiantes de tercer año cometen menos errores. Y entonces concluyes que una educación universitaria mejora las habilidades de escritura. ¿Correcto? Excepto que el gran problema de este experimento es que los estudiantes de tercer año son mayores y tienen más experiencia escribiendo cosas. Así que es difícil saber con certeza cuál es la relación causal. ¿Las personas mayores escriben mejor? ¿O personas que han tenido más experiencia escribiendo? ¿O personas que han tenido más educación? ¿Cuál de las anteriores es la verdadera causa del desempeño superior de los de tercer año? ¿Edad? ¿Experiencia? ¿Educación? No puedes saberlo. Este es un ejemplo de un fallo de validez interna, porque tu estudio no separa adecuadamente las relaciones causales entre las diferentes variables.

2.6.2 Validez externa

La validez externa se relaciona con la **generalizabilidad** o la **aplicabilidad** de tus hallazgos. Es decir, en qué medida esperas ver en la "vida real" el mismo patrón de resultados que viste en tu estudio. Para decirlo con un poco más de precisión, cualquier estudio que realices en psicología implicará un conjunto bastante específico de preguntas o tareas, ocurrirá en un entorno específico e involucrará a participantes que provienen de un subgrupo particular (lamentablemente, a menudo es alumnado universitario). Entonces, si resulta que los resultados en realidad no se generalizan ni se aplican a personas y situaciones más allá de las que estudiaste, lo que tienes es una falta de validez externa.

El ejemplo clásico de este problema es el hecho de que una gran proporción de los estudios de psicología utilizarán como participantes a estudiantes universitarios de psicología. Obviamente, sin embargo, los investigadores no se preocupan *solo* por el estudiantado de psicología. Se preocupan por la gente en general. Por ello, un estudio que utiliza como participantes únicamente a estudiantes de psicología siempre conlleva el riesgo de carecer de validez externa. Es decir, si hay algo "especial" en los estudiantes de psicología que los diferencia de la población general en algún aspecto relevante, entonces podemos comenzar a preocuparnos por la falta de validez externa.

Dicho esto, es absolutamente crítico darse cuenta de que un estudio que utiliza solo estudiantes de psicología no necesariamente tiene un problema con la validez externa. Volveré a hablar de esto más adelante, pero es un error tan común que lo mencionaré aquí. La validez externa de un estudio se ve amenazada por la elección de la población si (a) la población de la que tomas muestras de sus participantes es muy reducida (por ejemplo, estudiantes de psicología), y (b) la población reducida de la que tomas muestras es sistemáticamente diferente de la población general en algún aspecto que sea relevante para el *fenómeno psicológico que pretendes estudiar*. La parte en cursiva es la parte que mucha gente olvida. Es cierto que el alumnado de psicología difiere de la población general en muchos aspectos, por lo que un estudio que utilice solo estudiantes de psicología puede tener problemas con la validez externa. Sin embargo, si esas diferencias no son muy relevantes para el fenómeno que estás estudiando, entonces no hay de qué preocuparse. Para hacer esto un poco más concreto, aquí hay dos ejemplos extremos:

- Quieres medir las "actitudes del público en general hacia la psicoterapia", pero todos tus participantes son estudiantes de psicología. Es casi seguro que este estudio tendrá un problema con la validez externa.
- Quieres medir la efectividad de una ilusión visual y tus participantes son todos estudiantes de psicología. Es poco probable que este estudio tenga un problema con la validez externa.

Habiendo pasado los últimos dos párrafos centrándonos en la elección de los participantes, dado que es un tema importante que tiende a preocupar más a todos, vale la pena recordar que la validez externa es un concepto más amplio. Los siguientes también son ejemplos de cosas que podrían representar una amenaza para la validez externa, según el tipo de estudio que estés realizando:

- Las personas pueden responder un "cuestionario de psicología" de una manera que no refleja lo que harían en la vida real.
- Tu experimento de laboratorio sobre (digamos) "aprendizaje humano" tiene una estructura diferente a los problemas de aprendizaje que afrontan las personas en la vida real.

2.6.3 Validez de constructo

La validez de constructo es básicamente una cuestión de si estás midiendo lo que quieres medir. Una medida tiene una buena validez de constructo si en realidad mide el constructo teórico correcto y una mala validez de constructo si no lo hace. Para dar un ejemplo muy simple (aunque ridículo), supongamos que estoy tratando de investigar las tasas con las que el alumnado universitario hacen trampa en sus exámenes. Y la forma en que intento medirlo es pidiendo al alumnado que hace trampa que se ponga de pie en la sala de conferencias para que pueda contarlos. Cuando hago esto con una clase de 300 estudiantes, 0 personas afirman ser tramposos. Por lo tanto, concluyo que la proporción de tramposos en mi clase es del 0%. Claramente esto es un poco ridículo. Pero lo importante aquí no es que este sea un ejemplo metodológico muy profundo, sino más bien explicar qué es la validez de constructo. El problema con mi medida es que mientras trato de medir "la proporción de personas que hacen trampa", lo que en realidad estoy midiendo es "la proporción de personas lo suficientemente estúpidas como para reconocer que hacen trampa, o lo suficientemente estúpidas como para fingir que las hacen". ¡Obviamente, no es lo mismo! Así que mi estudio salió mal, porque mi

medida tiene una validez de constructo muy pobre.

2.6.4 Validez aparente

Validez aparente simplemente se refiere a si una medida "parece" que está haciendo lo que se supone que debe hacer, nada más. Si diseño una prueba de inteligencia, y la gente la mira y dice "no, esa prueba no mide la inteligencia", entonces la medida carece de validez aparente. Es tan simple como eso. Obviamente, la validez aparente no es muy importante desde una perspectiva puramente científica. Después de todo, lo que nos importa es si la medida *realmente* hace o no lo que se supone que debe hacer, no si *parece* que hace lo que se supone que debe hacer. Como consecuencia, generalmente no nos importa mucho la validez aparente. Dicho esto, el concepto de validez aparente tiene tres propósitos pragmáticos útiles:

- A veces, un científico experimentado tendrá la "corazonada" de que una medida en particular no funcionará. Si bien este tipo de corazonadas no tienen un valor probatorio estricto, a menudo vale la pena prestarles atención. Porque muchas veces las personas tienen conocimientos que no pueden verbalizar, por lo que puede haber algo de qué preocuparse, incluso si no puedes decir por qué. En otras palabras, cuando alguien de tu confianza critica la validez aparente de tu estudio, vale la pena tomarse el tiempo para pensar más detenidamente en tu diseño para ver si puedes pensar en las razones por las que podría salir mal. Eso sí, si no encuentras ningún motivo de preocupación, entonces probablemente no deberías preocuparte. Después de todo, la validez aparente realmente no importa mucho.
- A menudo (muy a menudo), las personas completamente desinformadas también tendrán la "corazonada" de que tu investigación es una porquería. Y lo criticarán en Internet o algo así. Si lo examinas detenidamente, puedes notar que estas críticas en realidad se centran por completo en cómo "se ve" el estudio, pero no en nada más profundo. El concepto de validez aparente es útil para explicar con delicadeza a las personas que necesitan fundamentar más sus argumentos.
- Ampliando el último punto, si las creencias de las personas no capacitadas son críticas (p. ej., este suele ser el caso de la investigación aplicada en la que realmente se quiere convencer a los responsables políticos de una cosa u otra), entonces hay que preocuparse por la validez aparente. Simplemente porque, te guste o no, mucha gente usará la validez aparente como un indicador de la validez real. Si quieres que el gobierno cambie una ley por razones psicológicas científicas, entonces no importará cuán buenos sean "realmente" tus estudios. Si carecen de validez aparente, encontrarás que los políticos lo ignoran. Por supuesto, es algo injusto que la política a menudo dependa más de la apariencia que de los hechos, pero así es como funcionan las cosas.

2.6.5 Validez ecológica

Validez ecológica es una noción diferente de validez, que es similar a la validez externa, pero menos importante. La idea es que, para que sea ecológicamente válido, toda la configuración del estudio debe aproximarse mucho al escenario del mundo real que se está investigando. En cierto sentido, la validez ecológica es una especie de validez aparente. Se relaciona principalmente con si el estudio "parece" correcto, pero con un poco más de rigor. Para ser ecológicamente válido, el estudio tiene que verse bien de una manera bastante específica. La idea detrás de esto es la intuición de que un estudio

que es ecológicamente válido tiene más probabilidades de ser válido externamente. No es una garantía, por supuesto. Pero lo bueno de la validez ecológica es que es mucho más fácil verificar si un estudio es ecológicamente válido que verificar si un estudio es válido externamente. Un ejemplo simple serían los estudios de identificación de testigos presenciales. La mayoría de estos estudios tienden a realizarse en un entorno universitario, a menudo con una serie bastante simple de caras para mirar, en lugar de una fila. El tiempo que transcurre entre ver al "criminal" y pedirle que identifique al sospechoso en la "fila" suele ser más corto. El "crimen" no es real, por lo que no hay posibilidad de que el testigo se asuste, y no hay policías presentes, por lo que no hay tanta posibilidad de sentirse presionado. Todas estas cosas significan que el estudio carece de validez ecológica. Podría (o no) significar que también carece de validez externa.

2.7 Factores de confusión, artefactos y otras amenazas a la validez

Si analizamos el tema de la validez en general, las dos mayores preocupaciones que tenemos son los *factores de confusión* y los artefactos. Estos dos términos se definen de la siguiente manera:

- **Factor de confusión:** Un confusor es una variable adicional, a menudo no medida[5] que resulta estar relacionada tanto con los predictores como con el resultado. La existencia de factores de confusión amenaza la validez interna del estudio porque no se puede saber si el predictor causa el resultado o si la variable de confusión lo causa.
- **Artefacto:** Se dice que un resultado es "artefacto" si solo se mantiene en la situación especial que probaste en tu estudio. La probabilidad de que tu resultado sea un artefacto describe una amenaza a su validez externa, porque plantea la posibilidad de que no puedas generalizar o aplicar tus resultados a la población real que te interesa.

Como regla general, los factores de confusión son una gran preocupación para los estudios no experimentales, precisamente porque no son experimentos adecuados. Por definición, se dejan muchas cosas sin controlar, por lo que hay mucha probabilidad de que los factores de confusión estén presentes en tu estudio. La investigación experimental tiende a ser mucho menos vulnerable a los factores de confusión. Cuanto más control tengas sobre lo que sucede durante el estudio, más podrás evitar que los factores de confusión afecten los resultados. Con la asignación aleatoria, por ejemplo, los factores de confusión se distribuyen de manera aleatoria y uniforme entre diferentes grupos.

Sin embargo, siempre hay ventajas y desventajas y cuando comenzamos a pensar en artefactos en lugar de factores de confusión, la situación es la contraria. En su mayor parte, los resultados de artefactos son una preocupación para los estudios experimentales más que para los estudios no experimentales. Para ver esto, es útil darse cuenta de que la razón por la que muchos estudios no son experimentales es precisamente porque lo que el investigador intenta hacer es examinar el comportamiento humano en un contexto

[5] la razón por la que digo que no se mide es que si lo has medido, puedes usar algunos trucos estadísticos sofisticados para lidiar con el factor de confusión. Debido a la existencia de estas soluciones estadísticas al problema de los factores de confusión, a menudo nos referimos a un factor de confusión que hemos medido y tratado como una covariable. Tratar con covariables es un tema más avanzado, pero pensé en mencionarlo de pasada ya que es un poco reconfortante saber al menos que esto existe.

más naturalista. Al trabajar en un contexto más real, pierde el control experimental (haciéndose vulnerable a los factores de confusión), pero debido a que estudia psicología humana "en el contexto natural", reduce la probabilidad de obtener un artefacto. O, para decirlo de otra manera, cuando sacas la psicología del contexto natural y la llevas al laboratorio (lo que generalmente tenemos que hacer para obtener nuestro control experimental), siempre corres el riesgo de estudiar accidentalmente algo diferente a lo que querías estudiar.

Sin embargo, ten cuidado. Lo anterior es solo una guía aproximada. Es absolutamente posible tener factores de confusión en un experimento y obtener resultados artefactos con estudios no experimentales. Esto puede ocurrir por varias razones, una de las cuales es un error del experimentador o del investigador. En la práctica, es realmente difícil pensar en todo antes de tiempo e incluso los mejores investigadores cometen errores.

Aunque hay un sentido en el que casi cualquier amenaza a la validez puede caracterizarse como un factor de confusión o un artefacto, son conceptos bastante vagos. Así que echemos un vistazo a algunos de los ejemplos más comunes.

2.7.1 Efectos de la historia

Los efectos de la historia se refieren a la posibilidad de que ocurran eventos específicos durante el estudio que puedan influir en la medida del resultado. Por ejemplo, algo podría suceder entre una prueba previa y una prueba posterior. O entre las pruebas del participante 23 y el participante 24. Alternativamente, podría ser que estés viendo un artículo de un estudio anterior que era perfectamente válido para su época, pero el mundo ha cambiado lo suficiente desde entonces y las conclusiones ya no son fiables. Ejemplos de cosas que contarían como efectos de historia son:

- Te interesa cómo piensa la gente sobre el riesgo y la incertidumbre. Empezaste la recopilación de datos en diciembre de 2010. Pero encontrar participantes y recopilar datos lleva tiempo, por lo que todavía estás encontrando nuevas personas en febrero de 2011. Desafortunadamente para ti (y aún más lamentablemente para otros), las inundaciones de Queensland ocurrieron en enero de 2011 y causaron miles de millones de dólares en daños y mataron a muchas personas. No es sorprendente que las personas evaluadas en febrero de 2011 expresen creencias bastante diferentes sobre el manejo del riesgo que las personas evaluadas en diciembre de 2010. ¿Cuál (si alguna) de estas refleja las creencias "verdaderas" de los participantes? Creo que la respuesta es probablemente ambas. Las inundaciones de Queensland cambiaron genuinamente las creencias del público australiano, aunque posiblemente solo temporalmente. La clave aquí es que la "historia" de las personas evaluadas en febrero es bastante diferente a la de las personas evaluadas en diciembre.

- Estás probando los efectos psicológicos de un nuevo medicamento contra la ansiedad. Entonces lo que haces es medir la ansiedad antes de administrar el fármaco (por ejemplo, por autoinforme y tomando medidas fisiológicas). Luego administras la droga y luego tomas las mismas medidas. Sin embargo, en el medio, debido a que tu laboratorio está en Los Ángeles, hay un terremoto que aumenta la ansiedad de los participantes.

2.7.2 Efectos de maduración

Al igual que con los efectos de la historia, los **efectos de maduración** tienen que ver fundamentalmente con el cambio a lo largo del tiempo. Sin embargo, los efectos de maduración no responden a eventos específicos. Más bien, se relacionan con cómo las personas cambian por sí mismas con el tiempo. Nos hacemos mayores, nos cansamos, nos aburrimos, etc. Algunos ejemplos de efectos de maduración son:

- Al realizar una investigación de psicología del desarrollo, debes tener en cuenta que los niños crecen con bastante rapidez. Entonces, supongamos que deseas averiguar si algún truco educativo ayuda con el tamaño del vocabulario entre los niños de 3 años. Una cosa que debes tener en cuenta es que el tamaño del vocabulario de los niños de esa edad está creciendo a un ritmo increíble (varias palabras por día) por sí solo. Si diseñas tu estudio sin tener en cuenta este efecto de maduración, entonces no podrás saber si tu truco educativo funciona.
- Cuando se ejecuta un experimento muy largo en el laboratorio (por ejemplo, algo que dure 3 horas), es muy probable que las personas comiencen a aburrirse y cansarse, y que este efecto madurativo provoque una disminución del rendimiento independientemente de cualquier otra cosa que suceda en el experimento

2.7.3 Efectos de las pruebas repetidas

Un tipo importante de efecto de la historia es el efecto de las **pruebas repetidas**. Supongamos que quiero tomar dos medidas de algún constructo psicológico (p. ej., ansiedad). Una cosa que podría preocuparme es si la primera medición tiene un efecto en la segunda medición. En otras palabras, ¡este es un efecto histórico en el que el "evento" que influye en la segunda medición es la primera medición en sí misma! Esto no es nada raro. Ejemplos de esto incluyen:

- Aprendizaje y práctica: por ejemplo, la "inteligencia" en el tiempo 2 podría parecer que aumenta en relación con el tiempo 1 porque los participantes aprendieron las reglas generales de cómo resolver preguntas del tipo "test de inteligencia" durante la primera sesión de pruebas.
- Familiaridad con la situación de la prueba: por ejemplo, si las personas están nerviosas en el momento 1, esto podría hacer que el rendimiento baje. Pero después de pasar por la primera situación de prueba, es posible que se calmen mucho precisamente porque han visto cómo es la prueba.
- Cambios auxiliares causados por las pruebas: por ejemplo, si un cuestionario que evalúa el estado de ánimo es aburrido, es más probable que la calificación del estado de ánimo en el tiempo de medición 2 sea "aburrida" precisamente por la medición aburrida realizada en el tiempo 1.

2.7.4 Sesgo de selección

Sesgo de selección es un término bastante amplio. Imagina que estás realizando un experimento con dos grupos de participantes en el que cada grupo recibe un "tratamiento" diferente y deseas ver si los diferentes tratamientos conducen a resultados diferentes. Sin embargo, supongamos que, a pesar de tus mejores esfuerzos, has terminado con un desequilibrio de género entre los grupos (por ejemplo, el grupo A tiene un 80 % de mujeres y el grupo B tiene un 50 % de mujeres). Puede parecer que esto nunca podría suceder, pero créeme, puede suceder. Este es un ejemplo de un sesgo de selección, en el que las

personas "seleccionadas en" los dos grupos tienen características diferentes. Si alguna de esas características resulta ser relevante (por ejemplo, tu tratamiento funciona mejor en mujeres que en hombres), entonces tienes un gran problema.

2.7.5 Abandono diferencial

Al pensar en los efectos del abandono, a veces es útil distinguir entre dos tipos diferentes. El primero es el **abandono homogéneo**, en el que el efecto del abandono es el mismo para todos los grupos, tratamientos o condiciones. En el ejemplo que di arriba, el abandono sería homogéneo si (y solo si) los participantes que se aburren fácilmente abandonan todas las condiciones de mi experimento aproximadamente al mismo ritmo. En general, es probable que el principal efecto del abandono homogéneo sea que hace que tu muestra no sea representativa. Como tal, la mayor preocupación que tendrás es que la generalización de los resultados disminuya. En otras palabras, pierde validez externa.

El segundo tipo de abandono es el **abandono heterogéneo**, en el que el efecto de abandono es diferente para diferentes grupos. Más a menudo llamado **abandono diferencial**, este es un tipo de sesgo de selección causado por el propio estudio. Supongamos que, por primera vez en la historia de la psicología, consigo encontrar la muestra de personas perfectamente equilibrada y representativa. Comienzo a ejecutar el "experimento increíblemente largo y tedioso de Dani" en mi muestra perfecta, pero luego, debido a que mi estudio es increíblemente largo y tedioso, muchas personas comienzan a abandonar. No puedo detener esto. Los participantes tienen absolutamente el derecho de dejar de hacer cualquier experimento, en cualquier momento, por cualquier motivo que deseen, y como investigadores estamos moralmente (y profesionalmente) obligados a recordar a las personas que tienen este derecho. Entonces, supongamos que el "experimento increíblemente largo y tedioso de Dani" tiene una tasa de abandono muy alta. ¿Cuáles crees que son las probabilidades de que este abandono sea aleatorio? Respuesta: cero. Es casi seguro que las personas que se quedan son más concienzudas, más tolerantes con el aburrimiento, etc., que las que se van. En la medida en que (digamos) la escrupulosidad sea relevante para el fenómeno psicológico que me importa, este abandono puede disminuir la validez de mis resultados.

Aquí hay otro ejemplo. Supongamos que diseño mi experimento con dos condiciones. En la condición de "tratamiento", el experimentador insulta al participante y luego le entrega un cuestionario diseñado para medir la obediencia. En la condición de "control", el experimentador se involucra en una charla sin sentido y luego les entrega el cuestionario. Dejando de lado los méritos científicos cuestionables y la ética dudosa de tal estudio, pensemos qué podría salir mal aquí. Como regla general, cuando alguien me insulta en la cara tiendo a cooperar mucho menos. Por lo tanto, hay muchas posibilidades de que muchas más personas abandonen la condición de tratamiento que la condición de control. Y este abandono no va a ser aleatorio. Las personas con más probabilidades de abandonar probablemente serían las personas a las que no les importa demasiado la importancia de permanecer obedientemente durante el experimento. Dado que las personas más malintencionadas y desobedientes abandonaron el grupo de tratamiento pero no el grupo de control, hemos introducido una confusión: las personas que realmente respondieron el cuestionario en el grupo de tratamiento ya eran más obedientes y cumplidoras que las personas en el grupo de control. En resumen, en este estudio insultar a las personas no las hace más obedientes. ¡Hace que las personas más desobedientes abandonen el experimento! La validez interna de este experimento está

completamente descartada.

2.7.6 Sesgo de no respuesta

El sesgo por falta de respuesta está estrechamente relacionado con el sesgo de selección y con el abandono diferencial. La versión más simple del problema es así. Envías una encuesta a 1000 personas, pero solo 300 de ellas responden. Es casi seguro que las 300 personas que respondieron no son una submuestra aleatoria. Las personas que responden a las encuestas son sistemáticamente diferentes a las personas que no lo hacen. Esto presenta un problema al tratar de generalizar a partir de esas 300 personas que respondieron a la población general, ya que ahora tienes una muestra no aleatoria. Sin embargo, el problema del sesgo por falta de respuesta es más general que esto. Entre las (digamos) 300 personas que respondieron a la encuesta, es posible que no todos respondan todas las preguntas. Si (digamos) 80 personas optaron por no responder a una de tus preguntas, ¿presenta esto problemas? Como siempre, la respuesta es quizás. Si la pregunta que no se contestó estaba en la última página del cuestionario y esas 80 encuestas se devolvieron sin la última página, es muy probable que los datos que faltan no sean un gran problema; probablemente las páginas simplemente se cayeron. Sin embargo, si la pregunta que 80 personas no respondieron fue la pregunta personal más conflictiva o invasiva del cuestionario, es casi seguro que tienes un problema. En esencia, se trata de lo que se denomina el problema de **datos faltantes**. Si los datos que faltan se "perdieron" al azar, entonces no es un gran problema. Si falta sistemáticamente, puede ser un gran problema.

2.7.7 Regresión a la media

La regresión a la media hace referencia a cualquier situación en la que selecciones datos en función de un valor extremo en alguna medida. Debido a que la variable tiene una variación natural, es casi seguro que significa que cuando tomas una medición posterior, la última medición será menos extrema que la primera, puramente por casualidad.

Aquí hay un ejemplo. Supongamos que me interesa saber si la educación en psicología tiene un efecto adverso en los chicos y chicas muy inteligentes. Para ello, busco a los 20 estudiantes de psicología I con las mejores notas de bachillerato y observo qué tal les va en la universidad. Resulta que les está yendo mucho mejor que el promedio, pero no son los mejores de la clase en la universidad a pesar de que sí fueron los mejores en bachillerato. ¿Que esta pasando? El primer pensamiento natural es que esto debe significar que las clases de psicología deben tener un efecto adverso en esos estudiantes. Sin embargo, si bien esa podría ser la explicación, es más probable que lo que estás viendo sea un ejemplo de "regresión a la media". Para ver cómo funciona, pensemos por un momento qué se requiere para obtener la mejor calificación en una clase, sin importar si esa clase es en bachillerato o en la universidad. Cuando tienes una clase grande, habrá muchas personas muy inteligentes inscritas. Para sacar la mejor nota tienes que ser muy inteligente, trabajar muy duro y tener un poco de suerte. El examen tiene que hacer las preguntas correctas para tus habilidades idiosincrásicas, y tienes que evitar cometer errores tontos (todos lo hacemos a veces) al responderlas. Y esa es la cuestión, mientras que la inteligencia y el trabajo duro son transferibles de una clase a otra, la suerte no lo es. Las personas que tuvieron suerte en la escuela secundaria no serán las mismas que las que tuvieron suerte en la universidad. Esa es la definición misma de "suerte". La

consecuencia de esto es que cuando seleccionas personas en los valores extremos de una medición (los 20 mejores estudiantes), estás seleccionando por trabajo duro, habilidad y suerte. Pero debido a que la suerte no se transfiere a la segunda medición (solo la habilidad y el trabajo), se espera que todas estas personas bajen un poco cuando las midas por segunda vez (en la universidad). Entonces sus puntuaciones retroceden un poco, hacia todos los demás. Esta es la regresión a la media.

La regresión a la media es sorprendentemente común. Por ejemplo, si dos personas muy altas tienen hijos, sus hijos tenderán a ser más altos que el promedio pero no tan altos como los padres. Lo contrario sucede con los padres muy bajos. Dos padres muy bajos tenderán a tener hijos pequeños, pero sin embargo esos niños tenderán a ser más altos que los padres. También puede ser extremadamente sutil. Por ejemplo, se han realizado estudios que sugieren que las personas aprenden mejor con comentarios negativos que con comentarios positivos. Sin embargo, la forma en que las personas intentaron mostrar esto fue dándoles un refuerzo positivo cada vez que lo hacían bien y un refuerzo negativo cuando lo hacían mal. Y lo que se ve es que después del refuerzo positivo la gente tendía a hacerlo peor, pero después del refuerzo negativo tendía a hacerlo mejor. ¡Pero fíjate que aquí hay un sesgo de selección! Cuando a las personas les va muy bien, estás seleccionando valores "altos", por lo que debes esperar, debido a la regresión a la media, que el rendimiento en la siguiente prueba sea peor, independientemente de si se da refuerzo. De manera similar, después de una mala prueba, las personas tenderán a mejorar por sí mismas. La aparente superioridad de la retroalimentación negativa es un artefacto causado por la regresión a la media (ver Kahneman & Tversky (1973), para discusión).

2.7.8 Sesgo del experimentador

El sesgo del experimentador puede presentarse de múltiples formas. La idea básica es que el experimentador, a pesar de sus mejores intenciones, puede terminar influenciando accidentalmente los resultados del experimento al comunicar sutilmente la "respuesta correcta" o el "comportamiento deseado" a los participantes. Por lo general, esto ocurre porque el experimentador tiene un conocimiento especial que el participante no tiene, por ejemplo, la respuesta correcta a las preguntas que se le hacen o el conocimiento del patrón de desempeño esperado para la condición en la que se encuentra el participante. El ejemplo clásico de esto es el caso de estudio de "Clever Hans", que data de 1907 (Pfungst, 1911). Clever Hans era un caballo que aparentemente podía leer y contar y realizar otras hazañas de inteligencia similares a las de los humanos. Después de que Clever Hans se hiciera famoso, los psicólogos comenzaron a examinar su comportamiento más de cerca. Resultó que, como era de esperar, Hans no sabía hacer matemáticas. Más bien, Hans estaba respondiendo a los observadores humanos que lo rodeaban, porque los humanos sí sabían contar y el caballo había aprendido a cambiar su comportamiento cuando la gente cambiaba el suyo.

La solución general al problema del sesgo del experimentador es participar en estudios doble ciego, en los que ni el experimentador ni el participante saben en qué condición se encuentra el participante ni cuál es el comportamiento deseado. Esto proporciona una muy buena solución al problema, pero es importante reconocer que no es del todo ideal y difícil de lograr a la perfección. Por ejemplo, la forma obvia en la que podría intentar construir un estudio doble ciego es tener uno de mis estudiantes de doctorado (uno que no sabe nada sobre el experimento) dirigiendo el estudio. Eso parece que debería ser suficiente. La única persona (yo) que conoce todos los detalles (p. ej., las respuestas

correctas a las preguntas, las asignaciones de los participantes a las condiciones) no interactúa con los participantes, y la persona que habla con la gente (el estudiante de doctorado) no sabe nada. Excepto por la realidad de que es muy poco probable que la última parte sea cierta. Para que el estudiante de doctorado pueda llevar a cabo el estudio de manera efectiva, deben haber sido informados por mí, el investigador. Y, como sucede, el estudiante también me conoce y sabe un poco acerca de mis creencias generales sobre las personas y la psicología (p. ej., tiendo a pensar que los humanos son mucho más inteligentes de lo que los psicólogos piensan). Como resultado de todo esto, es casi imposible que el experimentador deje de saber un poco sobre las expectativas que tengo. E incluso un poco de conocimiento puede tener un efecto. Supongamos que el experimentador transmite accidentalmente el hecho de que se espera que los participantes lo hagan bien en esta tarea. Bueno, hay una cosa llamada "efecto Pigmalión", donde si esperas grandes cosas de las personas, tenderán a estar a la altura de las circunstancias. Pero si esperas que fracasen, también lo harán. En otras palabras, las expectativas se convierten en una profecía autocumplida.

2.7.9 Efectos de la demanda y reactividad

Cuando se habla del sesgo del experimentador, la preocupación es que el conocimiento o los deseos del experimentador para el experimento se comuniquen a los participantes, y que estos puedan cambiar el comportamiento de las personas (Rosenthal, 1966). Sin embargo, incluso si logras evitar que esto suceda, es casi imposible evitar que las personas sepan que son parte de un estudio psicológico. Y el mero hecho de saber que alguien te está mirando o estudiando puede tener un efecto bastante grande en el comportamiento. Esto generalmente se conoce como **reactividad** o **efectos de demanda**. La idea básica se recoge en el efecto Hawthorne: las personas alteran su rendimiento debido a la atención que les presta el estudio. El efecto toma su nombre de un estudio que tuvo lugar en la fábrica "Hawthorne Works" en las afueras de Chicago (ver Adair (1984)). Este estudio, de la década de 1920, analizó los efectos de la iluminación de las fábricas en la productividad de los trabajadores. Pero, lo que es más importante, el cambio en el comportamiento de los trabajadores ocurrió porque los trabajadores sabían que estaban siendo estudiados, en lugar de cualquier efecto de la iluminación de la fábrica.

Para concretar un poco más algunas de las formas en que el mero hecho de estar en un estudio puede cambiar el comportamiento de las personas, ayuda pensar como un psicólogo social y observar algunos de los roles que las personas pueden adoptar durante un experimento, pero podría no adoptar si los eventos correspondientes estuvieran ocurriendo en el mundo real:

- El *buen participante* trata de ser demasiado útil para el investigador. Él o ella busca descifrar las hipótesis del experimentador y confirmarlas.
- El *participante negativo* hace exactamente lo contrario del buen participante. Él o ella busca romper o destruir el estudio o la hipótesis de alguna manera.
- El *participante fiel* es anormalmente obediente. Él o ella busca seguir las instrucciones a la perfección, independientemente de lo que podría haber sucedido en un entorno más realista.
- El *participante aprensivo* se pone nervioso acerca de ser evaluado o estudiado, tanto que su comportamiento se vuelve muy antinatural o demasiado socialmente deseable.

2.7.10 Efectos placebo

El **efecto placebo** es un tipo específico de efecto de demanda que nos preocupa mucho. Se refiere a la situación en la que el mero hecho de ser tratado provoca una mejora en los resultados. El ejemplo clásico proviene de los ensayos clínicos. Si le das a la gente un medicamento completamente inerte químicamente y les dices que es una cura para una enfermedad, tenderán a mejorar más rápido que las personas que no reciben ningún tratamiento. En otras palabras, es la creencia de las personas de que están siendo tratadas lo que produce mejores resultados, no el medicamento.

Sin embargo, el consenso actual en medicina es que los verdaderos efectos placebo son bastante raros y que la mayor parte de lo que antes se consideraba efecto placebo es, de hecho, una combinación de curación natural (algunas personas simplemente mejoran por sí solas), regresión a la media y otras peculiaridades de diseño de estudio. De interés para la psicología es que la evidencia más sólida de al menos algún efecto placebo se encuentra en los resultados autoinformados, sobre todo en el tratamiento del dolor (Hróbjartsson & Gøtzsche, 2010).

2.7.11 Efectos de situación, medición y subpoblación

En algunos aspectos, estos términos son un término general para "todas las demás amenazas a la validez externa". Se refieren al hecho de que la elección de la subpoblación de la que extraes a tus participantes, la ubicación, el momento y la forma en que llevas a cabo tu estudio (incluido quién recopila los datos) y las herramientas que utilizas para realizar tus mediciones pueden estar influyendo en los resultados. Específicamente, la preocupación es que estas cosas puedan influir en los resultados de tal manera que los resultados no se generalicen a una gama más amplia de personas, lugares y medidas.

2.7.12 Fraude, engaño y autoengaño

> *Es difícil lograr que un hombre entienda algo, cuando su salario depende de que no lo entienda.*
> - Upton Sinclair

Hay una última cosa que siento que debo mencionar. Mientras leía lo que los libros de texto a menudo tienen que decir sobre la evaluación de la validez de un estudio, no pude evitar notar que parecen asumir que el investigador es honesto. Me parece divertidísimo. Si bien la gran mayoría de los científicos son honestos, al menos según mi experiencia, algunos no lo son.[6] No solo eso, como mencioné anteriormente, los científicos no son inmunes al sesgo de creencias. Es fácil para un investigador terminar engañándose a sí mismo creyendo algo incorrecto, y esto puede llevarlos a realizar una investigación sutilmente defectuosa y luego ocultar esos defectos cuando la escriben. Por lo tanto, debes considerar no solo la posibilidad (probablemente poco probable) de un fraude absoluto, sino también la posibilidad (probablemente bastante común) de que la investigación esté "sesgada" sin querer. Abrí algunos libros de texto estándar y no encontré mucha discusión sobre este problema, así que aquí está mi propio intento de enumerar algunas formas en que pueden surgir estos problemas:

[6] Algunas personas podrían argumentar que si no eres honesto, entonces no eres un verdadero científico. Supongo que tiene algo de verdad, pero eso es falso (busque la falacia "No hay verdadero escocés"). El hecho es que hay muchas personas que están empleadas ostensiblemente como científicos, y cuyo trabajo tiene todas las trampas de la ciencia, pero que son totalmente fraudulentas. Pretender que no existen diciendo que no son científicos es solo un pensamiento confuso.

- **Fabricación de datos**. A veces, las personas simplemente inventan los datos. Esto se hace ocasionalmente con "buenas" intenciones. Por ejemplo, el investigador cree que los datos fabricados reflejan la verdad y, de hecho, pueden reflejar versiones "ligeramente limpias" de los datos reales. En otras ocasiones, el fraude es deliberado y malicioso. Algunos ejemplos destacados de presunta o demostrada falsificación de datos incluyen a Cyril Burt (un psicólogo que se cree que fabricó algunos de sus datos), Andrew Wakefield (a quien se acusó de fabricar sus datos conectando la vacuna MMR con el autismo) y Hwang Woo-suk (quien falsificó muchos de sus datos sobre la investigación con células madre).
- **Bulos**. Los bulos comparten muchas similitudes con la fabricación de datos, pero difieren en el propósito que persiguen. Un bulo es a menudo una broma, y muchos de ellos están destinados a ser (eventualmente) descubiertos. A menudo, el objetivo de un engaño es desacreditar a alguien o algún campo. A lo largo de los años se han producido bastantes bulos científicos (p. ej., el hombre de Piltdown) y algunos fueron intentos deliberados de desacreditar determinados campos de investigación (p. ej., el caso Sokal).
- **Tergiversación de datos**. Si bien el fraude ocupa la mayoría de los titulares, en mi experiencia es mucho más común ver datos tergiversados. Cuando digo esto no me refiero a que los periódicos se equivoquen (cosa que hacen, casi siempre). Me refiero al hecho de que a menudo los datos en realidad no dicen lo que los investigadores creen que dicen. Supongo que, casi siempre, esto no es el resultado de una deshonestidad deliberada, sino que se debe a una falta de sofisticación en los análisis de datos. Por ejemplo, piensa en el ejemplo de la paradoja de Simpson que analicé al principio de este libro. Es muy común ver a las personas presentar datos "agregados" de algún tipo y, a veces, cuando profundizas y encuentras los datos sin procesar, descubres que los datos agregados cuentan una historia diferente a los datos desagregados. Alternativamente, puede encontrar que algún aspecto de los datos está oculto, porque cuenta una historia inconveniente (p. ej., el investigador puede optar por no referirse a una variable en particular). Hay muchas variantes de esto, muchas de las cuales son muy difíciles de detectar.
- **Estudiar el "diseño erróneo"**. Vale, este es sutil. Básicamente, el problema aquí es que un investigador diseña un estudio que tiene fallos incorporadas y esos fallos nunca se informan en el artículo. Los datos que se reportan son completamente reales y están correctamente analizados, pero son producto de un estudio que en realidad está bastante mal elaborado. El investigador realmente quiere encontrar un efecto particular y, por lo tanto, el estudio se configura de tal manera que sea "fácil" observar (artefactualmente) ese efecto. Una forma astuta de hacer esto, en caso de que te apetezca hacer un poco de fraude, es diseñar un experimento en el que sea obvio para los participantes lo que "se supone" que deben hacer, y luego dejar que la reactividad haga su magia. Si lo deseas, puedes agregar todas las trampas de la experimentación doble ciego, pero no supondrá ninguna diferencia, ya que los propios materiales del estudio le están diciendo sutilmente a la gente lo que tú quieres que hagan. Cuando escribas los resultados, el fraude no será evidente para el lector. Lo que es obvio para el participante cuando está en el contexto experimental no siempre lo es para la persona que lee el artículo. Por supuesto, la forma en que lo he descrito hace que parezca que siempre es un fraude. Probablemente hay casos en los que esto se hace deliberadamente, pero en mi experiencia, la mayor preocupación ha sido el diseño erróneo no intencionado. El investigador cree y, por tanto, el estudio acaba teniendo un fallo incorporado,

y ese fallo se borra mágicamente cuando el estudio se redacta para su publicación.

- **Minería de datos y elaboración de hipótesis post hoc**. Otra forma en que los autores de un estudio pueden tergiversar más o menos los datos es participar en lo que se conoce como "minería de datos" (ver Gelman y Loken 2014, para una discusión más amplia de esto como parte del "jardín de caminos que se bifurcan" en el análisis estadístico). Como veremos más adelante, si sigues tratando de analizar los datos de muchas maneras diferentes, eventualmente encontrarás algo que "parece" un efecto real pero no lo es. Esto se conoce como "minería de datos". Antes era muy poco frecuente porque el análisis de datos solía llevar semanas, pero ahora que todo el mundo dispone de programas estadísticos potentes en sus ordenadores, se está convirtiendo en algo muy común. La minería de datos en sí no es "incorrecta", pero cuanto más se hace, mayor es el riesgo que se corre. Lo que está mal, y sospecho que es muy común, es la minería de datos no reconocida. Es decir, el investigador ejecuta todos los análisis posibles conocidos por la humanidad, encuentra el que funciona y luego finge que este fue el único análisis que realizó. Peor aún, a menudo "inventan" una hipótesis después de mirar los datos para encubrir la extracción de datos. Para que quede claro. No está mal cambiar de opinión después de analizar los datos y volver a analizar los datos con las nuevas hipótesis "post hoc". Lo que está mal (y sospecho que es común) es no reconocer lo que has hecho. Si reconoces que lo has hecho, otros investigadores podrán tener en cuenta tu comportamiento. Si no lo haces, no podrán hacerlo. Y eso convierte tu comportamiento en engañoso. Malo
- **Sesgo de publicación y autocensura**. Finalmente, un sesgo generalizado es la "no notificación" de los resultados negativos. Esto es casi imposible de prevenir. Las revistas no publican todos los artículos que se les envían. Prefieren publicar artículos que encuentran "algo". Así, si 20 personas realizan un experimento para ver si leer Finnegans Wake causa locura en los humanos, y 19 de ellos descubren que no es así, ¿cuál crees que se publicará? Obviamente, el único estudio que encontró que Finnegans Wake causa locura.[7] Este es un ejemplo de un sesgo de publicación. Dado que nadie publicó los 19 estudios que no encontraron un efecto, un lector ingenuo nunca sabría que existieron. Peor aún, la mayoría de los investigadores "internalizan" este sesgo y terminan autocensurándose en su investigación. Sabiendo que los resultados negativos no serán aceptados para su publicación, ni siquiera intentan informarlos. Como dice una amiga mí "por cada experimento que te publican, también tienes 10 fracasos". Y tiene razón. El problema es que, si bien algunos (quizás la mayoría) de esos estudios son fracasos por razones aburridas (por ejemplo, porque has metido la pata en algo), otros pueden ser auténticos resultados "nulos" que debes reconocer cuando escribes el experimento "bueno". Y a menudo es difícil saber cuál es cuál. Un buen punto de partida es un artículo de Ioannidis (2005) con el deprimente título "Por qué la mayoría de los hallazgos de investigación publicados son falsos". También sugeriría echar un vistazo al trabajo de Kühberger et al. (2014) que presenta evidencia estadística de que esto realmente sucede en psicología.

Probablemente haya muchos más problemas como este en los que pensar, pero eso servirá para empezar. Lo que realmente quiero señalar es la verdad cegadoramente obvia de que la ciencia del mundo real la realizan humanos reales, y solo las personas más crédulas asumen automáticamente que todos los demás son honestos e imparciales.

[7] Claramente, el efecto real es que solo las personas locas intentarían leer *Finnegans Wake*

Los científicos reales no suelen ser tan ingenuos, pero por alguna razón al mundo le gusta fingir que lo somos, y los libros de texto que solemos escribir parecen reforzar ese estereotipo.

2.8 Resumen

En realidad, este capítulo no pretende proporcionar una discusión exhaustiva de los métodos de investigación psicológica. Se necesitaría otro volumen tan largo como este para hacer justicia al tema. Sin embargo, en la vida real, la estadística y el diseño de estudios están tan estrechamente entrelazados que es muy útil discutir algunos de los temas clave. En este capítulo, he discutido brevemente los siguientes temas:

- Introducción a la medición psicológica. ¿Qué significa operativizar un constructo teórico? ¿Qué significa tener variables y tomar medidas?
- Escalas de medida y tipos de variables. Recuerda que hay dos distinciones diferentes aquí. Existe la diferencia entre datos discretos y continuos, y existe la diferencia entre los cuatro tipos de escala diferentes (nominal, ordinal, de intervalo y de razón).
- Evaluación de la fiabilidad de una medida. Si mido "lo mismo" dos veces, ¿debería esperar ver el mismo resultado? Sólo si mi medida es fiable. Pero, ¿qué significa hablar de hacer "lo mismo"? Bueno, es por eso que tenemos diferentes tipos de fiabilidad. Asegúrate de recordar cuáles son.
- El "rol" de las variables: predictores y resultados. ¿Qué papel juegan las variables en un análisis? ¿Puedes recordar la diferencia entre predictores y resultados? ¿Variables dependientes e independientes? Etc.
- Diseños [de investigación experimental y no experimental]. ¿Qué hace que un experimento sea un experimento? ¿Es una bonita bata blanca de laboratorio o tiene algo que ver con el control del investigador sobre las variables?
- Evaluar la validez de un estudio. ¿Tu estudio mide lo que tú quieres? ¿Cómo podrían salir mal las cosas? ¿Y es mi imaginación, o fue una lista muy larga de posibles formas en que las cosas pueden salir mal?

Todo esto debería dejarte claro que el diseño del estudio es una parte fundamental de la metodología de la investigación. Construí este capítulo a partir del librito clásico de Campbell & Stanley (1963), pero, por supuesto, hay una gran cantidad de libros de texto sobre diseños de investigación. Dedica unos minutos a tu motor de búsqueda favorito y encontrarás docenas.

Part II

Una introducción a jamovi

Chapter 3

Primeros pasos con jamovi

Es bueno trabajar con robots.
– Roger Zelazny[1]

En este capítulo hablaré de cómo empezar a utilizar jamovi. Hablaré brevemente sobre cómo descargar e instalar jamovi, pero la mayor parte del capítulo se centrará en que te familiarices con la interfaz gráfica de jamovi. Nuestro objetivo en este capítulo no es aprender ningún concepto estadístico: solo trataremos de aprender cómo funciona jamovi para sentirnos cómodas interactuando con el sistema. Para ello, dedicaremos algo de tiempo a ver conjuntos de datos y variables. Al hacerlo, te harás una idea de cómo es trabajar en jamovi.

Sin embargo, antes de entrar en detalles, merece la pena hablar un poco de por qué quieres usar jamovi. Dado que estás leyendo esto, probablemente tengas tus propias razones. Sin embargo, si esas razones son "porque es lo que se usa en mi clase de estadística", puede que merezca la pena explicar un poco por qué tu profesor o profesora ha elegido usar jamovi para la clase. Por supuesto, no sé realmente por qué *otras* personas eligen jamovi, así que realmente estoy hablando de por qué lo uso yo.

- Es algo obvio, pero vale la pena decirlo de todos modos: calcular los estadísticos en un ordenador es más rápido, más fácil y más potente que hacerlo a mano. Los ordenadores destacan en tareas repetitivas sin sentido, y muchos cálculos estadísticos son repetitivos y sin sentido. Para la mayoría de la gente, la única razón para hacer cálculos estadísticos con lápiz y papel es el aprendizaje. En mi clase sugiero de vez en cuando hacer algunos cálculos de esa manera, pero el único valor real es pedagógico. Hacer algunos cálculos te ayuda a "sentir" la estadística, así que vale la pena hacerlo una vez. Pero sólo una vez.
- Hacer estadística en una hoja de cálculo convencional (por ejemplo, Microsoft Excel) suele ser una mala idea a largo plazo. Aunque es probable que mucha gente se sienta más familiarizada con ellas, las hojas de cálculo son muy limitadas en cuanto a los cálculos que permiten realizar. Si te acostumbras a intentar hacer análisis de datos de la vida real usando hojas de cálculo, te habrás metido en un agujero muy profundo.

[1] Fuente: Dismal Light (1968).

- También existen "versiones para estudiantes" (versiones reducidas de la versión real) muy baratas, y luego venden "versiones educativas" completas a un precio que me hace estremecer. También venden licencias comerciales a un precio asombrosamente alto. El modelo de negocio consiste en engatusarte durante tu época de estudiante y dejarte dependiente de sus herramientas cuando salgas al mundo real. Es difícil culparles por intentarlo, pero personalmente no soy partidaria de desembolsar miles de dólares si puedo evitarlo. Y se puede evitar. Si utilizas paquetes como jamovi que son de código abierto y gratuitos, no tendrás que pagar licencias desorbitadas.
- Algo que tal vez no aprecies ahora, pero que te encantará más adelante si haces algo relacionado con el análisis de datos, es el hecho de que jamovi es básicamente una interfaz sofisticada para el lenguaje de programación estadística gratuita R. Cuando descargas e instalas R, obtienes todos los "paquetes" básicos que son muy potentes por sí solos. Sin embargo, como R es tan abierto y ampliamente utilizado, se ha convertido en una herramienta estándar en estadística, por lo que mucha gente escribe sus propios paquetes que amplían el sistema. Y estos también están disponibles gratuitamente. Una de las consecuencias de esto, he notado, es que si nos fijamos en los últimos libros de texto de análisis de datos avanzados, muchos de ellos utilizan R.

Esas son las principales razones por las que uso jamovi. Pero no está exento de defectos. Es relativamente nuevo[2], por lo que no hay una gran cantidad de libros de texto y otros recursos que lo apoyen, y tiene algunas peculiaridades molestas con las que todos estamos bastante atascados, pero en en general, creo que los puntos fuertes superan a las debilidades; más que cualquier otra opción que he encontrado hasta ahora.

3.1 Instalación de jamovi

Vale, basta de argumentos de venta. Empecemos. Como cualquier otro programa, jamovi debe instalarse en un "ordenador", que es una caja mágica que hace cosas chulas y reparte ponis gratis. O algo parecido; puede que esté confundiendo los ordenadores con las campañas de marketing del iPad. En cualquier caso, jamovi se distribuye gratuitamente en internet y se puede descargar desde la página principal de jamovi, que es: https://www.jamovi.org/

En la parte superior de la página, bajo el epígrafe "Descargar", verás enlaces separados para usuarios de Windows, Mac y Linux. Si sigues el enlace correspondiente, verás que las instrucciones en línea se explican por sí solas. En el momento de escribir estas líneas, la versión actual de jamovi es la 2.3, pero suelen publicar actualizaciones cada pocos meses, por lo que probablemente tendrás una versión más reciente.[3]

3.1.1 Poner en marcha jamovi

De un modo u otro, independientemente del sistema operativo que utilices, es hora de abrir jamovi y empezar. Al iniciar jamovi por primera vez, aparecerá una interfaz de

[2] En el momento de escribir esto por primera vez en agosto de 2018. Las versiones posteriores de este libro usarán versiones posteriores de jamovi.

[3] Aunque jamovi se actualiza con frecuencia, no suele suponer una gran diferencia para el tipo de trabajo que haremos en este libro. De hecho, durante la redacción de este libro lo actualicé varias veces y no supuso ninguna diferencia con respecto al contenido de este libro.

usuario parecida a Figure 3.1.

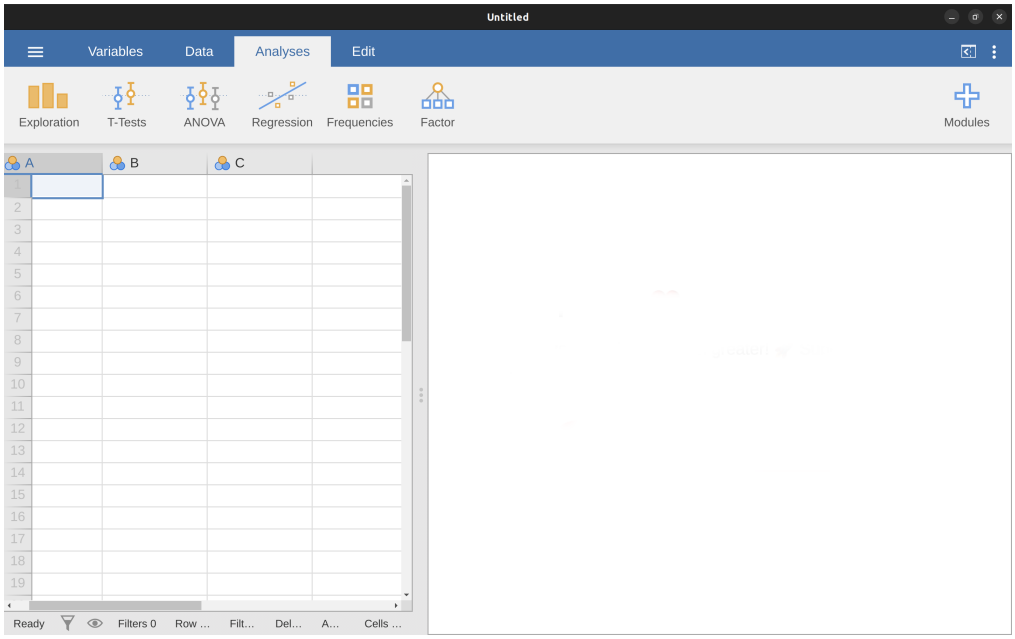


Figure 3.1: ¡jamovi se pone en marcha!

A la izquierda está la vista de hoja de cálculo, y a la derecha es donde aparecen los resultados de las pruebas estadísticas. En el centro hay una barra que separa estas dos regiones y se puede arrastrar hacia la izquierda o hacia la derecha para cambiar su tamaño.

Es posible simplemente empezar a escribir valores en la hoja de cálculo jamovi como lo harías en cualquier otro software de hoja de cálculo. Alternativamente, se pueden abrir archivos en formato CSV (.csv). Además, puedes importar fácilmente archivos SPSS, SAS, Stata y JASP directamente en jamovi. Para abrir un archivo, selecciona la pestaña Archivo (tres líneas horizontales indican esta pestaña) en la esquina superior izquierda, selecciona 'Abrir' y luego elige entre los archivos que aparecen en 'Examinar' en función de si deseas abrir un ejemplo o un archivo almacenado en tu ordenador.

3.2 Análisis

Los análisis se pueden seleccionar desde la cinta o el menú de análisis en la parte superior. Al seleccionar un análisis, aparecerá un "panel de opciones" para ese análisis en particular, que te permitirá asignar diferentes variables a distintas partes del análisis y seleccionar diferentes opciones. Al mismo tiempo, los resultados del análisis aparecerán en el 'Panel de resultados' de la derecha y se actualizarán en tiempo real a medida que modifiques las opciones.

Cuando hayas configurado correctamente el análisis, puedes descartar las opciones de análisis haciendo clic en la flecha en la parte superior derecha del panel opcional. Si

deseas volver a estas opciones, puedes hacer clic en los resultados que se produjeron. De esta forma, puedes volver a cualquier análisis que tú (o, por ejemplo, un colega) hayas creado anteriormente.

Si decides que ya no necesitas un análisis en particular, puedes eliminarlo con el menú contextual de resultados. Haciendo clic con el botón derecho del ratón en los resultados del análisis, aparecerá un menú y seleccionando 'Análisis' y luego 'Eliminar', se puede eliminar el análisis. Pero hablaremos de esto más adelante. Primero, echemos un vistazo más detallado a la vista de hoja de cálculo.

3.3 La hoja de cálculo

En jamovi, los datos se representan en una hoja de cálculo en la que cada columna representa una 'variable' y cada fila un 'caso' o 'participante'.

3.3.1 Variables

Las variables más utilizadas en jamovi son las 'variables de datos', que contienen datos cargados desde un archivo de datos o 'escritos' por el usuario. Las variables de datos pueden ser uno de varios niveles de medida (Figure 3.2).

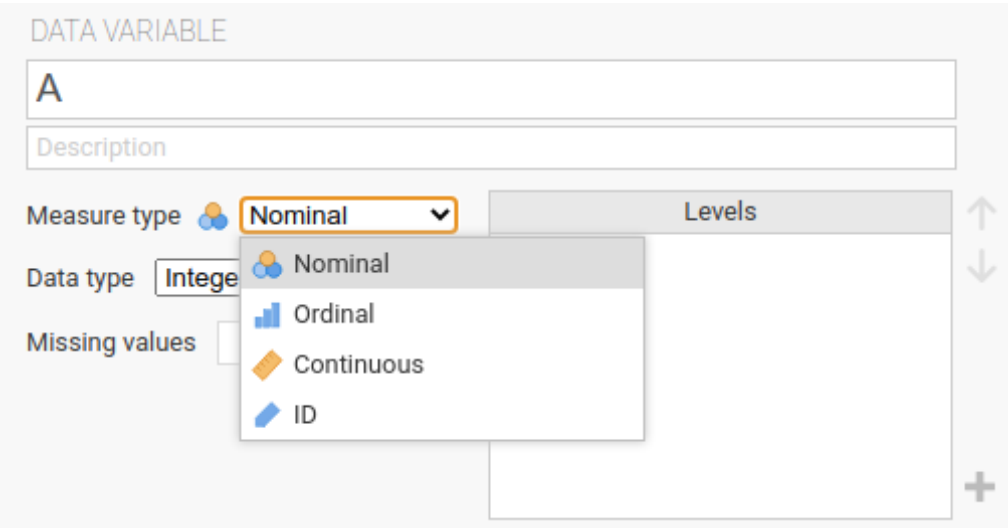


Figure 3.2: niveles de medida

Estos niveles se designan mediante el símbolo que aparece en la cabecera de la columna de la variable. El tipo de variable ID es exclusivo de jamovi. Está pensado para variables que contienen identificadores que casi nunca querrás analizar. Por ejemplo, el nombre de una persona o el ID de un participante. Especificar un tipo de variable de ID puede mejorar el rendimiento cuando se interactúa con conjuntos de datos muy grandes.

Las variables *nominales* son para variables categóricas que son etiquetas de texto, por ejemplo, una columna llamada Sexo con los valores Masculino y Femenino sería nominal. También lo sería el nombre de una persona. Los valores de las variables nominales

también pueden tener un valor numérico. Estas variables se utilizan más a menudo cuando se importan datos que codifican valores con números en lugar de texto. Por ejemplo, una columna de un conjunto de datos puede contener los valores 1 para hombres y 2 para mujeres. Es posible añadir etiquetas 'legibles' a estos valores con el editor de variables (más información más adelante).

Las variables *ordinales* son como las variables Nominales, excepto que los valores tienen un orden específico. Un ejemplo es una escala Likert en la que 3 es 'totalmente de acuerdo' y -3 es 'totalmente en desacuerdo'.

Las variables *continuas* son variables que existen en una escala continua. Por ejemplo, la altura o el peso. También se denomina 'Escala de intervalo' o 'Escala de razón'.

Además, también puedes especificar diferentes tipos de datos: las variables tienen un tipo de datos de 'Texto', 'Entero' o 'Decimal'.

Al empezar con una hoja de cálculo en blanco e introducir valores el tipo de variable cambiará automáticamente en función de los datos que introduzcas. Esta es una buena manera de hacerse una idea de qué tipos de variables van con qué tipo de datos. Del mismo modo, al abrir un archivo de datos, Jamovi intentará adivinar el tipo de variable a partir de los datos de cada columna. En ambos casos, este enfoque automático puede no ser correcto y puede ser necesario especificar manualmente el tipo de variable con el editor de variables.

El editor de variables se puede abrir seleccionando 'Configuración' en la pestaña de datos o haciendo doble clic en la cabecera de la columna de variables. El editor de variables permite cambiar el nombre de la variable y, en el caso de las variables de datos, el tipo de variable, el orden de los niveles y la etiqueta que aparece en cada nivel. Los cambios se pueden aplicar haciendo clic en la 'marca' situada en la parte superior derecha. Se puede salir del editor de variables haciendo clic en la flecha 'Ocultar'.

Se pueden insertar o añadir nuevas variables al conjunto de datos usando el botón 'añadir' de la cinta de datos. El botón 'añadir' también permite añadir variables calculadas.

3.3.2 Variables calculadas

Las Variables calculadas son aquellas que obtienen su valor realizando un cálculo sobre otras variables. Las variables calculadas se pueden utilizar para diversos fines, como transformaciones logarítmicas, puntuaciones z, puntuaciones de suma, puntuaciones negativas y medias.

Las variables calculadas pueden añadirse al conjunto de datos con el botón 'añadir' disponible en la pestaña de datos. Aparecerá un cuadro de fórmulas en el que podrás especificar la fórmula. Están disponibles los operadores aritméticos habituales. Algunos ejemplos de fórmulas son:

A + B LOG10(largo) MEDIA(A, B) (largo - VMEAN(largo)) / VSTDEV(largo)

En orden, son la suma de A y B, una transformación logarítmica (base 10) de len, la media de A y B, y la puntuación z de la variable len[4] . Figure 3.3 muestra la pantalla

[4]en versiones posteriores de jamovi hay una función predefinida 'Z' para calcular las puntuaciones z, que es mucho más fácil.

jamovi para la nueva variable calculada como la puntuación z de len (del conjunto de datos de ejemplo 'Crecimiento de dientes').

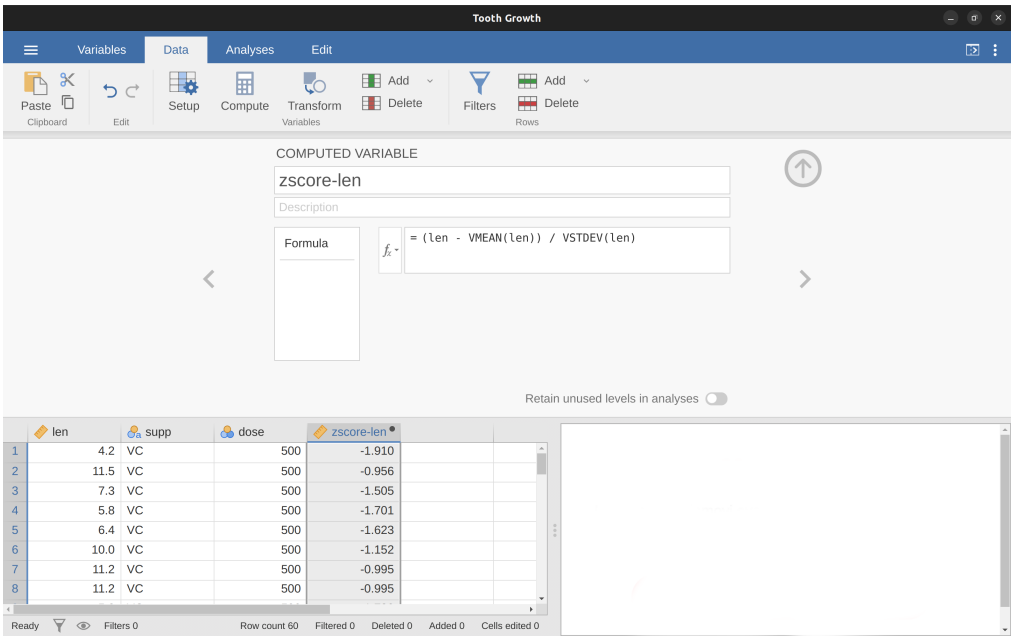


Figure 3.3: una nueva variable calculada, la puntuación z de 'dosis'

3.3.2.1 Funciones V

Varias funciones ya están disponibles en jamovi y están disponibles en el cuadro desplegable denominado fx. Varias funciones aparecen en pares, una con el prefijo V y la otra no. Las funciones V realizan sus cálculos sobre una variable en su conjunto, mientras que las funciones que no son V realizan su cálculo fila por fila. Por ejemplo, MEAN(A, B) producirá la media de A y B para cada fila. Donde como VMEAN(A) da la media de todos los valores de A.

3.3.3 Copiar y pegar

jamovi produce tablas con formato de la Asociación Americana de Psicología (APA) y gráficos atractivos. Suele ser útil poder copiarlos y pegarlos, por ejemplo en un documento de Word o en un correo electrónico a un colega. Para copiar los resultados, haz clic con el botón derecho en el objeto de interés y, en el menú, selecciona exactamente lo que deseas copiar. El menú te permite elegir entre copiar solo la imagen o todo el análisis. Al seleccionar "copiar" se copia el contenido en el portapapeles y este se puede pegar en otros programas de la forma habitual. Puedes practicar esto más adelante cuando hagamos algunos análisis.

3.3.4 Modo de sintaxis

jamovi también proporciona un "Modo de sintaxis R". En este modo, jamovi produce un código R equivalente para cada análisis. Para cambiar al modo de sintaxis, selecciona el menú Aplicación en la parte superior derecha de jamovi (un botón con tres puntos verticales) y haz clic en la casilla de verificación "Modo de sintaxis". Puedes desactivar el modo de sintaxis haciendo clic una segunda vez.

En el modo de sintaxis, los análisis siguen funcionando como antes, pero ahora producen sintaxis de R y 'salida ascii' como una sesión R. Como todos los objetos de resultados en jamovi, puedes hacer clic con el botón derecho en estos elementos (incluida la sintaxis de R) y copiarlos y pegarlos, por ejemplo, en una sesión de R. Actualmente, la sintaxis de R proporcionada no incluye el paso de importación de datos, por lo que debe realizarse manualmente en R. Existen muchos recursos que explican cómo importar datos en R y, si estás interesada, te recomendamos que los consultes; solo tienes que buscar en la interweb.

3.4 Carga de datos en jamovi

Existen varios tipos de archivos que pueden resultarnos útiles a la hora de analizar datos. Hay dos en particular que son especialmente importantes desde la perspectiva de este libro:

- Los *archivos jamovi* son los que tienen una extensión de archivo .omv. Este es el tipo de archivo estándar que utiliza jamovi para almacenar datos, variables y análisis.
- *Los archivos de valores separados por comas (csv)* son los que tienen una extensión .csv. Son archivos de texto normales y corrientes que pueden abrirse con muchos programas de software diferentes. Es bastante habitual almacenar datos en archivos csv, precisamente por su sencillez.

También hay otros tipos de archivos de datos que puedes importar a jamovi. Por ejemplo, puedes abrir hojas de cálculo de Microsoft Excel (archivos .xls) o archivos de datos guardados en formatos nativos de otros programas estadísticos, como SPSS o SAS. Sea cual sea el formato de archivo que utilices, es una buena idea crear una carpeta o carpetas especialmente para tus conjuntos de datos y análisis de jamovi y asegurarte de que realizas copias de seguridad con regularidad.

3.4.1 Importar datos de archivos csv

Un formato de datos bastante utilizado es el humilde archivo de "valores separados por comas", también llamado archivo csv, y que suele llevar la extensión de archivo .csv. Los archivos csv son archivos de texto antiguos y lo que almacenan es básicamente una tabla de datos. Esto se ilustra en Figure 3.4, que muestra un archivo llamado booksales.csv que he creado. Como puedes ver, cada fila representa los datos de ventas de libros de un mes. Sin embargo, la primera fila no contiene datos reales, sino los nombres de las variables.

Es fácil abrir archivos csv en jamovi. En el menú superior izquierdo (el botón con tres líneas paralelas), selecciona 'Abrir' y busca en tu ordenador el lugar donde has guardado el archivo csv. Si estás en un Mac, se parecerá a la ventana del Finder habitual que

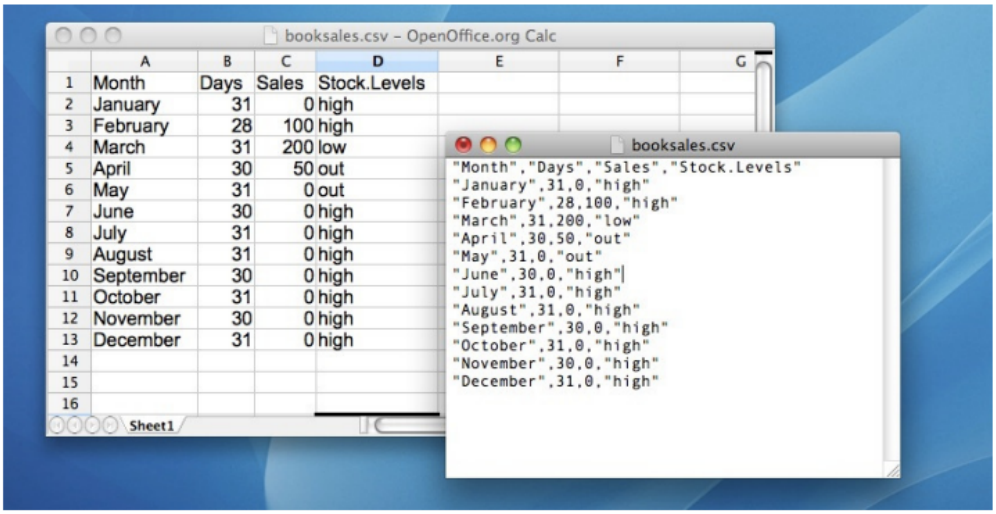


Figure 3.4: el archivo de datos booksales.csv. A la izquierda, he abierto el archivo usando un programa de hoja de cálculo (OpenOffice), que muestra que el archivo es básicamente una tabla. A la derecha, el mismo archivo está abierto en un editor de texto estándar (el programa TextEdit en un Mac), que muestra cómo está formateado el archivo. Las entradas de la tabla van entre comillas y separadas por comas.

usas para elegir un archivo; en Windows se parece a una ventana del Explorador. En Figure 3.5 se muestra un ejemplo del aspecto en un Mac. Asumo que estás familiarizada con tu ordenador, así que no deberías tener ningún problema para encontrar el archivo csv que quieres importar. Busca el que quieras y haz clic en el botón "Abrir".

Hay algunas cosas que puedes comprobar para asegurarte de que los datos se importan correctamente:

- Encabezamiento. ¿La primera fila del archivo contiene los nombres de cada variable, una fila de "encabezado"? El archivo booksales.csv tiene un encabezado, así que eso es un sí.
- Decimal. ¿Qué carácter se utiliza para especificar el punto decimal? En los países de habla inglesa, es casi siempre un punto (es decir, .). Sin embargo, esto no es universalmente cierto, muchos países europeos usan una coma.
- Cita. ¿Qué carácter se utiliza para denotar un bloque de texto? Suele ser una comilla doble ("). Lo es para el archivo booksales.csv.

3.5 Importación de archivos de datos inusuales

A lo largo de este libro he asumido que tus datos están almacenados como un archivo jamovi .omv o como un archivo csv "correctamente" formateado. Sin embargo, en la vida real no es una suposición muy plausible, así que será mejor que hable de otras posibilidades con las que te puedes encontrar.

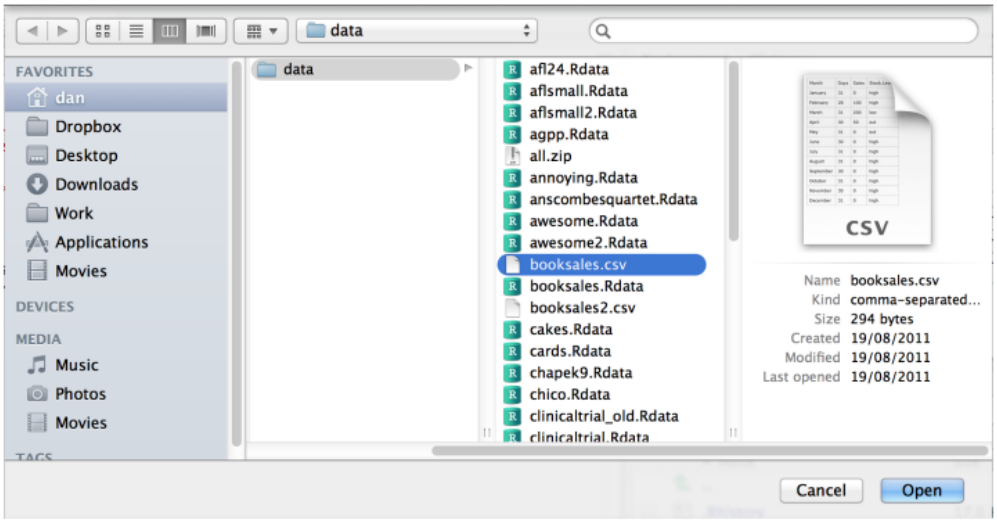


Figure 3.5: Un cuadro de diálogo en un Mac pidiéndote que selecciones el archivo csv que jamovi debe intentar importar. Los usuarios de Mac reconocerán esto inmediatament, es la forma habitual en que un Mac te pide que busques un archivo. Los usuarios de Windows no verán esto, en su lugar verán la ventana del explorador habitual que Windows siempre te ofrece cuando quiere que selecciones un archivo.

3.5.1 Carga de datos de archivos de texto

Lo primero que debo señalar es que si tus datos se guardan como un archivo de texto pero no están en el formato csv adecuado, es muy probable que jamovi pueda abrirlo. Sólo tienes que probarlo y ver si funciona. Sin embargo, a veces es necesario cambiar el formato. Los que a menudo he tenido que cambiar son:

- cabecera. Muchas veces, cuando se almacenan datos en un archivo csv, la primera fila contiene los nombres de las columnas y no los datos. Si es así, es una buena idea abrir el archivo csv en un programa de hoja de cálculo como Open Office y añadir la fila de encabezado manualmente.
- sep. Como indica el nombre "valor separado por comas", los valores de una fila de un archivo csv suelen estar separados por comas. Sin embargo, esto no es universal. En Europa, el punto decimal generalmente se escribe como , en lugar de . y por consiguiente sería un poco incómodo usar , como separador. Por lo tanto, no es raro usar ; en lugar de , como separador. En otras ocasiones, he visto que se utiliza el carácter TAB.
- entrecomillado. Es habitual en los archivos csv incluir un carácter de comillas para los datos textuales. Como puedes ver en el archivo booksales.csv, se trata normalmente de un carácter de comillas dobles, ". Pero a veces no hay ningún carácter de comillas, o puede que en su lugar se utilice una comilla simple '.
- saltar. En realidad, es muy habitual recibir archivos CSV en los que las primeras filas no tienen nada que ver con los datos reales. En su lugar, proporcionan un resumen legible para el ser humano de dónde proceden los datos, o quizás incluyen alguna información técnica que no guarda relación con los datos.

- valores perdidos. A menudo recibirás datos con valores omitidos. Por una razón u otra, faltan algunas entradas en la tabla. El archivo de datos debe incluir un valor "especial" para indicar que falta la entrada. Por defecto, jamovi asume que este valor es 99[5], tanto para datos numéricos como de texto, por lo que debes asegurarte de que, cuando sea necesario, todos los valores que faltan en el archivo csv se reemplacen con 99 (o -9999; lo que elijas) antes de abrir/importar el archivo en jamovi. Una vez que hayas abierto/importado el archivo en jamovi, todos los valores que falten se convertirán en celdas en blanco o sombreadas en la vista de hoja de cálculo de jamovi. También puedes cambiar el valor que falta para cada variable como una opción en la vista Datos - Configuración.

3.5.2 Carga de datos desde SPSS (y otros paquetes estadísticos)

Los comandos enumerados anteriormente son los principales que necesitaremos para los archivos de datos en este libro. Pero en la vida real tenemos muchas más posibilidades. Por ejemplo, es posible que quieras leer archivos de datos de otros programas estadísticos. Dado que SPSS es probablemente el paquete estadístico más utilizado en psicología, vale la pena mencionar que jamovi también puede importar archivos de datos de SPSS (extensión de archivo .sav). Simplemente sigue las instrucciones anteriores para abrir un archivo csv, pero esta vez debes ir al archivo .sav que deseas importar. Para los archivos SPSS, jamovi considerará todos los valores como faltantes si se consideran como archivos "faltantes del sistema" en SPSS. El valor 'faltas por defecto' no parece funcionar como se espera cuando se importan archivos SPSS, así que tenlo en cuenta - puede que necesites otro paso: importa el archivo SPSS a jamovi, luego expórtalo como un archivo csv antes de volver a abrirlo en jamovi. [6]

Y eso es todo, al menos en lo que respecta a SPSS. En cuanto a otros programas estadísticos, jamovi también puede abrir/importar directamente archivos SAS y STATA.

3.5.3 Carga de archivos Excel

Otro problema lo plantean los archivos Excel. A pesar de llevar años pidiendo a la gente que me envíen datos codificados en un formato de datos propietario, me envían muchos archivos de Excel. La forma de tratar los archivos Excel es abrirlos primero en Excel u otro programa de hoja de cálculo que pueda tratar archivos Excel y luego exportar los datos como un archivo csv antes de abrir/importar el archivo csv a jamovi.

3.6 Cambio de datos de un nivel a otro

A veces se desea cambiar el nivel de la variable. Esto puede ocurrir por muchas razones. A veces, al importar datos de archivos, pueden llegar en un formato incorrecto. Los números a veces se importan como valores de nominales de texto. Las fechas se pueden importar como texto. Los valores de ID de participante a veces se pueden leer como

[5] Puedes cambiar el valor por defecto para los valores perdidos en jamovi desde el menú superior derecho (tres puntos verticales), pero esto solo funciona en el momento de importar los archivos de datos a jamovi. El valor omitido por defecto en el conjunto de datos no debe ser un número válido asociado a ninguna de las variables, por ejemplo, podrías usar -9999 ya que es poco probable que sea un valor válido.

[6] Sé que esto es una chapuza, pero funciona y espero que se arregle en una versión posterior de jamovi.

continuos: los valores nominales a veces se pueden leer como ordinales o incluso continuos. Es muy probable que a veces quieras convertir una variable de un nivel de medida a otro. O, para utilizar el término correcto, quieres **coaccionar** la variable de una clase a otra.

Anteriormente vimos cómo especificar diferentes niveles de variables, y si deseas cambiar el nivel de medida de una variable, puedes hacerlo en la vista de datos de jamovi para esa variable. Simplemente haz clic en la casilla de verificación del nivel de medida que desees: continuo, ordinal o nominal.

3.7 Instalación de módulos adicionales en jamovi

Una característica realmente interesante de jamovi es la posibilidad de instalar módulos adicionales de la biblioteca de jamovi. Estos módulos adicionales han sido desarrollados por la comunidad jamovi, es decir, usuarios y desarrolladores de jamovi que han creado complementos de software especiales que realizan otros análisis, generalmente más avanzados, que van más allá de las capacidades del programa jamovi base.

Para instalar módulos adicionales, haz clic en el + grande de la parte superior derecha de la ventana de jamovi, selecciona "jamovi-library" y navega por los diversos módulos adicionales disponibles. Elige los que quieras e instálalos, como en Figure 3.6. Así de fácil. Podrás acceder a los módulos recién instalados desde la barra de botones "Análisis". Pruébalo... entre los módulos complementarios útiles para instalar se incluyen "scatr" (añadido en "Descriptivos") y R_j.

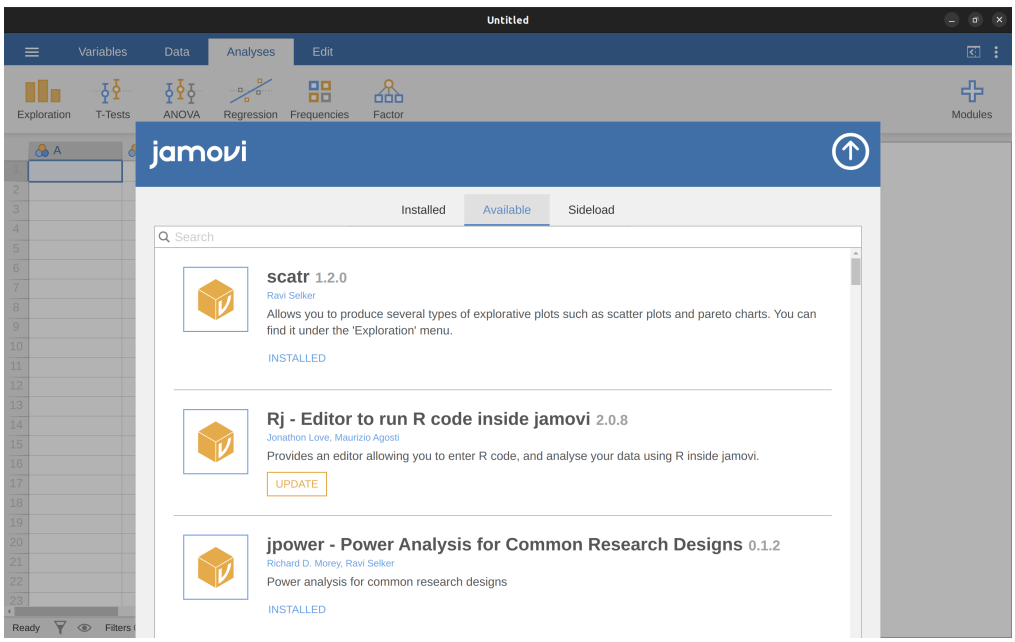


Figure 3.6: Instalación de módulos adicionales en jamovi

3.8 Salir de Jamovi

Hay una última cosa que debería cubrir en este capítulo: cómo salir de jamovi. No es difícil, simplemente cierra el programa de la misma forma que lo harías con cualquier otro programa. Sin embargo, lo que quizás quieras hacer antes de salir es guardar tu trabajo. Esto consta de dos partes: guardar cualquier cambio en el conjunto de datos y guardar los análisis que ejecutaste.

Es una buena práctica guardar cualquier cambio en el conjunto de datos como un *nuevo* conjunto de datos. De este modo, siempre podrás volver a los datos originales. Para guardar cualquier cambio en jamovi, selecciona 'Exportar'...'Datos' en el menú principal de jamovi (botón con tres barras horizontales en la parte superior izquierda) y crea un nuevo nombre de archivo para el conjunto de datos modificado.

También puedes guardar *tanto* los datos modificados como los análisis que hayas realizado guardándolos como un archivo jamovi. Para ello, desde el menú principal de jamovi selecciona 'Guardar como' y escribe un nombre de archivo para este 'archivo jamovi (.omv)'. Recuerda guardar el archivo en un lugar donde puedas encontrarlo más tarde. Yo suelo crear una carpeta nueva para conjuntos de datos y análisis específicos.

3.9 Resumen

Todos los libros que intentan enseñar un nuevo programa de software estadístico a los principiantes tienen que cubrir más o menos los mismos temas y más o menos en el mismo orden. El nuestro no es una excepción así que, siguiendo la gran tradición de hacerlo de la misma manera que todos los demás, este capítulo cubre los siguientes temas:

- [Instalando jamovi]. Descargamos e instalamos jamovi y lo ponemos en marcha.
- Análisis. Nos orientamos muy brevemente hacia la parte de jamovi en la que se realizan los análisis y aparecen los resultados, pero lo aplazamos hasta más adelante en el libro.
- La hoja de cálculo. Dedicamos más tiempo a la parte de la hoja de cálculo de jamovi, y consideramos diferentes tipos de variables y cómo calcular nuevas variables.
- Carga de datos en jamovi. También vimos cómo cargar archivos de datos en jamovi.
- [Importanción de archivos de datos inusuales]. Luego vimos cómo abrir otros archivos de datos, de diferentes tipos de archivos.
- Cambio de datos de un nivel a otro. Y vimos que a veces necesitamos coaccionar datos de un tipo a otro.
- [Instalando módulos adicionales en jamovi]. La instalación de módulos adicionales de la comunidad jamovi realmente amplía las capacidades de jamovi.
- Salir de jamovi. Por último, examinamos las buenas prácticas en términos de guardar el conjunto de datos y los análisis cuando se ha terminado y se está a punto de salir de jamovi.

Todavía no hemos llegado a nada que se parezca al análisis de datos. Quizá el próximo capítulo nos acerque un poco más.

Part III

Trabajar con datos

Chapter 4

Estadística descriptiva

Cuando se dispone de un nuevo conjunto de datos, una de las primeras tareas que hay que hacer es encontrar la manera de resumirlos de forma compacta y fácil de entender. En eso consiste la **estadística descriptiva** (a diferencia de la estadística inferencial). De hecho, para mucha gente el término "estadística" es sinónimo de estadística descriptiva. Este es el tema que trataremos en este capítulo, pero antes de entrar en detalles, tomemos un momento para entender por qué necesitamos la estadística descriptiva. Para ello, abramos el archivo *aflsmall_margins* y veamos qué variables están almacenadas en él, véase Figure 4.1.

De hecho, aquí solo hay una variable, afl.margins. Nos centraremos un poco en esta variable en este capítulo, así que será mejor que te diga lo que es. A diferencia de la mayoría de los conjuntos de datos de este libro, se trata en realidad de datos reales relativos a la Liga de fútbol australiana (AFL).[1] La variable afl.margins contiene el margen ganador (número de puntos) de los 176 partidos jugados en casa y fuera de casa durante la temporada 2010.

Este resultado no facilita la comprensión de lo que dicen realmente los datos. Simplemente "mirar los datos" no es una forma muy eficaz de entenderlos. Para hacernos una idea de lo que dicen realmente los datos, tenemos que calcular algunos estadísticos descriptivos (este capítulo) y dibujar algunas imágenes bonitas (Chapter 5). Dado que los estadísticos descriptivos son los más fáciles de los dos temas, comenzaremos con ellos, pero sin embargo, vamos a mostrar un histograma de los datos de afl.margins, ya que debería ayudar a tener una idea de cómo son los datos que estamos tratando de describir, ver Figure 4.2. Hablaremos mucho más sobre cómo dibujar histogramas en Section 5.1 en el próximo capítulo. Por ahora, basta con mirar el histograma y observar que proporciona una representación bastante interpretable de los datos de afl.margins.

4.1 Medidas de tendencia central

Dibujar los datos, como en Figure 4.2, es una excelente manera de transmitir la "esencia" de lo que los datos intentan decirnos. Suele ser muy útil tratar de condensar los datos

[1] Nota para los no australianos: la AFL es una competición de fútbol de reglas australianas. No es necesario saber nada sobre las reglas australianas para seguir esta sección.

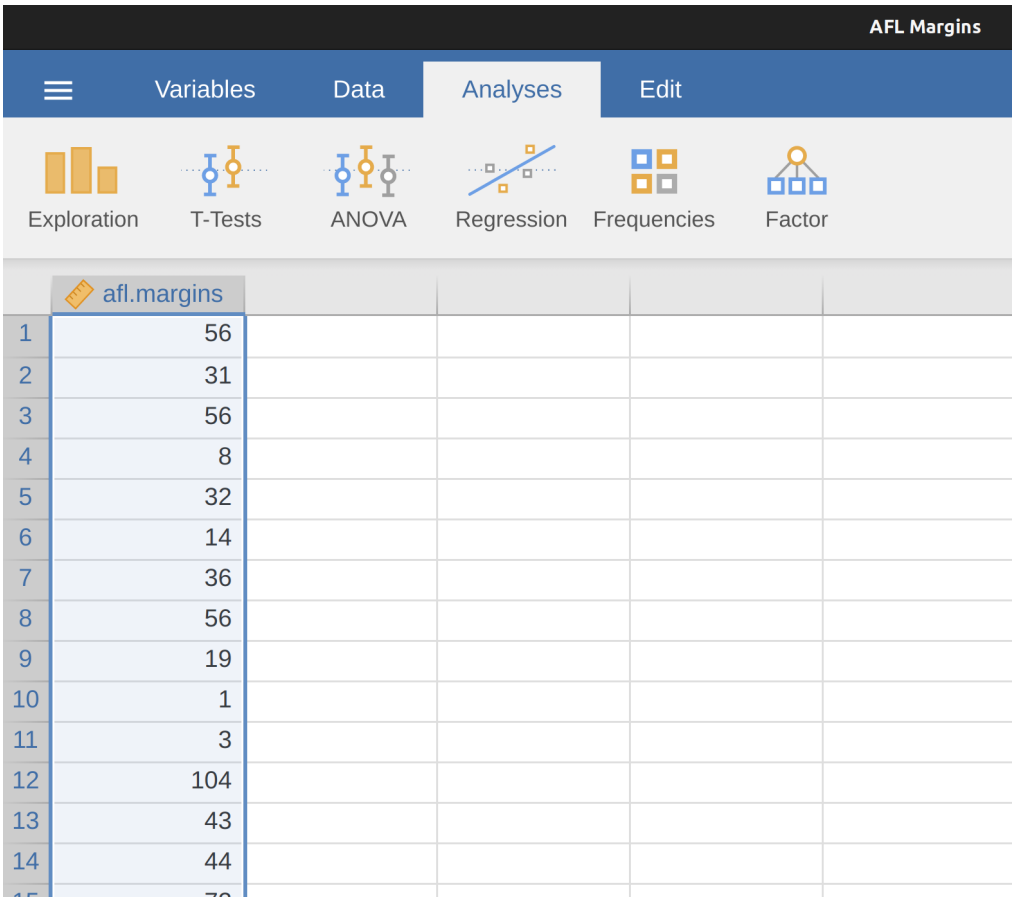


Figure 4.1: una captura de pantalla de jamovi que muestra las variables almacenadas en el archivo aflsmallmargins.csv

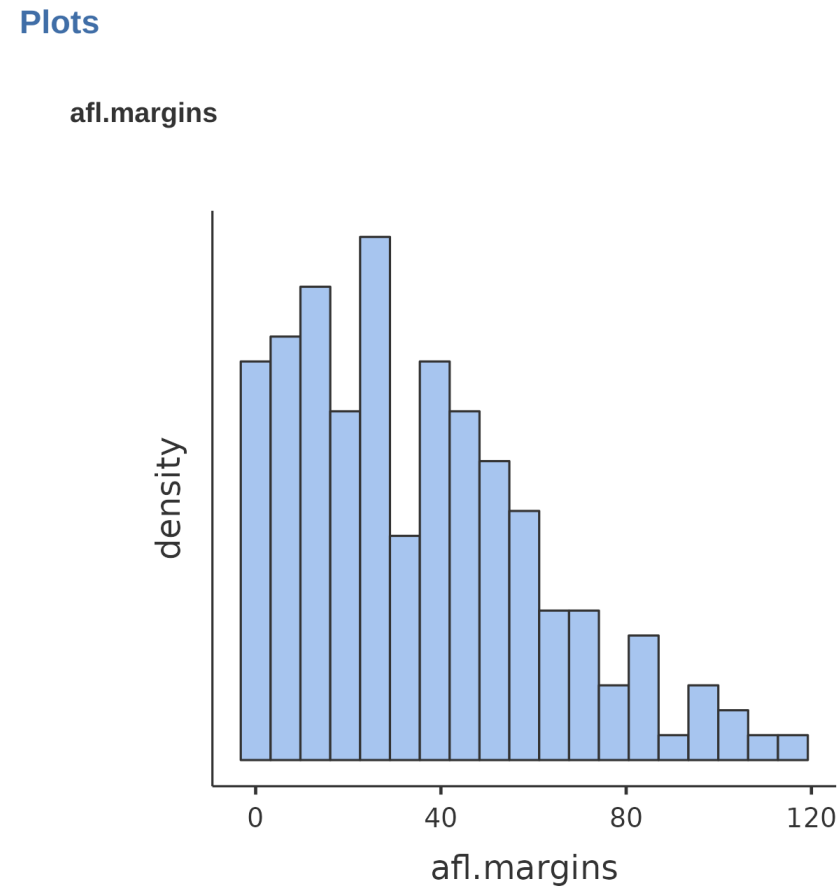


Figure 4.2: Un histograma de los datos del margen ganador de la AFL 2010 (la variable afl.margins). Como era de esperar, cuanto mayor sea el margen de victoria, con menos frecuencia se tiende a verlo.

en unos cuantos estadísticos "resumidos" sencillos. En la mayoría de las situaciones, lo primero que querrás calcular es una medida de **tendencia central**. Es decir, te gustaría saber dónde se encuentra el "promedio" o el "medio" de tus datos. Las tres medidas más utilizadas son la media, la mediana y la moda. Explicaremos cada una de ellas, y luego discutiremos cuándo es útil cada una de ellas.

4.1.1 La media

La **media** de un conjunto de observaciones no es más que un promedio normal y corriente. Se suman todos los valores y se dividen por el número total de valores. Los cinco primeros márgenes ganadores de la AFL fueron 56, 31, 56, 8 y 32, por lo que la media de estas observaciones es simplemente:

$$\frac{56 + 31 + 56 + 8 + 32}{5} = \frac{183}{5} = 36,60$$

Por supuesto, esta definición de la media no es nueva para nadie. Los promedios (es decir, las medias) se usan tan a menudo en la vida cotidiana que se trata de algo bastante familiar. Sin embargo, dado que el concepto de media es algo que todo el mundo entiende, usaré esto como excusa para empezar a introducir algo de la notación matemática que los estadísticos utilizan para describir este cálculo y hablar de cómo se harían los cálculos en jamovi.

La primera notación que hay que introducir es N, que usaremos para referirnos al número de observaciones que estamos promediando (en este caso, $N = 5$). A continuación, debemos adjuntar una etiqueta a las observaciones. Es habitual usar X para esto y utilizar subíndices para indicar de qué observación estamos hablando. Es decir, usaremos X_1 para referirnos a la primera observación, X_2 para referirnos a la segunda observación, y así sucesivamente hasta llegar a X_N para la última. O, para decir lo mismo de una manera un poco más abstracta, usamos X_i para referirnos a la i-ésima observación. Solo para asegurarnos de que tenemos clara la notación, Table 4.1 enumera las 5 observaciones en la variable afl.margins, junto con el símbolo matemático utilizado para referirse a ella y el valor real al que corresponde la observación.

Table 4.1: Observaciones en la variable afl.margins

the observation	its symbol	the observed value
winning margin, game 1	X_1	56 points
winning margin, game 2	X_2	31 points
winning margin, game 3	X_3	56 points
winning margin, game 4	X_4	8 points
winning margin, game 5	X_5	32 points

[Detalle técnico adicional[2]]

4.1.2 Cálculo de la media en jamovi

Bien, esas son las matemáticas. Entonces, ¿cómo conseguimos que la caja mágica de la informática haga el trabajo por nosotros? Cuando el número de observaciones comienza a ser grande, es mucho más fácil hacer este tipo de cálculos con un ordenador. Para calcular la media usando todos los datos podemos utilizar jamovi. El primer paso es hacer clic en el botón 'Exploración' y luego hacer clic en 'Descriptivos'. A continuación, marca la variable afl.margins y haz clic en la flecha hacia la derecha para moverla al cuadro 'Variables'. En cuanto lo hagas, aparecerá una tabla en la parte derecha de la pantalla con información por defecto sobre los 'Descriptivos'; ver Figure 4.3.

Como puedes ver en Figure 4.3, el valor medio de la variable afl.margins es 35,30. Otra información presentada incluye el número total de observaciones (N=176), el número de valores perdidos (ninguno) y los valores de la mediana, mínimo y máximo de la variable.

4.1.3 La mediana

La segunda medida de tendencia central que la gente usa mucho es la **mediana**, y es incluso más fácil de describir que la media. La mediana de un conjunto de observaciones es simplemente el valor medio. Como antes, imaginemos que solo nos interesan los primeros 5 márgenes ganadores de la AFL: 56, 31, 56, 8 y 32. Para calcular la mediana ordenamos estos números en orden ascendente:

[2]Bien, ahora intentemos escribir una fórmula para la media. Por tradición, usamos $\bar{X}$ como notación para la media. Así que el cálculo de la media podría expresarse mediante la siguiente fórmula:

$$\bar{X} = \frac{X_1 + X_2 ... + X_{N-1} + X_N}{N}$$

Esta fórmula es completamente correcta pero es terriblemente larga, por lo que usamos el símbolo del sumatorio $\sum$ para acortarla.[a] Si quiero sumar las cinco primeras observaciones, podría escribir la suma de la forma larga, $X_1 + X_2 + X_3 + X_4 + X_5$ o podría usar el símbolo de suma para acortarla a esto:

$$\sum_{i=1}^{5} X_i$$

Tomado al pie de la letra, esto podría leerse como "la suma, tomada sobre todos los valores i del 1 al 5, del valor X_i". Pero básicamente lo que significa es "sumar las primeras cinco observaciones". En cualquier caso, podemos usar esta notación para escribir la fórmula de la media, que tiene este aspecto:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Sinceramente, no creo que toda esta notación matemática ayude a aclarar el concepto de la media en absoluto. De hecho, no es más que una forma elegante de escribir lo mismo que dije con palabras: sumar todos los valores y dividirlos por el número total de elementos. Sin embargo, esa no es realmente la razón por la que entré en tanto detalle. Mi objetivo era tratar de asegurarme de que todo el mundo leyendo este libro tenga clara la notación que usaremos a lo largo del mismo: $\bar{X}$ para la media, $\sum$ para la idea del sumatorio, $ X_i$ para la i-ésima observación y N para el número total de observaciones. Vamos a reutilizar estos símbolos un poco, por lo que es importante que los entiendas lo suficientemente bien como para poder "leer" las ecuaciones y poder ver que solo está diciendo "suma muchas cosas y luego divide por otra cosa". —[a] *La elección de usar* $\sum$ para denotar el sumatorio no es arbitraria. Es la letra mayúscula griega sigma, que es el análogo de la letra S en ese alfabeto. De manera similar, hay un símbolo equivalente que se usa para denotar la multiplicación de muchos números, dado que las multiplicaciones también se llaman "productos" usamos el símbolo $\prod$ para esto (la pi mayúscula griega, que es el análogo de la letra P).

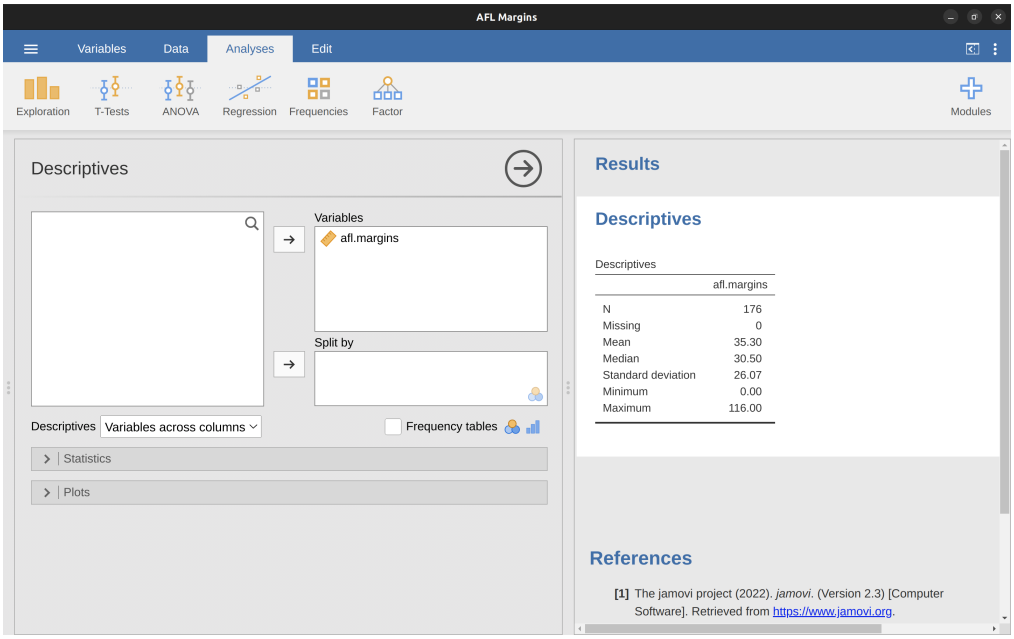


Figure 4.3: Descriptivos por defecto para los datos del margen ganador de AFL 2010 (la variable afl.margins)

8, 31, **32**, 56, 56

A partir de la inspección, es obvio que el valor mediano de estas 5 observaciones es 32, ya que es el del medio en la lista ordenada (lo he puesto en negrita para que sea aún más obvio). Esto es fácil. Pero, ¿qué debemos hacer si nos interesan los primeros 6 partidos en lugar de los primeros 5? Como el sexto partido de la temporada tuvo un margen de ganancia de 14 puntos, nuestra lista ordenada ahora es

8, 14, **31**, **32**, 56, 56

y hay dos números en el medio, 31 y 32. La mediana se define como el promedio de esos dos números, que por supuesto es 31,5. Como antes, es muy tedioso hacer esto a mano cuando se tienen muchos números. En la vida real, por supuesto, nadie calcula la mediana ordenando los datos y buscando el valor medio. En la vida real usamos un ordenador para que haga el trabajo pesado por nosotras, y jamovi nos ha proporcionado un valor de mediana de 30.50 para la variable afl.margins (Figure 4.3).

4.1.4 ¿Media o mediana? ¿Cuál es la diferencia?

Saber calcular medias y medianas es solo una parte de la historia. También hay que entender qué dice cada una de ellas sobre los datos y lo que eso implica en relación a cuándo se debe usar cada una. Esto se ilustra en Figure 4.4. La media es algo así como el "centro de gravedad" del conjunto de datos, mientras que la mediana es el "valor medio" de los datos. Lo que esto implica, en cuanto a cuál deberías usar, depende un poco del tipo de datos que tengas y de lo que estés intentando conseguir. Como guía aproximada:

- Si tus datos son de escala nominal, probablemente no deberías usar ni la media ni la mediana. Tanto la media como la mediana se basan en la idea de que los números asignados a los valores son significativos. Si el esquema de numeración es arbitrario, probablemente sea mejor usar la Moda en su lugar.
- Si tus datos son de escala ordinal, es más probable que quieras usar la mediana que la media. La mediana solo utiliza la información de orden de los datos (es decir, qué números son mayores) pero no depende de los números exactos involucrados. Esta es exactamente la situación que se da cuando tus datos son de escala ordinal. La media, por otro lado, utiliza los valores numéricos precisos asignados a las observaciones, por lo que no es realmente apropiada para datos ordinales.
- Para datos de escala de intervalo y razón, cualquiera de los dos suele ser aceptable. La elección depende un poco de lo que se quiera conseguir. La media tiene la ventaja de que utiliza toda la información de los datos (lo cual es útil cuando no se dispone de muchos datos). Pero es muy sensible a los valores extremos y atípicos.

Vamos a ampliar un poco esta última parte. Una consecuencia es que hay diferencias sistemáticas entre la media y la mediana cuando el histograma es asimétrico (asimetría y apuntamiento). Esto se ilustra en Figure 4.4. Observa que la mediana (a la derecha) se sitúa más cerca del "cuerpo" del histograma, mientras que la media (a la izquierda) se arrastra hacia la "cola" (donde están los valores extremos). Por poner un ejemplo concreto, supongamos que Bob (ingreso $50 000), Kate (ingreso $60 000) y Jane (ingreso $65 000) están sentados en una mesa. La renta media de la mesa es $58,333 y la renta mediana es $60,000. Entonces Bill se sienta con ellos (ingresos $100,000,000). La renta media ha subido a $25,043,750 pero la mediana sube solo a $62,500. Si lo que te interesa es ver la renta total en la tabla, la media podría ser la respuesta correcta. Pero si lo que te interesa es lo que se considera una renta típica en la mesa, la mediana sería una mejor opción.

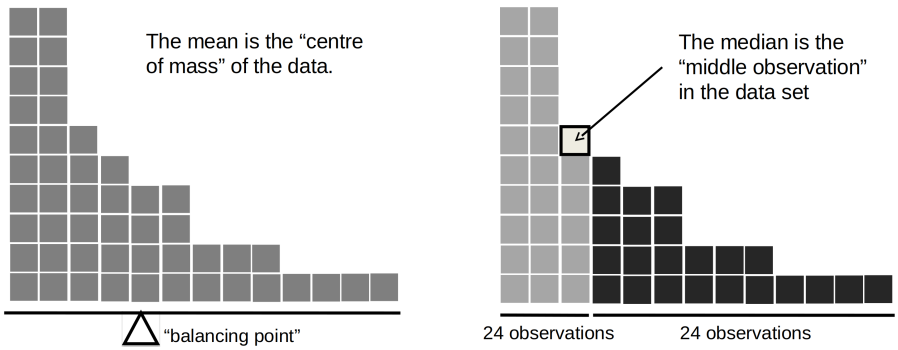


Figure 4.4: Ilustración de la diferencia entre cómo deben interpretarse la media y la mediana. La media es básicamente el "centro de gravedad" del conjunto de datos. Si imaginamos que el histograma de los datos es un objeto sólido, el punto sobre el que podríamos equilibrarlo (como en un balancín) es la media. En cambio, la mediana es la observación intermedia, con la mitad de las observaciones más pequeñas y la otra mitad más grandes.

4.1.5 Un ejemplo de la vida real

Para intentar entender por qué hay que prestar atención a las diferencias entre la media y la mediana, veamos un ejemplo de la vida real. Como suelo burlarme de los periodistas por sus escasos conocimientos científicos y estadísticos, debo dar crédito a quienes lo merecen. Este es un excelente artículo de la web de noticias ABC[3] del 24 de septiembre de 2010:

> En las últimas semanas ejecutivos sénior del Commonwealth Bank han viajado por todo el mundo con una presentación en la que muestran cómo los precios de la vivienda en Australia y las principales relaciones precio-renta clave se comparan favorablemente con países similares. "La asequibilidad de la vivienda en realidad ha ido a la baja en los últimos cinco o seis años", afirmó Craig James, economista jefe de CommSec, la división de negociación del banco.

Esto probablemente sea una gran sorpresa para cualquiera que tenga una hipoteca, o que quiera una hipoteca, o que pague el alquiler, o que no sea completamente ajeno a lo que ha estado ocurriendo en el mercado inmobiliario australiano en los últimos años. De vuelta al artículo:

> CBA ha declarado la guerra a los que considera agoreros de la vivienda con gráficos, cifras y comparaciones internacionales. En su presentación, el banco rechaza los argumentos de que la vivienda en Australia es relativamente cara en comparación con los ingresos. Afirma que la relación entre el precio de la vivienda y la renta familiar en Australia, de 5,6 en las principales ciudades y de 4,3 en todo el país, es comparable a la de muchos otros países desarrollados. Dice que San Francisco y Nueva York tienen una relación de 7, Auckland es 6.7 y Vancouver de 9.3.

¡Más excelentes noticias! Excepto que el artículo continúa haciendo la observación de que:

> Muchos analistas afirman que eso ha llevado al banco a utilizar cifras y comparaciones engañosas. Si se va a la página cuatro de la presentación del CBA y se lee la información de la fuente en la parte inferior del gráfico y la tabla, se observará que hay una fuente adicional en la comparación internacional: Demographia. Sin embargo, si el Commonwealth Bank también hubiera utilizado el análisis de Demographia en la relación entre el precio de la vivienda y la renta en Australia, habría obtenido una cifra más cercana a 9 en lugar de 5,6 o 4,3.

Se trata de una discrepancia bastante seria. Un grupo de personas dice 9, otro dice 4-5. ¿Deberíamos dividir la diferencia y decir que la verdad está en algún punto intermedio? Por supuesto que no. Esta es una situación en la que hay una respuesta correcta y otra incorrecta. Demographia tiene razón y el Commonwealth Bank está equivocado. Como señala el artículo:

> [Un] problema obvio de las cifras de precios e ingresos del Commonwealth Bank es que comparan los ingresos medios con los precios medianos de la vivienda (a diferencia de las cifras de Demographia que comparan los ingresos medianos con los precios medianos). La mediana es el punto medio,

[3] www.abc.net.au/news/stories/2010/09/24/3021480.htm

> reduciendo eficazmente los altibajos, y eso significa que el promedio es generalmente más alto cuando se trata de ingresos y precios de activos, porque incluye los ingresos de las personas más ricas de Australia. Dicho de otro modo: las cifras del Commonwealth Bank tienen en cuenta el sueldo multimillonario de Ralph Norris en lo que respecta a los ingresos, pero no su (sin duda) carísima casa en las cifras del precio de los bienes inmuebles, con lo que subestiman la relación entre el precio de la vivienda y los ingresos de los australianos con rentas medias.

Yo no lo habría expresado mejor. La forma en que Demographia calculó la proporción es la correcta. La forma en que lo hizo el Banco es incorrecta. En cuanto a por qué una organización tan sofisticada cuantitativamente como un gran banco cometió un error tan elemental, bueno... No puedo asegurarlo ya que no tengo ningún conocimiento especial de su forma de pensar. Pero el propio artículo menciona los siguientes hechos, que pueden o no ser relevantes:

> [Como] el mayor prestamista hipotecario de Australia, el Commonwealth Bank, tiene uno de los mayores intereses creados en el aumento de los precios de la vivienda. Efectivamente, posee una gran cantidad de viviendas australianas como garantía de sus préstamos hipotecarios, así como de muchos préstamos a pequeñas empresas.

Vaya, vaya.

4.1.6 Moda

La moda de una muestra es muy sencilla. Es el valor que aparece con más frecuencia. Podemos ilustrar la moda utilizando una variable diferente de la AFL: ¿quién ha jugado más finales? Abre el archivo de finalistas de aflsmall y echa un vistazo a la variable afl.finalists, ver Figure 4.5. Esta variable contiene los nombres de los 400 equipos que jugaron en las 200 finales disputadas durante el período de 1987 a 2010.

Lo que podríamos hacer es leer las 400 entradas y contar el número de veces en las que aparece el nombre de cada equipo en nuestra lista de finalistas, produciendo así una **tabla de frecuencias**. Sin embargo, sería una tarea aburrida y sin sentido: exactamente el tipo de tarea para la que los ordenadores son excelentes. Así que usemos jamovi para que lo haga por nosotros. En 'Exploración' - 'Descriptivos', haz clic en la pequeña casilla de verificación etiquetada como 'Tablas de frecuencias' y obtendrás algo como Figure 4.6.

Ahora que tenemos nuestra tabla de frecuencias, podemos mirarla y ver que, en los 24 años de los que tenemos datos, Geelong ha jugado más finales que cualquier otro equipo. Por lo tanto, la moda de los datos de afl.finalists es "Geelong". Podemos ver que Geelong (39 finales) jugó más finales que cualquier otro equipo durante el período 1987-2010. También vale la pena señalar que en la tabla 'Descriptivos' no se calculan los resultados de Media, Mediana, Mínimo o Máximo. Esto se debe a que la variable afl.finalists es una variable de texto nominal, por lo que no tiene sentido calcular estos valores.

Una última observación sobre la moda. Aunque la moda se calcula con mayor frecuencia cuando se tienen datos nominales, porque las medias y las medianas son inútiles para ese tipo de variables, hay algunas situaciones en las que realmente se desea conocer la moda de una variable de escala ordinal, de intervalo o de escala de razón. Por ejemplo,

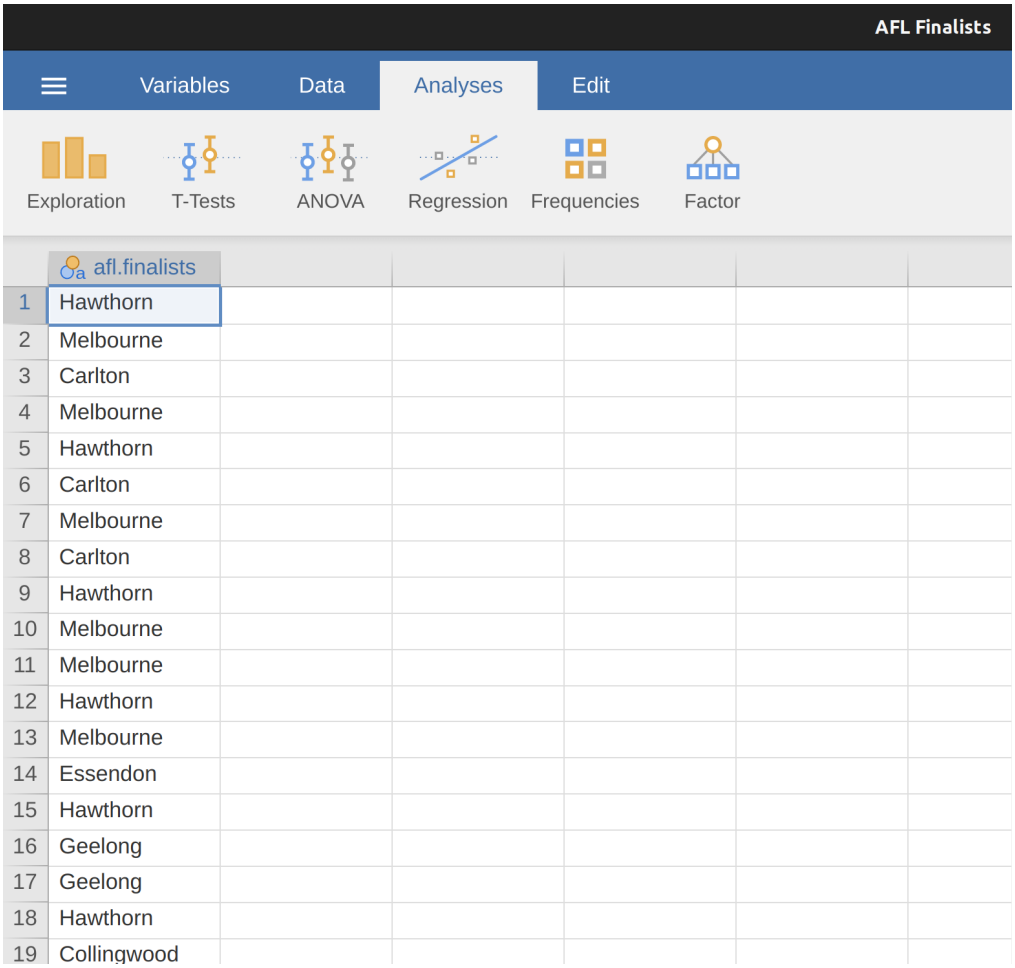


Figure 4.5: Una captura de pantalla de jamovi que muestra las variables almacenadas en el archivo aflsmall finalists.csv

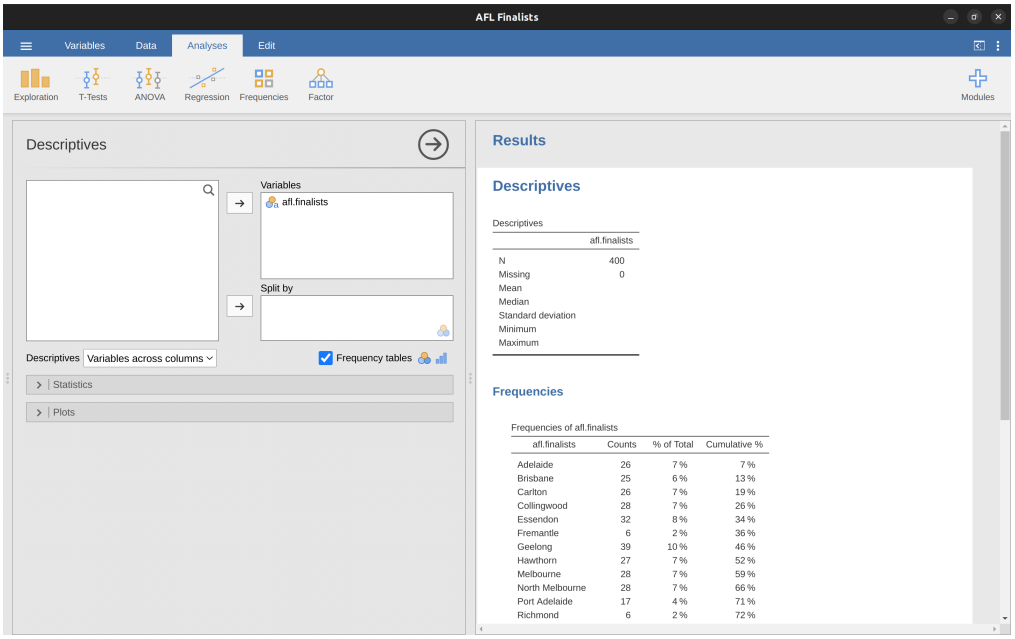


Figure 4.6: Una captura de pantalla de jamovi que muestra la tabla de frecuencias para la variable afl.finalists

volvamos a nuestra variable afl.margins. Esta variable es claramente una escala de razón (si no te queda claro, puede que te ayude volver a leer la sección sobre Section 2.2), por lo que en la mayoría de las situaciones la media o la mediana es la medida de tendencia central que quieres. Pero considera esta situación: un amigo te ofrece una apuesta y elige un partido de fútbol al azar. Sin saber quién juega, debes adivinar el margen ganador exacto. Si aciertas, ganas $50. Si no aciertas, pierdes $1. No hay premios de consolación por "casi" acertar. Tienes que acertar exactamente el margen ganador exacto. Para esta apuesta, la media y la mediana no te sirven para nada. Debes apostar por la moda. Para calcular la moda de la variable afl.margins en jamovi, vuelve a ese conjunto de datos y en la pantalla 'Exploración' - 'Descriptivos' verás que puedes ampliar la sección marcada como 'Estadísticas'. Haz clic en la casilla de verificación marcada como 'Moda' y verás el valor modal presentado en la tabla 'Descriptivos', como en Figure 4.7. Así, los datos de 2010 sugieren que deberías apostar por un margen de 3 puntos.

4.2 Medidas de variabilidad

Todos los estadísticos que hemos analizado hasta ahora se refieren a la tendencia central. Es decir, todos hablan de qué valores están "en el medio" o son "populares" en los datos. Sin embargo, la tendencia central no es el único tipo de resumen estadístico que queremos calcular. Lo segundo que realmente queremos es una medida de la **variabilidad** de los datos. Es decir, ¿cuán "dispersos" están los datos? ¿Cómo de "alejados" de la media o la mediana tienden a estar los valores observados? Por ahora, vamos a suponer que los datos son de escala de intervalo o de razón, y seguiremos utilizando los datos de afl.margins. Usaremos estos datos para discutir varias medidas diferentes de dispersión,

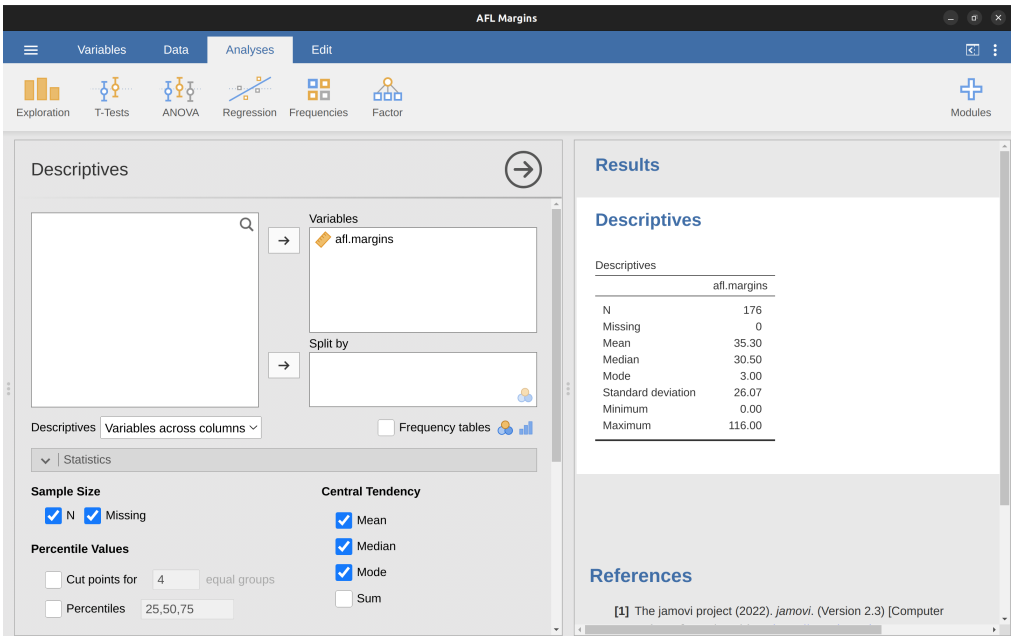


Figure 4.7: Una captura de pantalla de jamovi que muestra el valor modal para la variable afl.margins

cada una con diferentes puntos fuertes y débiles.

4.2.1 Rango

El **rango** de una variable es muy sencillo. Es el valor mayor menos el valor menor. Para los datos de márgenes ganadores de la AFL, el valor máximo es 116 y el valor mínimo es 0. Aunque el rango es la forma más sencilla de cuantificar la noción de "variabilidad", es una de las peores. Recuerda de nuestra discusión sobre la media que queremos que nuestra medida de resumen sea robusta. Si el conjunto de datos tiene uno o dos valores extremadamente malos, nos gustaría que nuestros estadísticos no se vean excesivamente influidos por estos casos. Por ejemplo, en una variable que contenga valores atípicos muy extremos

-100, 2, 3, 4, 5, 6, 7, 8, 9, 10

está claro que el rango no es robusto. Esta variable tiene un rango de 110, pero si se eliminara el valor atípico, tendríamos un rango de solo 8.

4.2.2 Rango intercuartílico

El **rango intercuartílico** (RIC) es como el rango, pero en lugar de la diferencia entre el valor mayor y el menor se toma la diferencia entre el percentil 25 y el percentil 75. Si aún no sabes qué es un **percentil**, el percentil 10 de un conjunto de datos es el número x más pequeño tal que el 10 % de los datos es menor que x. De hecho, ya nos hemos topado con la idea. La mediana de un conjunto de datos es su percentil 50. En jamovi

puedes especificar fácilmente los percentiles 25, 50 y 75 haciendo clic en la casilla de verificación 'Cuartiles' de la pantalla 'Exploración' - 'Descriptivas' - 'Estadísticas'.

Y no es sorprendente que en Figure 4.8 el percentil 50 sea el mismo que el valor de la mediana. Y, observando que \$50,50 - 12,75 = \$37,75, podemos ver que el rango intercuartílico para los datos de los márgenes ganadores de la AFL de 2010 es 37,75. Si bien es obvio cómo interpretar el rango, es un poco menos obvio cómo interpretar el RIC. La forma más sencilla de pensar en ello es la siguiente: el rango intercuartílico es el rango que abarca la "mitad central" de los datos. Es decir, una cuarta parte de los datos cae por debajo del percentil 25 y una cuarta parte de los datos está por encima del percentil 75, por lo que la "mitad central" de los datos se encuentra entre ambos. Y el RIC es el rango cubierto por esa mitad central.

4.2.3 Desviación absoluta media

Las dos medidas que hemos visto hasta ahora, el rango y el rango intercuartílico, se basan en la idea de que podemos medir la dispersión de los datos observando los percentiles de los datos. Sin embargo, esta no es la única manera de abordar el problema. Un enfoque diferente consiste en seleccionar un punto de referencia significativo (generalmente la media o la mediana) y, a continuación, informar las desviaciones "típicas" con respecto a ese punto de referencia. ¿Qué entendemos por desviación "típica"? Por lo general, se trata del valor medio o mediano de estas desviaciones. En la práctica, esto da lugar a dos medidas diferentes: la "desviación absoluta media" (de la media) y la "desviación absoluta mediana" (de la mediana). Por lo que he leído, la medida basada en la mediana parece usarse en estadística y parece ser la mejor de las dos. Pero, para ser sincera, no creo haber visto que se utilice mucho en psicología. Sin embargo, la medida basada en la media sí aparece ocasionalmente en psicología. En esta sección hablaré de la primera, y volveré a hablar de la segunda más adelante.

Como el párrafo anterior puede sonar un poco abstracto, vamos a repasar la **desviación absoluta media** de la media un poco más despacio. Un aspecto útil de esta medida es que su nombre indica exactamente cómo calcularla. Pensemos en nuestros datos de márgenes de victorias en la AFL y, una vez más, comenzaremos imaginando que solo hay 5 partidos en total, con márgenes de victorias de 56, 31, 56, 8 y 32. Dado que nuestros cálculos se basan en un examen de la desviación de algún punto de referencia (en este caso la media), lo primero que debemos calcular es la media, $\bar{X}$. Para estas cinco observaciones, nuestra media es $\bar{X} = 36,6$. El siguiente paso es convertir cada una de nuestras observaciones X_i en una puntuación de desviación. Para ello, calculando la diferencia entre la observación X_i y la media $\bar{X}$. Es decir, la puntuación de desviación se define como $X_i - \bar{X}$. Para la primera observación de nuestra muestra, esto equivale a \$56 - 36,6 = 19,4 \$. Bien, eso es bastante sencillo. El siguiente paso en el proceso es convertir estas desviaciones en desviaciones absolutas, y lo hacemos mediante la conversión de cualquier valor negativo en positivo. Matemáticamente, denotaremos el valor absoluto de -3 como $|-3|$, por lo que decimos que $|-3| = 3$. Usamos el valor absoluto aquí porque realmente no nos importa si el valor es mayor o menor que la media, solo nos interesa lo cerca que está de la media. Para ayudar a que este proceso sea lo más obvio posible, Table 4.2 muestra estos cálculos para las cinco observaciones.

Ahora que hemos calculado la puntuación de la desviación absoluta para cada observación del conjunto de datos, todo lo que tenemos que hacer es calcular la media de estas puntuaciones. Vamos a hacerlo:

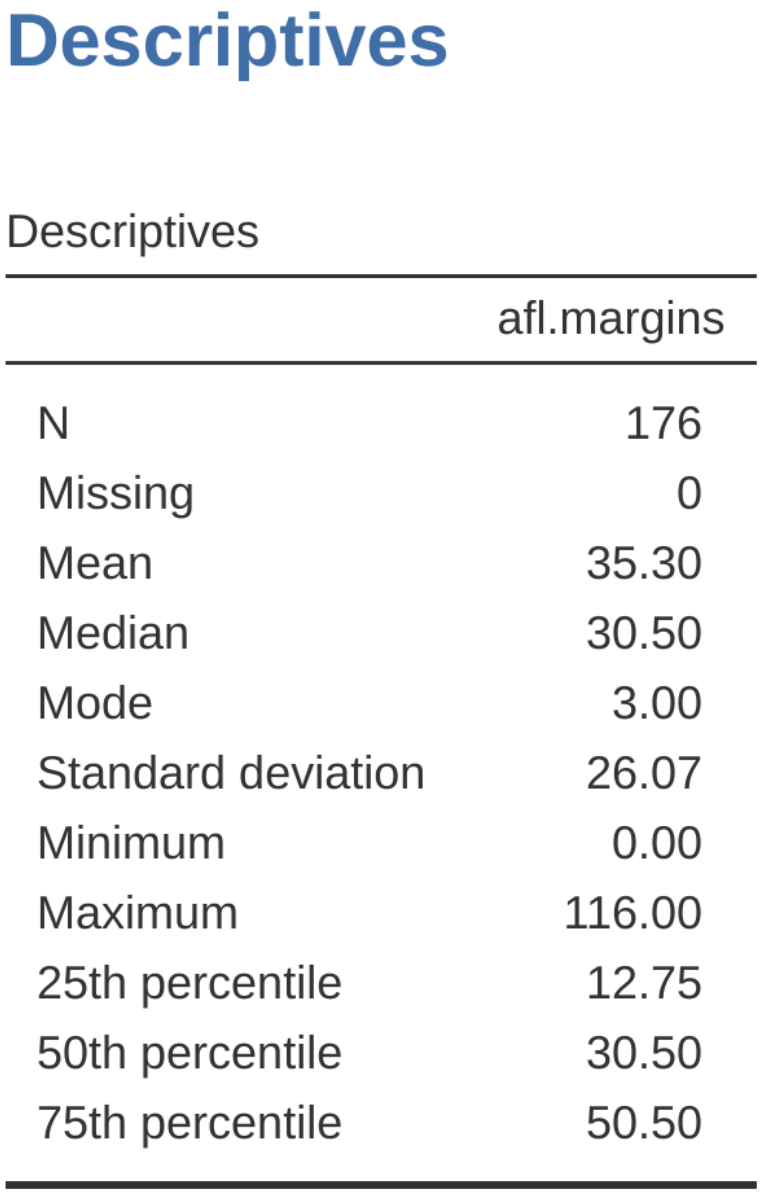


Figure 4.8: Una captura de pantalla de jamovi que muestra los cuartiles para la variable afl.margins

Table 4.2: Medidas de variabilidad

English	**notation**	**value**	**deviation from mean**	**absolute deviation**
notation:	i	X_i	$X_i - \bar{X}$	$\mid X_i - \bar{X} \mid$
	1	56	19.4	19.4
	2	31	-5.6	5.6
	3	56	19.4	19.4
	4	8	-28.6	28.6
	5	32	-4.6	4.6

$$\frac{19,4 + 5,6 + 19,4 + 28,6 + 4,6}{5} = 15,52$$

Y hemos terminado. La desviación absoluta media de estas cinco puntuaciones es 15,52.

[Detalle técnico adicional[4]]

4.2.4 Variancia

Aunque la medida de la desviación absoluta media tiene su utilidad, no es la mejor medida de variabilidad que se puede utilizar. Desde una perspectiva puramente matemática, hay algunas razones sólidas para preferir las desviaciones al cuadrado en lugar de las desviaciones absolutas. Si lo hacemos obtenemos una medida llamada **variancia**, que tiene muchas propiedades estadísticas realmente buenas que voy a ignorar,[5] y un defecto psicológico del que voy a hacer un gran problema en un momento. La variancia de un conjunto de datos X a veces se escribe como Var(X), pero se denota más comúnmente como s^2 (la razón de esto se aclarará en breve).

[4]Sin embargo, aunque nuestros cálculos para este pequeño ejemplo han llegado a su fin, nos quedan un par de cosas de las que hablar. En primer lugar, deberíamos intentar escribir una fórmula matemática adecuada. Pero para ello necesito una notación matemática para referirme a la desviación absoluta media. "Desviación absoluta media" y "desviación absoluta mediana" tienen el mismo acrónimo (MAD en inglés), lo que genera cierta ambigüedad, así que mejor me invento algo diferente para la desviación absoluta media. Lo que haré es usar DAP en su lugar, abreviatura de desviación absoluta promedio. Ahora que tenemos una notación inequívoca, esta es la fórmula que describe lo que acabamos de calcular:

$$AAD(X) = \frac{1}{N} \sum_{i=1}^{N} \mid X_i - \bar{X} \mid = 15.52$$

[5]Bueno, mencionaré muy brevemente la que me parece más guay, para una definición muy particular de "guay", claro. Las variancias son aditivas. Esto es lo que eso significa. Supongamos que tengo dos variables X y Y, cuyas variancias son $Var(X)$ y $Var(Y)$ respectivamente. Ahora imagina que quiero definir una nueva variable Z que sea la suma de las dos, $Z = X + Y$. Resulta que la variancia de Z es igual a $Var(X) + Var(Y)$. Esta es una propiedad muy útil, pero no es cierta para las otras medidas de las que hablo en esta sección.

[Detalle técnico adicional[6]]

Ahora que ya tenemos la idea básica, veamos un ejemplo concreto. Una vez más, utilizaremos como datos los cinco primeros juegos de la AFL. Si seguimos el mismo planteamiento que la última vez, obtendremos la información que se muestra en Table 4.3.

Table 4.3: medidas de variabilidad para los cinco primeros juegos de la AFL

English	**maths:**	**value**	**deviation from mean**	**absolute deviation**
notation:	i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
	1	56	19.4	376.36
	2	31	-5.6	31.36
	3	56	19.4	376.36
	4	8	-28.6	817.96
	5	32	-4.6	21.16

Esa última columna contiene todas nuestras desviaciones al cuadrado, así que todo lo que tenemos que hacer es promediarlas. Si lo hacemos a mano, es decir, con una calculadora, obtenemos una variancia de $324,64$. Emocionante, ¿verdad? Por el momento, vamos a ignorar la pregunta candente que probablemente todas estáis pensando (es decir, ¿qué diablos significa realmente una variancia de $ 324.64 $?) Y en su lugar hablemos un poco más sobre cómo hacer los cálculos en jamovi, porque esto revelará algo muy extraño. Inicia una nueva sesión de jamovi haciendo clic en el botón del menú principal (tres líneas horizontales en la esquina superior izquierda) y seleccionan 'Nuevo'. Ahora escribe los cinco primeros valores del conjunto de datos de afl.margins en la columna A (56, 31 , 56, 8, 32. Cambia el tipo de variable a 'Continua' y, en 'Descriptivas', haz clic en la casilla de verificación 'Variancia' y obtendrás los mismos valores de variancia que calculamos a mano ($324,64$). No, espera, obtienes una respuesta completamente diferente ($ 405.80 $) - mira Figure 4.9. Eso es muy raro. ¿Jamovi no funciona? ¿Es un error tipográfico? ¿Soy idiota?

La respuesta es no.[7] No es un error tipográfico y jamovi no está cometiendo un error. De hecho, es muy sencillo explicar lo que hace jamovi aquí, pero es un poco más complicado explicar por qué lo hace. Así que empecemos con el "qué". Lo que jamovi está haciendo es evaluar una fórmula ligeramente diferente a la que te mostré anteriormente. En lugar de promediar las desviaciones al cuadrado, lo que requiere dividir por el número de puntos de datos N, jamovi eligió dividir por $N - 1$.

[6] La fórmula que usamos para calcular la variancia de un conjunto de observaciones es la siguiente:

$$VAR(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Como puedes ver, es básicamente la misma fórmula que usamos para calcular la desviación absoluta media, salvo que en lugar de usar "desviaciones absolutas" usamos "desviaciones al cuadrado". Es por esta razón que la variancia a veces se denomina "desviación cuadrática media".

[7] Con la posible excepción de la tercera pregunta.

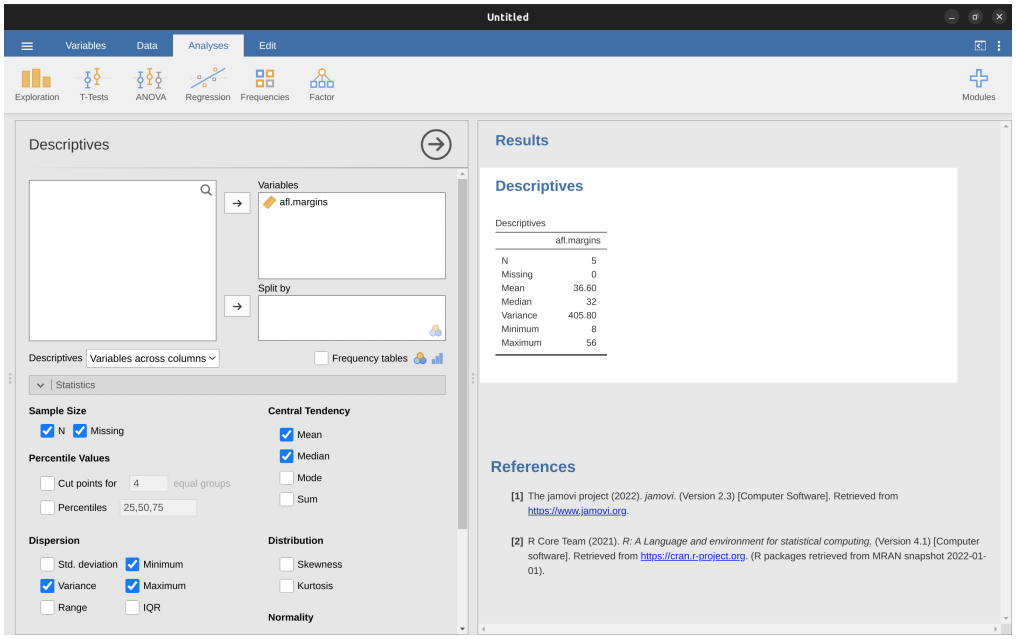


Figure 4.9: Captura de pantalla de jamovi que muestra la variancia de los 5 primeros valores de la variable afl.margins

[Detalle técnico adicional[8]]

Así que ese es el *qué*. La verdadera pregunta es por qué jamovi está dividiendo por $N - 1$ y no por N. Después de todo, se supone que la variancia es la desviación cuadrática media, ¿verdad? Entonces, ¿no deberíamos dividir por N, el número real de observaciones en la muestra? Bueno, sí, deberíamos. Sin embargo, como veremos en el capítulo sobre Chapter 8, existe una distinción sutil entre "describir una muestra" y "hacer conjeturas sobre la población de la que procede la muestra". Hasta este punto, ha sido una distinción sin diferencia. Independientemente de si se describe una muestra o se hagan inferencias sobre la población, la media se calcula exactamente igual. No ocurre lo mismo con la variancia, la desviación estándar, o muchas otras medidas. Lo que describí anteriormente (es decir, tomar la media real y dividirla por N) asume que literalmente pretendes calcular la variancia de la muestra. Sin embargo, la mayoría de las veces, no te interesa la muestra en sí misma. Más bien, la muestra existe para decirte algo sobre el mundo. Si es así, estás comenzando a alejarte del cálculo de un "estadístico muestral" y acercándote a la idea de estimar un "parámetro poblacional". Pero me estoy adelantando. Por ahora, confiemos en que jamovi sabe lo que hace, y revisaremos esta cuestión más adelante cuando hablemos de la estimación en Chapter 8.

Bien, una última cosa. Esta sección hasta ahora se ha leído un poco como una novela

[8] En otras palabras, la fórmula que está usando jamovi es esta:

$$\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

de misterio. Te he enseñado a calcular la variancia, he descrito el extraño "$N - 1$" que hace jamovi y he insinuado la razón por la que está ahí, pero no he mencionado lo más importante. ¿Cómo interpretas la variancia? Después de todo, se supone que la estadística descriptiva describe cosas, y ahora mismo la variancia no es más que un número incomprensible. Desafortunadamente, la razón por la que no te he dado una interpretación humana de la variancia es que realmente no existe. Este es el problema más grave de la variancia. Aunque tiene algunas propiedades matemáticas elegantes que sugieren que realmente es una cantidad fundamental para expresar la variación, es completamente inútil si quieres comunicarte con un ser humano real. Las variancias son completamente imposibles de interpretar en términos de la variable original. Todos los números se han elevado al cuadrado y ya no significan nada. Es un problema enorme. Por ejemplo, según Table 4.3, el margen del partido 1 fue "376,36 puntos cuadrados superior al margen medio". Esto es *exactamente* tan estúpido como suena, y por eso cuando calculamos una variancia de 324.64 estamos en la misma situación. He visto muchos partidos de fútbol y en ningún momento nadie se ha referido a "puntos al cuadrado". No es una unidad de medida real, y dado que la variancia se expresa en términos de esta unidad de galimatías, carece totalmente de sentido para un humano.

4.2.5 Desviación Estándar

De acuerdo, supongamos que te gusta la idea de utilizar la variancia por esas bonitas propiedades matemáticas de las que no he hablado, pero como eres un ser humano y no un robot, te gustaría tener una medida que se exprese en el mismas unidades que los propios datos (es decir, puntos, no puntos al cuadrado). ¿Qué debes hacer? La solución al problema es obvia. Tomar la raíz cuadrada de la variancia, conocida como **desviación estándar**, también llamada "raíz de la desviación cuadrática media". Esto resuelve nuestro problema bastante bien. Aunque nadie tiene ni idea de lo que realmente significa "una variancia de 324,68 puntos al cuadrado", es mucho más fácil entender "una desviación estándar de 18,01 puntos" ya que se expresa en las unidades originales. Es tradicional referirse a la desviación estándar de una muestra de datos como s, aunque a veces también se utilizan "de" y "desv est".

[Detalle técnico adicional[9]]

Sin embargo, como habrás adivinado por nuestra discusión sobre la variancia, lo que jamovi calcula en realidad es ligeramente diferente a la fórmula anterior. Al igual que vimos con la variancia, lo que jamovi calcula es una versión que divide por $N - 1$ en lugar de N.

[Detalle técnico adicional[10]]

[9] dado que la desviación estándar es igual a la raíz cuadrada de la variancia, probablemente no te sorprenda ver que la fórmula es:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

y en jamovi hay una casilla de verificación para 'Desv. Estándar' justo encima de la casilla de verificación de 'Variancia'. Al seleccionarla se obtiene un valor de 26.07 para la desviación estándar.

[10] Por razones que tendrán sentido cuando volvamos a este tema en el capítulo sobre [Estimación de cantidades desconocidas de una muestra] me referiré a esta nueva cantidad como $\hat{\sigma}$ (léase como: "sombrero sigma"), y la fórmula para esto es:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Interpretar las desviaciones estándar es un poco más complejo. Como la desviación estándar se obtiene a partir de la variancia, y la variancia es una cantidad que tiene poco o ningún significado para nosotros, los humanos, la desviación estándar no tiene una interpretación sencilla. En consecuencia, la mayoría de nosotras nos basamos en una simple regla empírica. En general, cabe esperar que el 68 % de los datos se sitúen dentro de 1 desviación estándar de la media, el 95 % de los datos dentro de 2 desviaciones estándar de la media y el 99,7 % de los datos dentro de 3 desviaciones estándar de la media. Esta regla suele funcionar bastante bien la mayoría de las veces, pero no es exacta. En realidad, se calcula basándose en la suposición de que el histograma es simétrico y tiene "forma de campana".[ˆ04.5] Como puedes ver en el histograma de márgenes ganadores de la AFL en Figure 4.2, esto no es exactamente cierto en nuestros datos. Aun así, la regla es aproximadamente correcta. Resulta que el 65,3 % de los datos de los márgenes de la AFL caen dentro de una desviación estándar de la media. Esto se muestra visualmente en Figure 4.10.

4.2.6 ¿Qué medida hay que utilizar?

Hemos discutido varias medidas de dispersión: rango, RIC, desviación absoluta media, variancia y desviación estándar; y hemos aludido a sus puntos fuertes y débiles. He aquí un resumen rápido:

- *Rango.* Proporciona la dispersión completa de los datos. Es muy vulnerable a los valores atípicos y, en consecuencia, no se suele utilizar a menos que haya buenas razones para preocuparse por los extremos de los datos.
- *Rango intercuartílico.* Indica dónde se sitúa la "mitad central" de los datos. Es bastante robusto y complementa muy bien a la mediana. Se usa mucho.
- *Desviación media absoluta.* Indica la distancia "media" entre las observaciones y la media. Es muy interpretable pero tiene algunos problemas menores (no discutidos aquí) que la hacen menos atractiva para los estadísticos que la desviación estándar. Se usa a veces, pero no a menudo.
- *Variancia.* Indica la desviación media al cuadrado de la media. Es matemáticamente elegante y probablemente la forma "correcta" de describir la variación alrededor de la media, pero es completamente ininterpretable porque no usa las mismas unidades que los datos. Casi nunca se usa excepto como una herramienta matemática, pero está oculta "bajo el capó" de una gran cantidad de herramientas estadísticas.
- *Desviación Estándar.* Es la raíz cuadrada de la variancia. Es bastante elegante desde el punto de vista matemático y se expresa en las mismas unidades que los datos, por lo que se puede interpretar bastante bien. En situaciones en las que la media es la medida de tendencia central, esta es la medida por defecto. Es, con mucho, la medida de variación más popular.

En resumen, el RIC y la desviación estándar son fácilmente las dos medidas más utilizadas para informar de la variabilidad de los datos. Pero hay situaciones en las que se utilizan las otras. Las he descrito todas en este libro porque es muy probable que te encuentres con la mayoría de ellas en alguna parte.

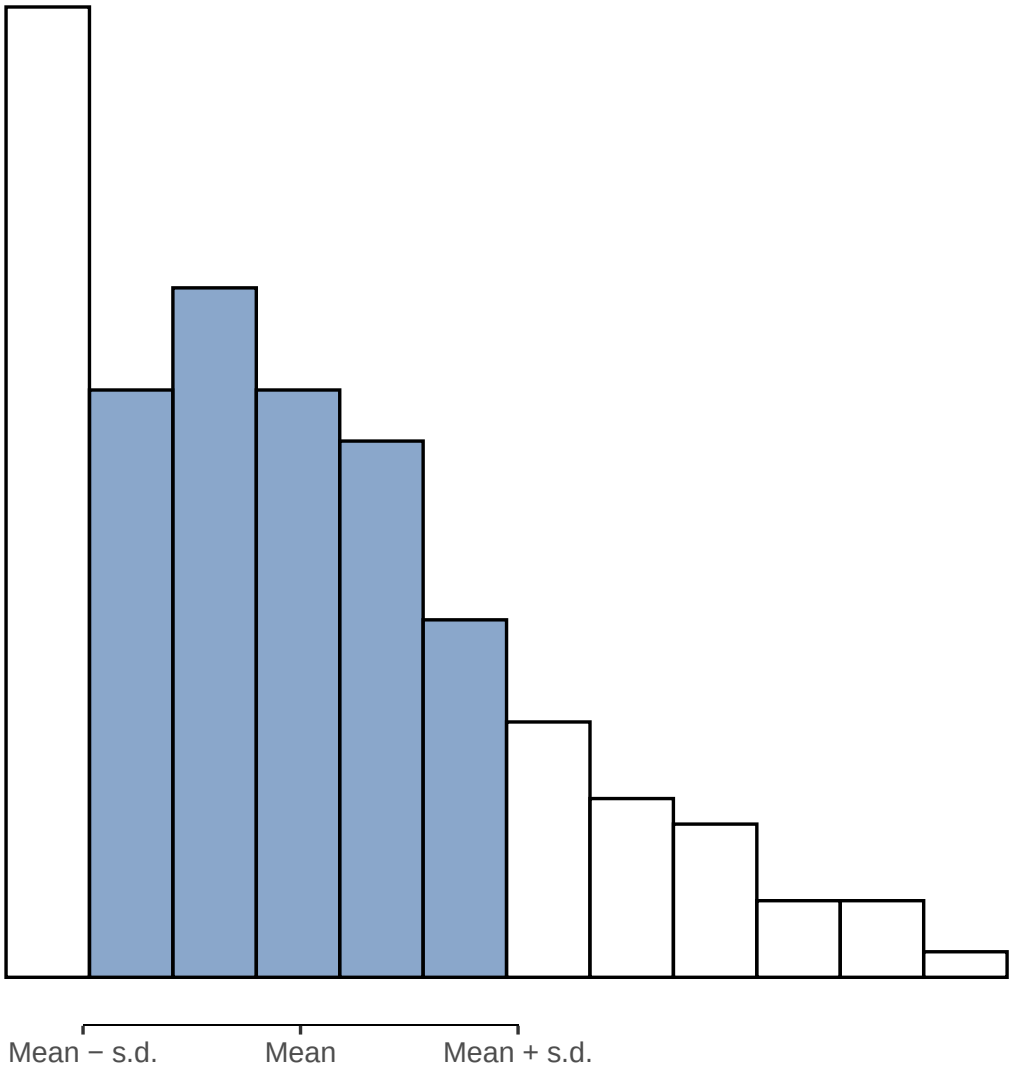


Figure 4.10: Ilustración de la desviación estándar de los datos de los márgenes ganadores de la AFL. Las barras sombreadas del histograma muestran la proporción de datos que se sitúan dentro de una desviación estándar de la media. En este caso, el 65,3 % del conjunto de datos se encuentra dentro de este intervalo, lo que es bastante consistente con la "regla de aproximadamente el 68%" comentada en el texto principal.

4.3 Asimetría y apuntamiento

Hay otros dos estadísticos descriptivos que a veces aparecen en la literatura psicológica: la asimetría y el apuntamiento. En la práctica, ninguno de los dos se usa con tanta frecuencia como las medidas de tendencia central y variabilidad de las que hemos hablado. La asimetría es bastante importante, por lo que se menciona a menudo, pero nunca he visto el apuntamiento en un artículo científico hasta la fecha.

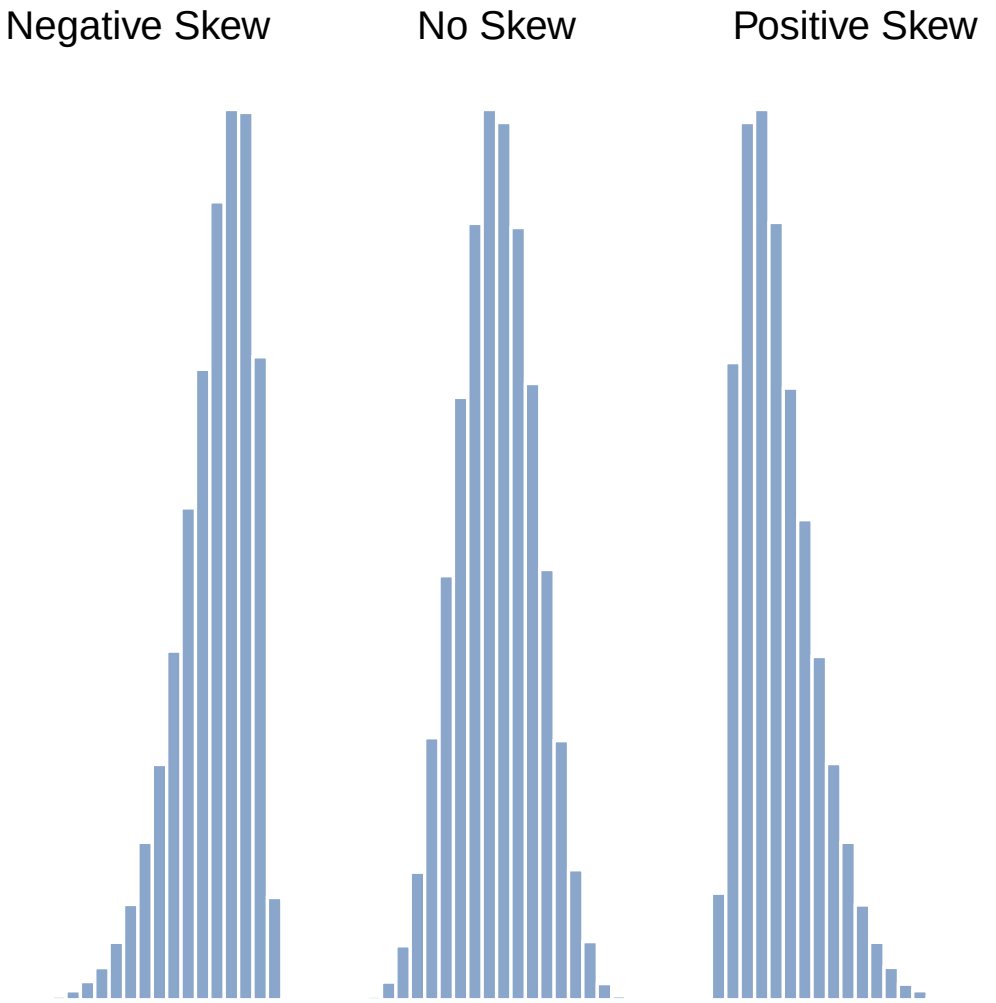


Figure 4.11: Ilustración de la asimetría. A la izquierda tenemos un conjunto de datos con asimetría negativa, en el medio tenemos un conjunto de datos sin asimetría y a la derecha tenemos un conjunto de datos con asimetría positiva

Como es el más interesante de los dos, comencemos hablando de la **asimetría**. La asimetría es básicamente una medida de asimetría y la forma más fácil de explicarla es haciendo algunos dibujos. Como ilustra Figure 4.11, si los datos tienden a tener muchos valores extremadamente pequeños (es decir, la cola inferior es "más larga" que la cola superior) y no tantos valores extremadamente grandes (panel izquierdo), entonces

decimos que los datos presentan *asimetría negativa*. En cambio, si hay más valores extremadamente grandes que extremadamente pequeños (panel derecho), decimos que los datos presentan asimetría positiva. Esa es la idea cualitativa que subyace a la asimetría. Si hay relativamente más valores que son muy superiores a la media, la distribución tiene una asimetría positiva o una asimetría hacia la derecha, con una cola que se extiende hacia la derecha. La asimetría negativa o izquierda es lo contrario. Una distribución simétrica tiene una asimetría de 0. El valor de asimetría para una distribución con asimetría positiva es positivo y el valor para una distribución con asimetría negativa es negativo.

[Detalle técnico adicional[11]]

Quizás sea más útil usar jamovi para calcular la asimetría: es una casilla de verificación en las opciones de 'Estadísticas' en 'Exploración' - 'Descriptivos'. Para la variable afl.margins, la asimetría es de 0.780. Si divides la estimación de la asimetría por el error estándar de la asimetría, tendrás una indicación del grado de asimetría de los datos. Especialmente en muestras pequeñas (N $<$ 50), una regla general sugiere que un valor de 2 o menos puede significar que los datos no son muy asimétricos, y un valor de más de 2 sugiere que hay suficiente asimetría en los datos para posiblemente limitar su uso en algunos análisis estadísticos. Aunque no hay un acuerdo claro sobre esta interpretación. Dicho esto, esto indica que los datos de márgenes ganadores de la AFL son algo asimétricos ($\frac{0.780}{0.183} = 4.262$).

La última medida a la que a veces se hace referencia, aunque muy raramente en la práctica, es el apuntamiento de un conjunto de datos. En pocas palabras, el apuntamiento es una medida de lo delgadas o gruesas que son las colas de una distribución, como se ilustra en Figure 4.12. Por convención, decimos que la "curva normal" (líneas negras) tiene apuntamiento cero, por lo que el grado de apuntamiento se evalúa en relación con esta curva.

En esta figura, los datos de la izquierda tienen una distribución bastante plana, con colas finas, por lo que el apuntamiento es negativo y decimos que los datos son platicúrticos. Los datos de la derecha tienen una distribución con colas gruesas, por lo que el apuntamiento es positivo y decimos que los datos son leptocúrticos. Pero los datos del medio no tienen colas gruesas ni gordas, por lo que decimos que son mesocúrticos y tienen apuntamiento cero. Esto se resume en Table 4.4:

Table 4.4: Colas finas a gruesas para ilustrar la curtosis

English	**informal term**	**kurtosis value**
"tails too thin"	platykurtic	negative
"tails neither thin or fat"	mesokurtic	zero
"tails too fat"	leptokurtic	positive

[11] Una fórmula para la asimetría de un conjunto de datos es la siguiente

$$asimetra(X) = \frac{1}{N\hat{\sigma}^3} \sum_{i=1}^{N} (X_i - \bar{X})^3$$

donde N es el número de observaciones, $\bar{X}$ es la media muestral y $\hat{\sigma}$ es la desviación estándar (es decir, la versión "dividir por $N-1$").

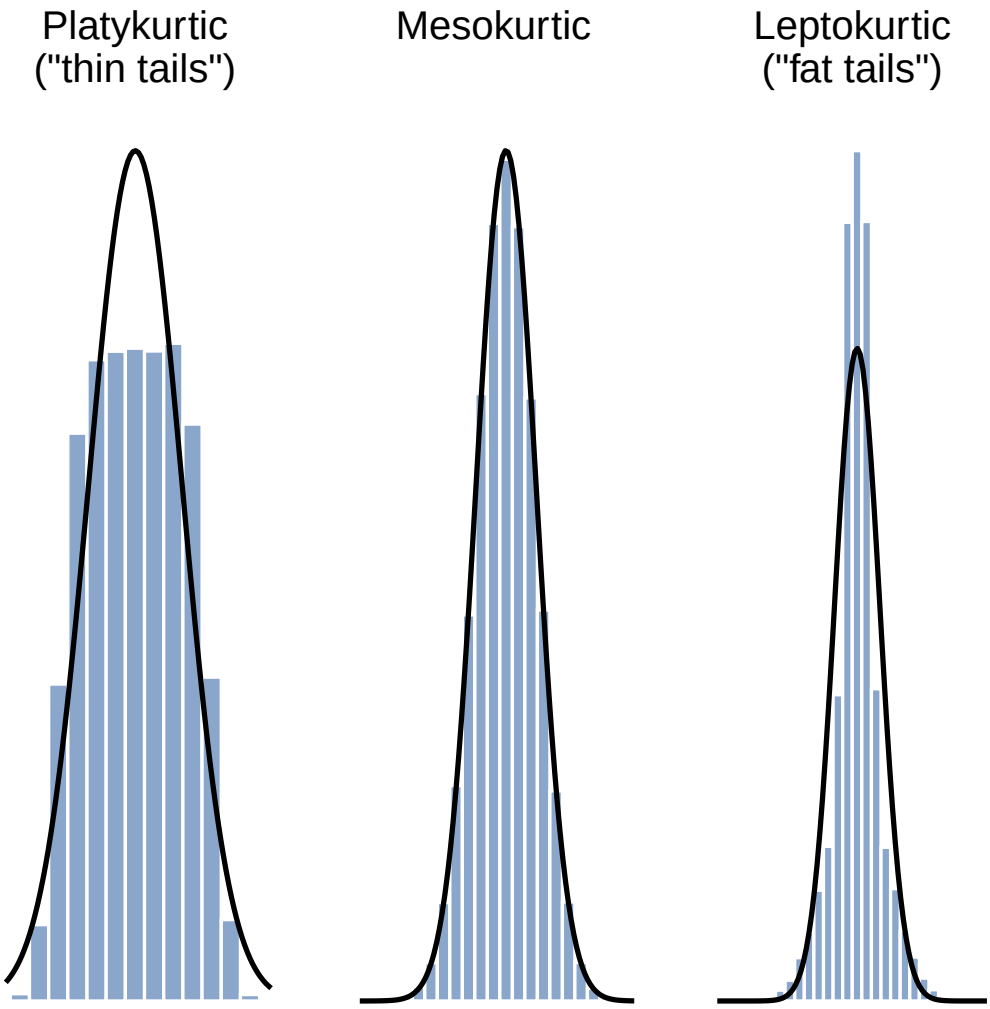


Figure 4.12: Un ejemplo del apuntamiento. A la izquierda, tenemos una distribución 'platicúrtica' (curtosis = -0,95), lo que significa que la distribución tiene colas 'finas' o planas. En el centro tenemos una distribución 'mesocúrtica' (la curtosis es casi exactamente 0) lo que significa que las colas no son ni finas ni gruesas. Finalmente, a la derecha, tenemos una distribución 'leptocúrtica' (curtosis = 2.12) que indica que la distribución tiene colas 'gordas'. Observa que el apuntamiento se mide con respecto a una curva normal (línea negra)

[Detalle técnico adicional[12]]

Más concretamente, jamovi tiene una casilla de verificación para el apuntamiento justo debajo de la casilla de verificación para la asimetría, y esto da un valor para el apuntamiento de 0.101 con un error estándar de 0.364. Esto significa que los datos de márgenes ganadores de la AFL tienen solo un pequeño apuntamiento, lo cual está bien.

4.4 Estadísticos descriptivos para cada grupo

Es muy frecuente encontrarse con la necesidad de consultar estadísticos descriptivos desglosados por alguna variable de agrupación. Esto es bastante fácil de hacer en jamovi. Por ejemplo, supongamos que quiero ver los estadísticos descriptivos de algunos datos de ensayos clínicos, desglosados por separado según el tipo de terapia. Se trata de un nuevo conjunto de datos, que no habías visto antes. Los datos están almacenados en el archivo Clinicaltrial.csv y los usaremos mucho más adelante en Chapter 13 (puedes encontrar una descripción completa de los datos al principio de ese capítulo). Vamos a cargarlo y ver lo que tenemos (Figure 4.13):

Evidentemente, había tres fármacos: un placebo, algo llamado "anxifree" y algo llamado "joyzepam", y cada fármaco se le administró a 6 personas. Hubo 9 personas tratadas con terapia cognitiva conductual (TCC) y 9 personas que no recibieron tratamiento psicológico. Y podemos ver al mirar las 'Descriptivas' de la variable mood.gain que la mayoría de las personas mostraron una mejora en el estado de ánimo ($media = 0.88$), aunque sin saber cuál es la escala aquí es difícil decir mucho más que eso. Aún así, no está nada mal. En general, creo que he aprendido algo.

También podemos seguir adelante y ver otros estadísticos descriptivos, y esta vez por separado para cada tipo de terapia. En jamovi, marca Desviación Estándar, Asimetría y Apuntamiento en las opciones de 'Estadísticas'. Al mismo tiempo, mueve la variable de terapia al cuadro 'Dividir por' y deberías obtener algo como Figure 4.14.

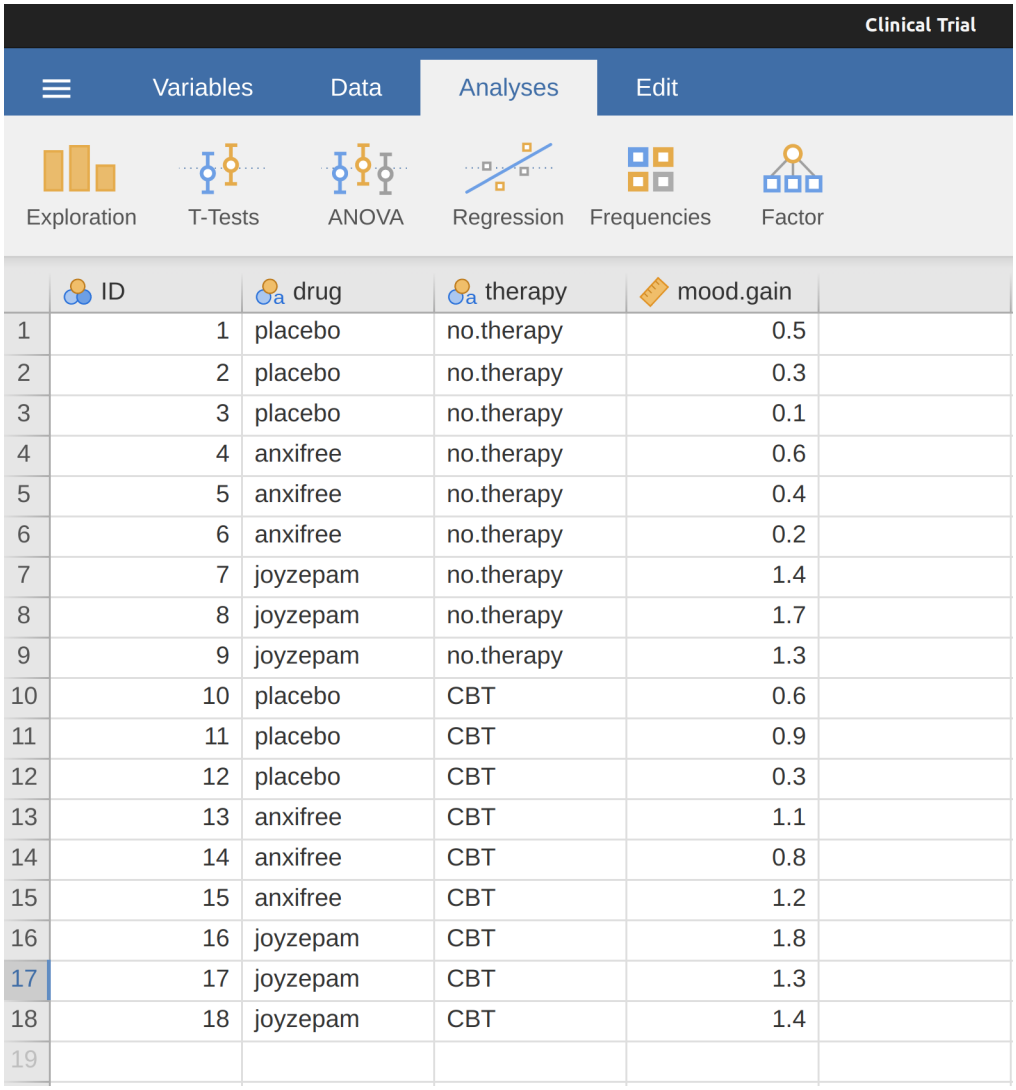
¿Qué ocurre si tienes múltiples variables de agrupación? Supongamos que deseas observar el aumento medio del estado de ánimo por separado para todas las combinaciones posibles de fármaco y terapia. Es posible hacerlo añadiendo otra variable, fármaco, en el cuadro 'Dividir por'. Fácil, aunque a veces si divides demasiado no hay suficientes datos en cada combinación de desglose para hacer cálculos significativos. En este caso, jamovi lo indica diciendo algo como NaN o Inf. [13]

[12] La ecuación para el apuntamiento es bastante similar en espíritu a las fórmulas que ya hemos visto para la variancia y la asimetría. Salvo que donde la variancia incluía desviaciones al cuadrado y la asimetría incluía desviaciones al cubo, la curtosis implica elevar las desviaciones a la cuarta potencia: [b]

$$curtosis(X) = \frac{1}{N\hat{\sigma}^4}\sum_{i=1}^{N}(X_i - \bar{X})^4 - 3$$

Lo sé, a mí tampoco me interesa mucho. — [b] El "-3" es algo que los estadísticos añaden para asegurarse de que la curva normal tenga un apuntamiento cero. Parece un poco estúpido poner un "-3" al final de la fórmula, pero existen buenas razones matemáticas para hacerlo.

[13] A veces, jamovi también presenta los números de una forma inusual. Si un número es muy pequeño, o muy grande, jamovi cambia a una forma exponencial. Por ejemplo, 6,51e-4 es lo mismo que decir que el punto decimal se mueve 4 posiciones a la izquierda, por lo que el número real es 0,000651. Si hay un signo más (es decir, 6,51e+4), entonces el punto decimal se desplaza hacia la derecha, es decir, 65.100,00. Normalmente, solo se expresan de este modo los números muy pequeños o muy grandes, por ejemplo, 6,51e-16, que sería bastante difícil de escribir de la manera normal.



	ID	drug	therapy	mood.gain
1	1	placebo	no.therapy	0.5
2	2	placebo	no.therapy	0.3
3	3	placebo	no.therapy	0.1
4	4	anxifree	no.therapy	0.6
5	5	anxifree	no.therapy	0.4
6	6	anxifree	no.therapy	0.2
7	7	joyzepam	no.therapy	1.4
8	8	joyzepam	no.therapy	1.7
9	9	joyzepam	no.therapy	1.3
10	10	placebo	CBT	0.6
11	11	placebo	CBT	0.9
12	12	placebo	CBT	0.3
13	13	anxifree	CBT	1.1
14	14	anxifree	CBT	0.8
15	15	anxifree	CBT	1.2
16	16	joyzepam	CBT	1.8
17	17	joyzepam	CBT	1.3
18	18	joyzepam	CBT	1.4
19				

Figure 4.13: Captura de pantalla de jamovi que muestra las variables almacenadas en el archivo clinictrial.csv

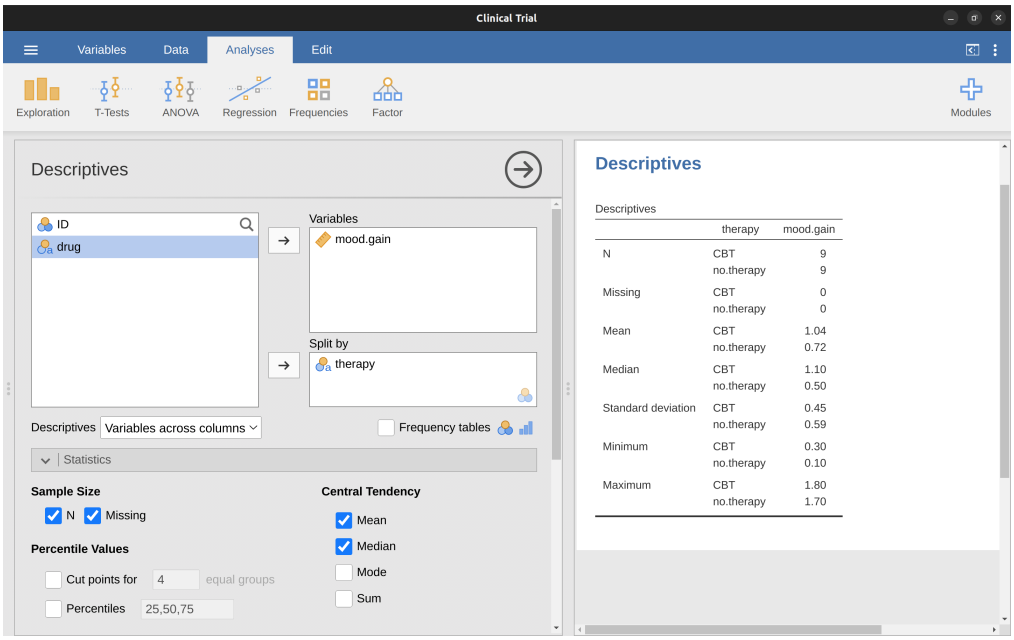


Figure 4.14: Captura de pantalla de jamovi que muestra descriptivos divididos por tipo de terapia

4.5 Puntuaciones estándar

Supongamos que mi amigo está elaborando un nuevo cuestionario destinado a medir el "mal humor". La encuesta tiene 50 preguntas a las que puedes responder de forma malhumorada o no. En una muestra grande (hipotéticamente, imaginemos un millón de personas más o menos) los datos se distribuyen de forma bastante normal: la puntuación media de mal humor es de 17 de cada $50 preguntas respondidas de forma malhumorada, y la desviación estándar es de 5. En cambio, cuando hago el cuestionario respondo 35 de 50 preguntas de una forma malhumorada. Entonces, ¿hasta qué punto soy gruñona? Una forma de verlo sería decir que tengo un mal humor de $\frac{35}{50}$, por lo que podría decir que soy un 70% gruñona. Pero eso es un poco raro, si lo piensas. Si mi amigo hubiera formulado sus preguntas de otra manera, la gente podría haberlas respondido de una manera diferente, por lo que la distribución general de las respuestas podría subir o bajar dependiendo de la forma precisa en que se formularan las preguntas. Por lo tanto, solo tengo un 70% de mal humor *con respecto a este conjunto de preguntas de la encuesta*. Incluso si fuera un cuestionario muy bueno, no es una afirmación muy informativa.

Una forma más sencilla de describir mi mal humor es compararme con otras personas. Sorprendentemente, de la muestra de mi amigo de $1,000,000$ de personas, solo 159 personas eran tan malhumoradas como yo (lo cual no es nada irreal, francamente) lo que sugiere que estoy en el 0.016% superior de la gente malhumorada. Esto tiene mucho más sentido que intentar interpretar los datos en bruto. Esta idea, la de que deberíamos describir mi mal humor en términos de la distribución general del mal humor de los seres humanos, es la idea cualitativa a la que intenta llegar la estandarización. Una forma de hacerlo es hacer exactamente lo que acabo de mostrar y describirlo todo en términos de

percentiles. Sin embargo, el problema es que "estás solo en la cima". Supongamos que mi amigo solo había recogido una muestra de 1000 personas (todavía es una muestra bastante grande a efectos de probar un nuevo cuestionario, me gustaría añadir), y esta vez hubiera obtenido, digamos, una media de 16 sobre 50 con una desviación estándar de 5. El problema es que es casi con toda seguridad ni una sola persona en esa muestra sería tan gruñona como yo.

Sin embargo, no todo está perdido. Un enfoque diferente es convertir mi puntuación de mal humor en una **puntuación estándar**, también conocida como puntuación z. La puntuación estándar se define como el número de desviaciones estándar por encima de la media en la que se encuentra mi puntuación de mal humor. Para expresarlo en "pseudomatemáticas", la puntuación estándar se calcula así:

$$\text{puntaje estándar} = \frac{\text{puntaje bruto} - \textit{media}}{\text{desviación estándar}}$$

[Detalle técnico adicional[14]]

Así que, volviendo a los datos de mal humor, ahora podemos transformar el mal humor en bruto de Dani en una puntuación de mal humor estandarizada.

$$z = \frac{35 - 17}{5} = 3,6$$

Para interpretar este valor, recuerda la heurística aproximada que proporcioné en la sección sobre Desviación estándar en la que señalé que se espera que el 99,7 % de los valores se encuentran dentro de 3 desviaciones estándar de la media. Así que el hecho de que mi mal humor corresponda a una puntuación z de 3,6 indica que soy muy gruñona. De hecho, esto sugiere que soy más gruñóna que el 99,98% de las personas. Me parece correcto.

Además de permitirte interpretar una puntuación bruta en relación con una población más amplia (y, por tanto, darle sentido a variables que se sitúan en escalas arbitrarias), las puntuaciones estándar cumplen una segunda función útil. Las puntuaciones estándar se pueden comparar entre sí en situaciones en las que las puntuaciones brutas no pueden. Supongamos, por ejemplo, que mi amigo también tiene otro cuestionario que mide la extraversión utilizando un cuestionario de 24 ítems. La media general de esta medida resulta ser 13 con una desviación estándar de 4 y yo obtuve una puntuación de 2. Como puedes imaginar, no tiene mucho sentido intentar comparar mi puntuación bruta de 2 en el cuestionario de extraversión con mi puntuación bruta de 35 en el cuestionario de mal humor. Las puntuaciones brutas para las dos variables son "sobre" cosas fundamentalmente diferentes, así que sería como comparar manzanas con naranjas.

¿Y las puntuaciones estándar? Bueno, esto es un poco diferente. Si calculamos las puntuaciones estándar obtenemos ($z = \frac{(35-17)}{5} = 3,6$) para el mal humor y ($z = \frac{(2-13)}{4} = -2,75$) para la extraversión. Estos dos números se pueden comparar entre sí.[15] Soy

[14] En matemáticas reales, la ecuación para la puntuación z es

$$z_i = \frac{X_i - \bar{X}}{\hat{\sigma}}$$

[15] Aunque suele estar justificada con cautela. No siempre se da el caso de que una desviación estándar en la variable A corresponda al mismo "tipo" de cosas que una desviación estándar en la variable B. Usa

mucho menos extrovertida que la mayoría de la gente ($z = -2,75$) y mucho más gruñóna que la mayoría de la gente ($z = 3,6$). Pero el alcance de mi rareza es mucho más extremo en el caso del mal humor, ya que $3,6$ es un número mayor que $2,75$. Dado que cada puntuación estandarizada es una afirmación sobre el lugar que ocupa una observación en relación con su propia población, es posible comparar puntuaciones estandarizados entre variables completamente diferentes.

4.6 Resumen

Calcular algunos estadísticos descriptivos básicos es una de las primeras cosas que se hacen cuando se analizan datos reales, y los estadísticos descriptivos son mucho más sencillos de entender que los estadísticaos inferenciales, así que, como cualquier otro libro de texto de estadística, he empezado con los descriptivos. En este capítulo, hablamos de los siguientes temas:

- Medidas de tendencia central. En términos generales, las medidas de tendencia central indican dónde se encuentran los datos. Hay tres medidas que suelen aparecer en la literatura: la media, la mediana y la moda.
- Medidas de variabilidad. Por el contrario, las medidas de variabilidad indican la "dispersión" de los datos. Las medidas clave son: rango, desviación estándar y rango intercuartílico.
- Asimetría y apuntamiento. También analizamos la asimetría en la distribución de una variable y las distribuciones con colas finas o gruesas (apuntamiento).
- Estadísticos descriptivos para cada grupo. Dado que este libro se centra en el análisis de datos en jamovi, dedicamos un poco de tiempo a hablar de cómo se calculan los estadísticos descriptivos para los diferentes subgrupos.
- Puntuaciones estándar. La puntuación z es una bestia un poco inusual. No es exactamente un estadístico descriptivo ni tampoco una inferencia. Asegúrate de entender esta sección. Volverá a aparecer más adelante.

En el próximo capítulo hablaremos de cómo hacer dibujos. A todo el mundo le gustan los dibujos bonitos, ¿verdad? Pero antes quiero terminar con un punto importante. Un primer curso tradicional de estadística dedica solo una pequeña parte de la clase a la estadística descriptiva, tal vez una o dos clases como mucho. La inmensa mayoría del tiempo del profesorado se dedica a la estadística inferencial porque ahí es donde está todo lo difícil. Eso tiene sentido, pero oculta la importancia práctica cotidiana de elegir buenos descriptivos. Teniendo esto en cuenta…

el sentido común cuando intentes determinar si las puntuaciones z de dos variables se pueden comparar significativamente o no.

Chapter 5

Dibujando gráficos

> *Sobre todo mostrar los datos.*
> – Edward Tufte[1]

La visualización de datos es una de las tareas más importantes a las que se enfrenta el analista de datos. Es importante por dos razones distintas pero estrechamente relacionadas. En primer lugar, está la cuestión de dibujar "gráficos de presentación", mostrar tus datos de una manera limpia y visualmente atractiva hace que sea más fácil para el lector entender lo que estás tratando de decirles. Igualmente importante, quizás aún más importante, es el hecho de que dibujar gráficos te ayuda a comprender los datos. Con ese fin, es importante dibujar "gráficos exploratorios" que te ayuden a aprender sobre los datos a medida que los analizas. Estos puntos pueden parecer bastante obvios, pero no puedo contar la cantidad de veces que he visto a la gente olvidarlos.

Para dar una idea de la importancia de este capítulo, quiero comenzar con una ilustración clásica de cuán poderoso puede ser un buen gráfico. Con ese fin, Figure 5.1 muestra un nuevo dibujo de una de las visualizaciones de datos más famosas de todos los tiempos. Este es el mapa de muertes por cólera de John Snow de 1854. El mapa es elegante en su simplicidad. De fondo tenemos un callejero que ayuda a orientar al espectador. En la parte superior vemos una gran cantidad de pequeños puntos, cada uno de los cuales representa la ubicación de un caso de cólera. Los símbolos más grandes muestran la ubicación de las bombas de agua, etiquetadas por su nombre. Incluso la inspección más casual del gráfico deja muy claro que la fuente del brote es casi con certeza la bomba de Broad Street. Al ver este gráfico, el Dr. Snow hizo arreglos para quitar el mango de la bomba y puso fin al brote que había matado a más de 500 personas. Tal es el poder de una buena visualización de datos.

Los objetivos de este capítulo son dos. Primero, discutir varios gráficos bastante estándar que usamos mucho al analizar y presentar datos, y segundo, mostrarte cómo crear estos gráficos en jamovi. Los gráficos en sí tienden a ser bastante sencillos, por lo que, en cierto sentido, este capítulo es bastante simple. Donde la gente suele tener dificultades es en aprender a producir gráficos y, especialmente, aprender a producir buenos gráficos. Afortunadamente, aprender a dibujar gráficos en jamovi es razonablemente

[1] El origen de esta cita es el encantador libro de Tufte *The Visual Display of Quantitative Information.*

simple, siempre y cuando no seas demasiado exigente con el aspecto de tu gráfico. Lo que quiero decir cuando digo esto es que jamovi tiene muchos gráficos predeterminados muy buenos, o tramas, que la mayoría de las veces producen un gráfico limpio y de alta calidad. Sin embargo, si deseas hacer algo no estándar, o si necesitas realizar cambios muy específicos en la figura, la funcionalidad gráfica en jamovi aún no es capaz de admitir trabajos avanzados o edición de detalles.

5.1 Histogramas

Comencemos con el humilde **histograma**. Los histogramas son una de las formas más sencillas y útiles de visualizar datos. Tienen más sentido cuando tienes una variable de escala de intervalo o razón (p. ej., los datos de afl.margins de Chapter 4 y lo que quieres hacer es obtener una impresión general de la variable. La mayoría probablemente sabéis cómo funcionan los histogramas). Funcionan, ya que se usan mucho, pero para que estén completos, los describiré. Todo lo que debes hacer es dividir los valores posibles en **contenedores** y luego contar el número de observaciones que caen dentro de cada contenedor. Este conteo se conoce como la frecuencia o densidad del contenedor y se muestra como una barra vertical.En los datos de márgenes ganadores de la AFL, hay 33 juegos en los que el margen ganador fue inferior a 10 puntos y es este hecho el que está representado por la altura de la barra más a la izquierda que mostramos anteriormente en Chapter 4, Figure 4.2. Con los gráficos anteriores, usamos un paquete de trazado avanzado en R que, por ahora, va más allá de la capacidad de jamovi. Pero jamovi nos acerca, y dibujar este histograma en jamovi es bastante sencillo. Abre las opciones de 'gráficos' en 'Exploración' - 'Descriptivas' y haz clic en la casilla de verificación 'histograma', como en Figure 5.1. jamovi por defecto etiqueta el eje y como 'densidad' y el eje x con el nombre de la variable. Los **contenedores** se seleccionan automáticamente y no hay información de escala o conteo en el eje y, a diferencia de la Figure 4.2 anterior. Pero esto no importa demasiado porque después de todo lo que realmente nos interesa es nuestra impresión de la forma de la distribución: ¿se distribuye normalmente o hay sesgo o curtosis? Nuestras primeras impresiones de estas características provienen de dibujar un **histograma**.

Una característica adicional que proporciona jamovi es la capacidad de trazar una curva de 'Densidad'. Puedes hacer esto haciendo clic en la casilla de verificación 'Densidad' debajo de las opciones de 'Gráficos' (y desmarcando 'Histograma'), y esto nos da el gráfico que se muestra en Figure 5.3. Un gráfico de densidad visualiza la distribución de datos en un intervalo continuo o período de tiempo. Este gráfico es una variación de un histograma que usa **suavizado de kernel** para trazar valores, lo que permite distribuciones más suaves al suavizar el ruido. Los picos de una gráfica de densidad ayudan a mostrar dónde se concentran los valores en el intervalo. Una ventaja que tienen los gráficos de densidad sobre los histogramas es que son mejores para determinar la forma de distribución porque no se ven afectados por la cantidad de contenedores utilizados (cada barra utilizada en un histograma típico). Un histograma compuesto por solo 4 contenedores no produciría una forma de distribución lo suficientemente distinguible como lo haría un histograma de 20 contenedores. Sin embargo, con gráficos de densidad, esto no es un problema.

Aunque esta imagen necesitaría mucha limpieza para hacer un buen gráfico de presentación (es decir, uno que incluirías en un informe), hace un buen trabajo al describir los datos. De hecho, la gran fortaleza de un histograma o gráfico de densidad es que (uti-

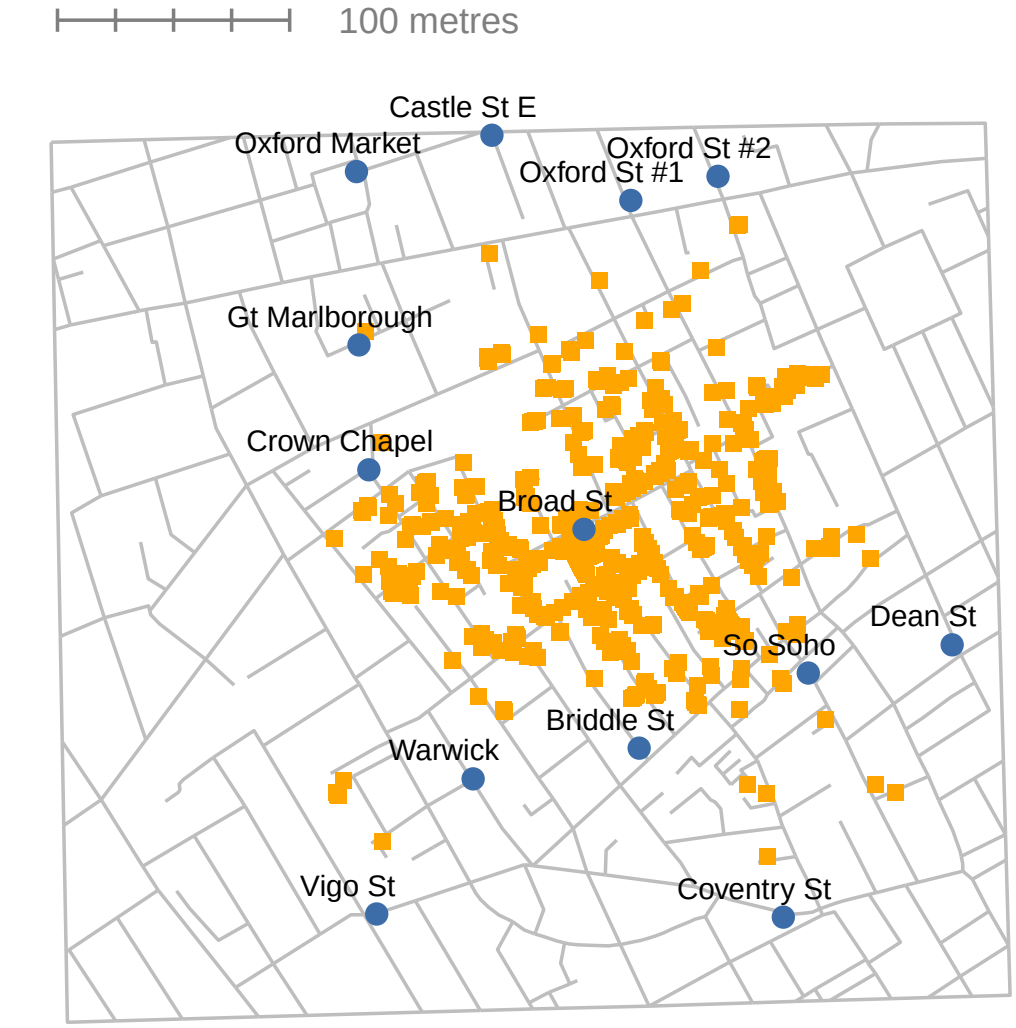


Figure 5.1: Redibujo estilizado del mapa original del cólera de John Snow. Cada pequeño cuadrado naranja representa la ubicación de una muerte por cólera y cada círculo azul muestra la ubicación de una bomba de agua. Como se aprecia claramente en el gráfico, el brote de cólera se concentra en el surtidor de la calle Broad.

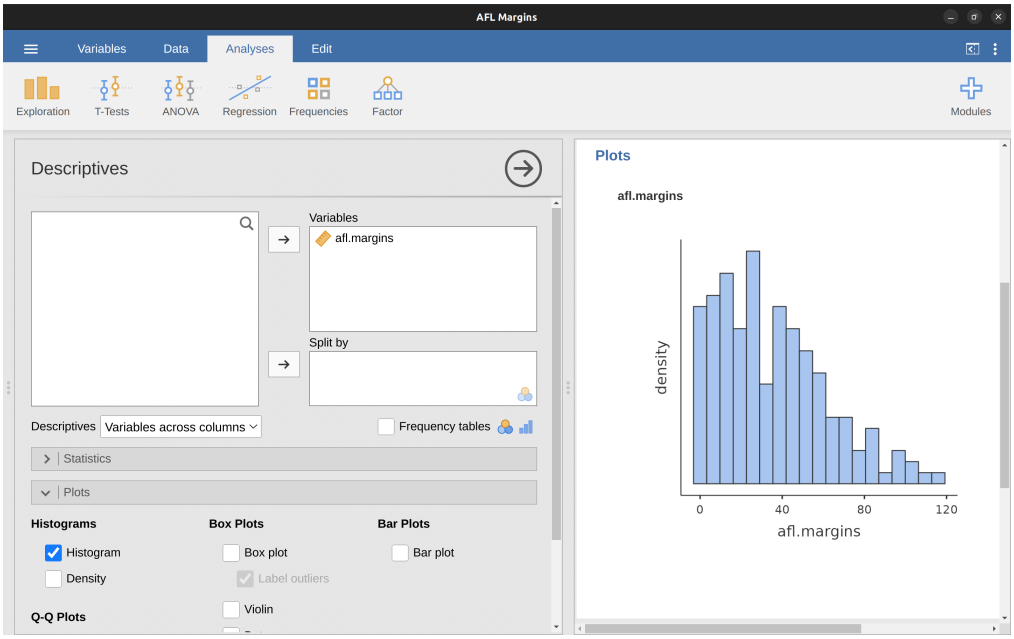


Figure 5.2: pantalla jamovi que muestra la casilla de verificación del histograma

lizado correctamente) muestra la distribución completa de los datos, por lo que puedes tener una idea bastante clara de cómo se ve. La desventaja de los histogramas es que no son muy compactos. A diferencia de algunas de las otras tramas de las que hablaré, es difícil meter 20-30 histogramas en una sola imagen sin abrumar al espectador. Y, por supuesto, si tus datos son de escala nominal, los histogramas son inútiles.

5.2 Diagramas de caja

Otra alternativa a los histogramas es un **diagrama de caja**, a veces llamado diagrama de "caja y bigotes". Al igual que los histogramas, son más adecuados para datos de escala de razón o de intervalo. La idea que subyace a un diagrama de caja es proporcionar una representación visual simple de la mediana, el rango intercuartílico y el rango de los datos. Y debido a que lo hacen de una manera bastante compacta, los diagramas de caja se han convertido en un gráfico estadístico muy popular, especialmente durante la etapa exploratoria del análisis de datos cuando intentas comprender los datos tú misma. Echemos un vistazo a cómo funcionan, usando nuevamente los datos de afl.margins como ejemplo.

La forma más sencilla de describir un diagrama de caja es dibujar uno. Haz clic en la casilla de verificación 'Diagrama de caja' y obtendrás el gráfico que se muestra en la parte inferior derecha de Figure 5.4. jamovi ha dibujado el diagrama de caja más básico posible. Así es como debe interpretarse este gráfico: la línea gruesa en el medio del cuadro es la mediana; el cuadro en sí abarca el rango del percentil 25 al percentil 75; y los "bigotes" salen al punto de datos más extremo que no excede un cierto límite. De forma predeterminada, este valor es 1,5 veces el rango intercuartílico (RIC), calculado

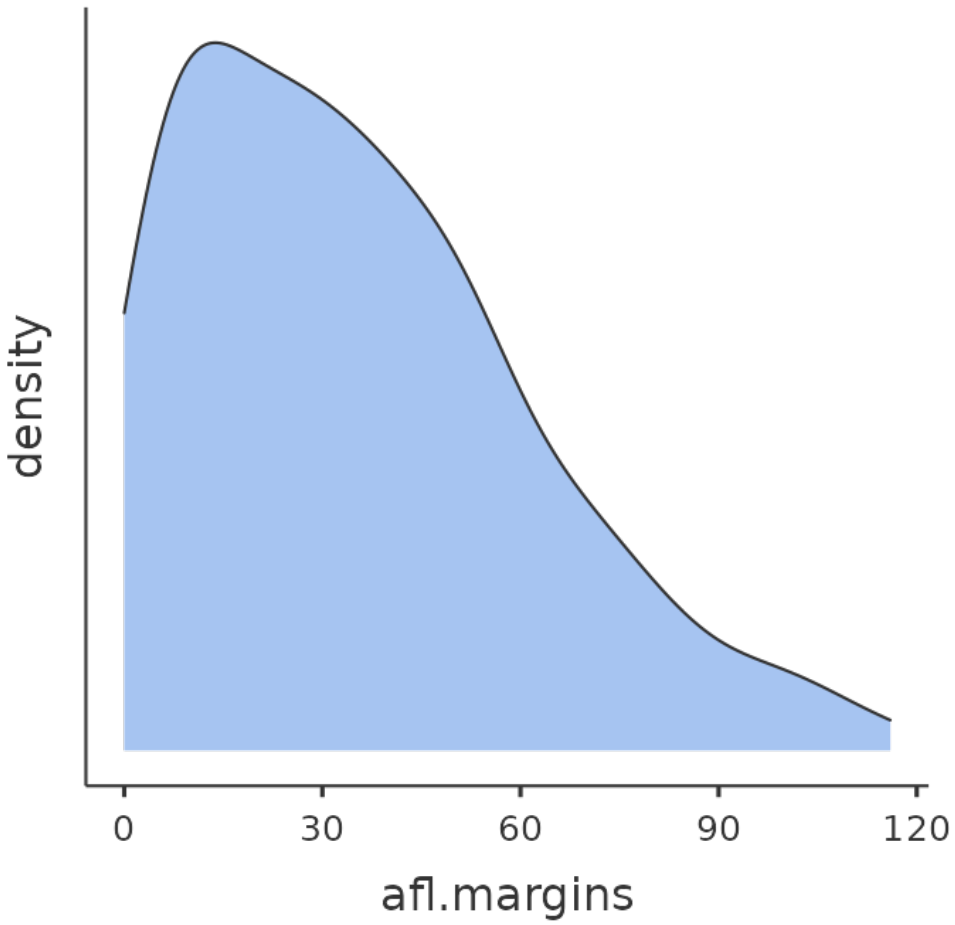


Figure 5.3: Un gráfico de densidad de la variable afl.margins trazada en jamovi

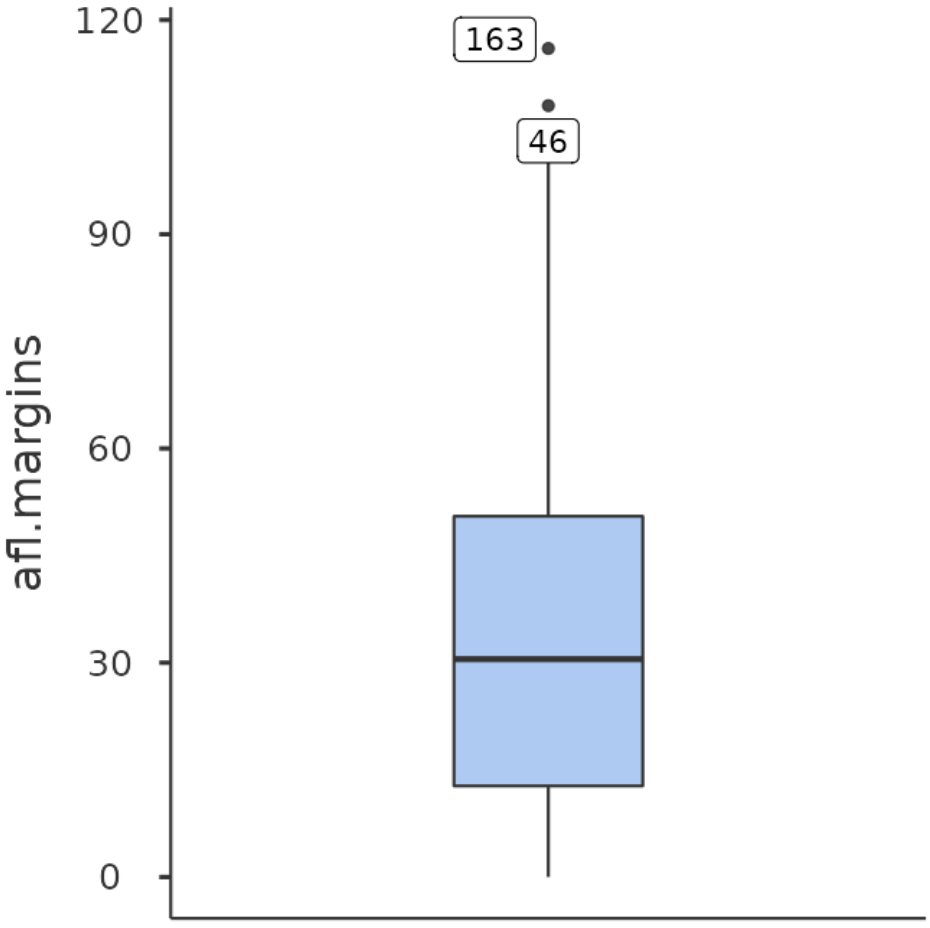


Figure 5.4: un gráfico de caja de la variable afl.margins trazada en jamovi

como el percentil 25 - (1,5*IQR) para el límite inferior y el percentil 75 + (1,5*IQR) para el límite superior. Cualquier observación cuyo valor se encuentre fuera de este rango se traza como un círculo o un punto en lugar de estar cubierto por los bigotes, y normalmente se lo denomina **valor atípico**. En nuestros datos de márgenes AFL hay dos observaciones que caen fuera de este rango, y estas observaciones se trazan como puntos (el límite superior es 107, y mirando la columna de datos en la hoja de cálculo hay dos observaciones con valores más altos que este, 108 y 116, así que estos son los puntos).

5.2.1 Diagramas de violín

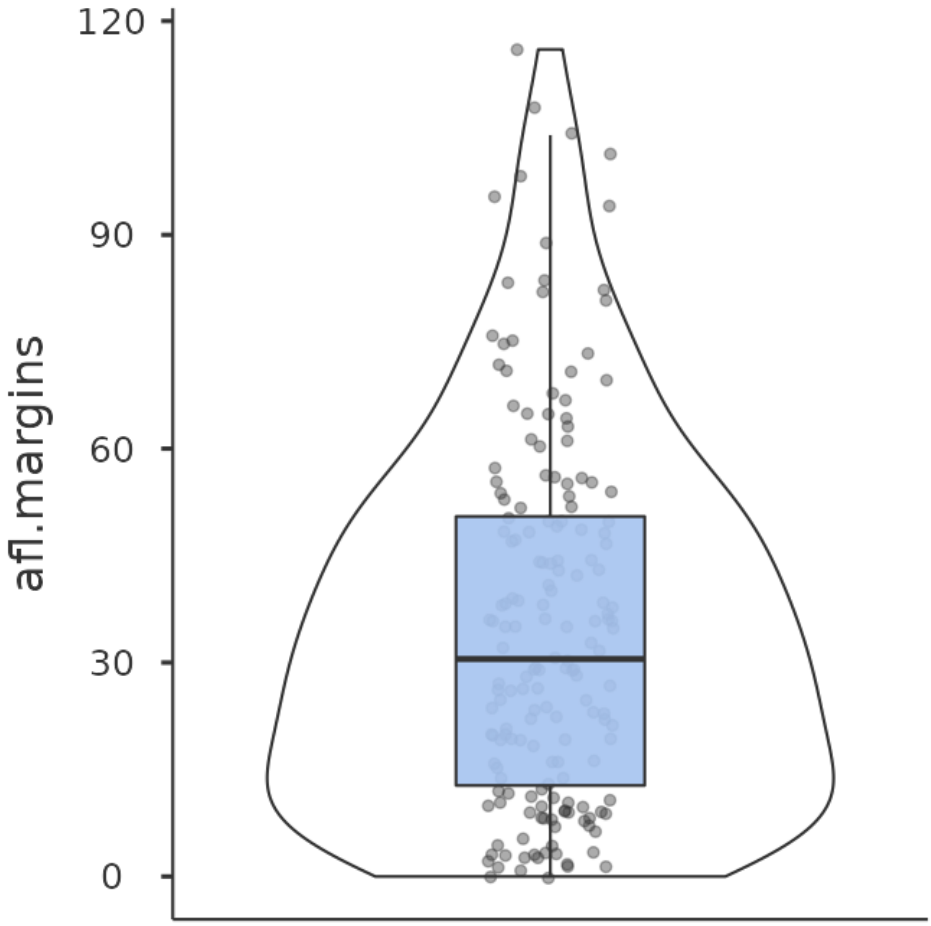


Figure 5.5: un diagrama de violín de la variable afl.margins trazado en jamovi, que también muestra un diagrama de caja y puntos de datos

Una variación del diagrama de caja tradicional es el diagrama de violín. Los diagramas de violín son similares a los diagramas de caja, excepto que también muestran la den-

sidad de probabilidad de kernel de los datos en diferentes valores. Por lo general, los diagramas de violín incluirán un marcador para la mediana de los datos y un cuadro que indica el rango intercuartílico, como en los diagramas de caja estándar. En jamovi, puedes conseguir este tipo de funcionalidad marcando las casillas de verificación 'Violín' y 'Box plot'. Consulta Figure 5.5, que también tiene activada la casilla de verificación 'Datos' para mostrar los puntos de datos reales en el gráfico. Sin embargo, esto hace que el gráfico esté demasiado recargado, en mi opinión. La claridad es simplicidad, por lo que en la práctica sería mejor usar un simple diagrama de caja.

5.2.2 Dibujar múltiples diagramas de caja

Una última cosa. ¿Qué sucede si quieres dibujar varios diagramas de caja a la vez? Supongamos, por ejemplo, que quisieras diagramas de caja separados que mostraran los márgenes de AFL no solo para 2010 sino para todos los años entre 1987 y 2010. Para hacer eso, lo primero que tendremos que hacer es encontrar los datos. Estos se almacenan en el archivo **aflmarginbyyear.csv**. Carguémoslo en jamovi y veamos qué contiene. Verás que es un conjunto de datos bastante grande. Contiene 4296 juegos y las variables que nos interesan. Lo que queremos hacer es que jamovi dibuje diagramas de caja para la variable de margen, pero graficados por separado para cada año. Para hacer esto tienes que mover la variable del año al cuadro 'Dividir por', como en Figure 5.6.

El resultado se muestra en Figure 5.7. Esta versión del diagrama de caja, dividida por año, da una idea de por qué a veces es útil elegir diagramas de caja en lugar de histogramas. Es posible tener una buena idea del aspecto de los datos de un año a otro sin abrumarse con demasiados detalles. Imagina lo que hubiera pasado si hubieras intentado meter 24 histogramas en este espacio: no hay ninguna posibilidad de que el lector aprenda nada útil.

5.2.3 Uso de diagramas de caja para detectar valores atípicos

Dado que el diagrama de caja separa automáticamente aquellas observaciones que se encuentran fuera de un cierto rango, representándolas con un punto en jamovi, la gente a menudo los usa como un método informal para detectar **valores atípicos**: observaciones que están "sospechosamente" distantes del resto de los datos. Aquí hay un ejemplo. Supongamos que dibujé el diagrama de caja para los datos de márgenes de AFL y apareció como Figure 5.8. Está bastante claro que algo raro ocurre con dos de las observaciones. ¡Aparentemente, hubo dos juegos en los que el margen superó los 300 puntos! Eso no me suena bien. Ahora que he comenzado a sospechar, es hora de mirar un poco más de cerca los datos. En jamovi, puedes averiguar rápidamente cuáles de estas observaciones son sospechosas y luego puedes volver a los datos sin procesar para ver si ha habido un error en la entrada de datos. Una forma de hacer esto es decirle a jamovi que etiquete los valores atípicos, marcando la casilla junto a la casilla de verificación Diagrama de caja. Esto agrega una etiqueta de número de fila al lado del valor atípico en el diagrama de caja, para que puedas mirar esa fila y encontrar el valor extremo. Otra forma, más flexible, es configurar un filtro para que solo se incluyan aquellas observaciones con valores por encima de un cierto umbral. En nuestro ejemplo, el umbral es superior a 300, por lo que ese es el filtro que crearemos. Primero, haz clic en el botón 'Filtros' en la parte superior de la ventana jamovi y luego escribe 'margen $>$ 300' en el campo de filtro, como en Figure 5.9.

Este filtro crea una nueva columna en la vista de hoja de cálculo donde solo se incluyen

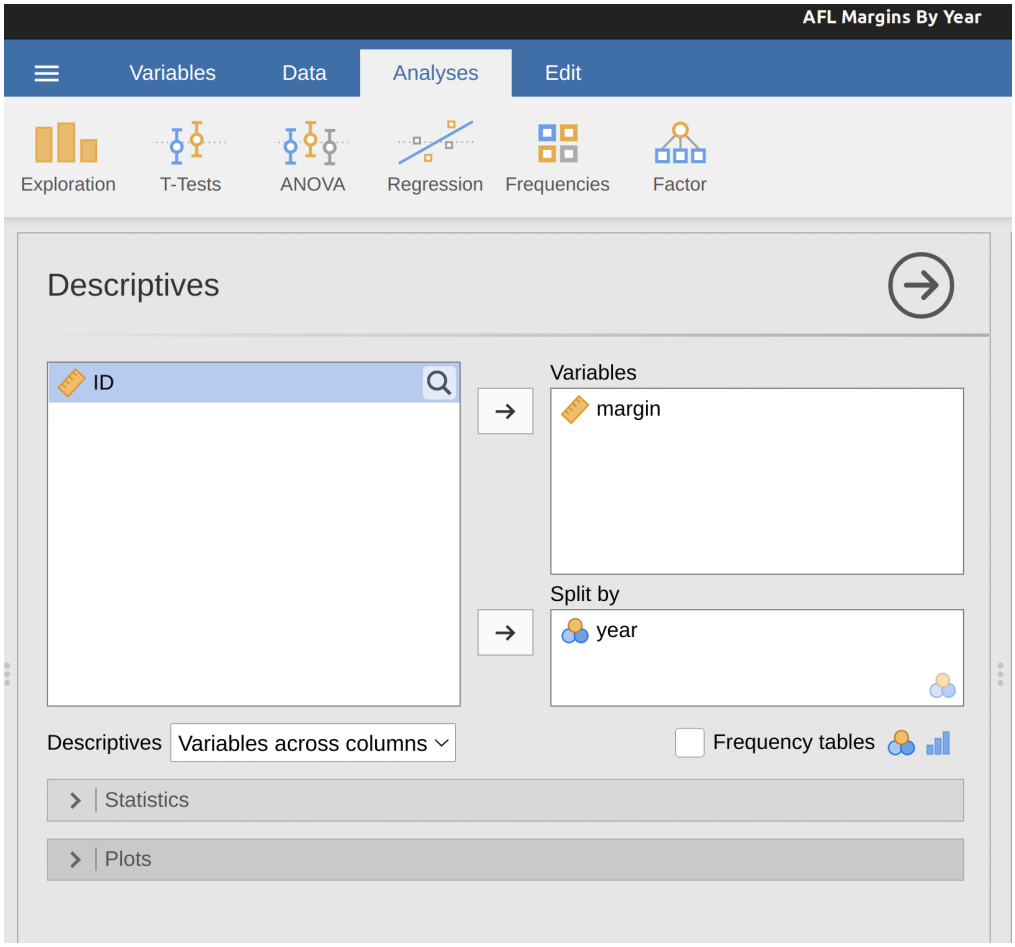


Figure 5.6: captura de pantalla de jamovi que muestra la ventana 'Dividir por'

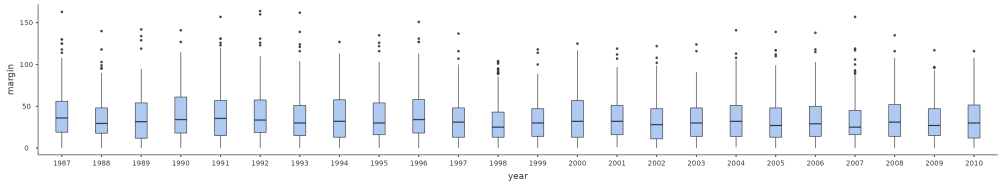


Figure 5.7: diagramas de caja múltiples trazados en jamovi, para las variables de margen por año

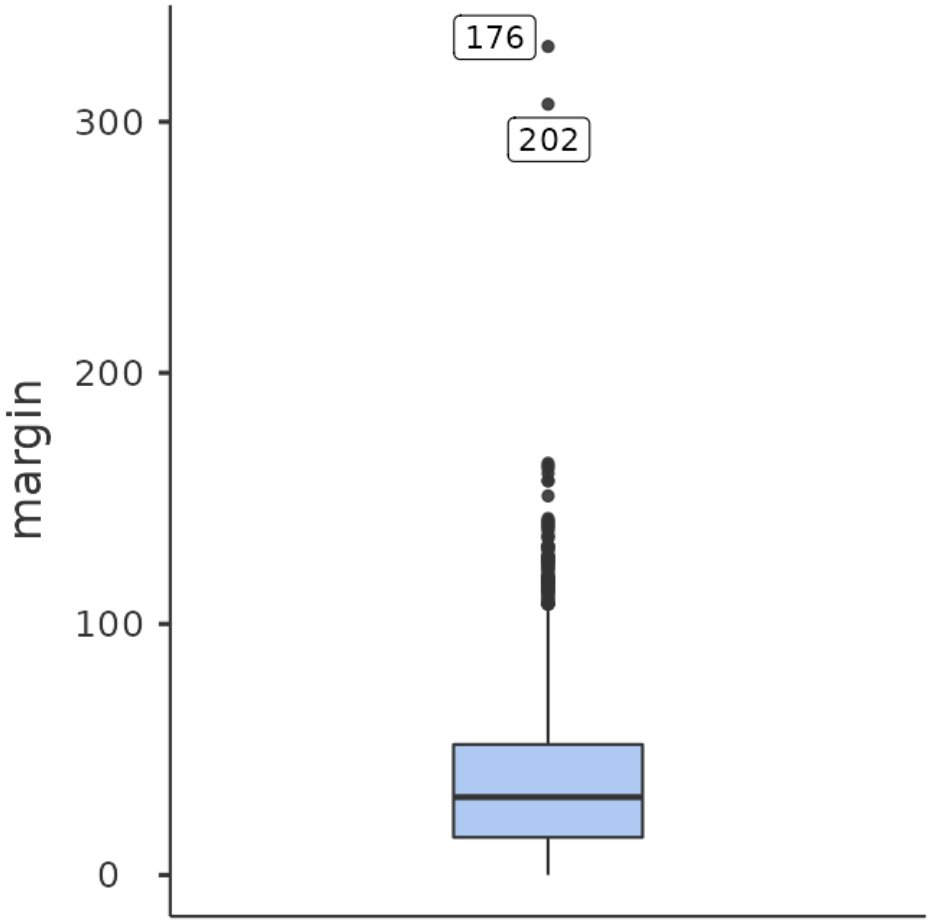


Figure 5.8: Un diagrama de caja que muestra dos valores atípicos muy sospechosos

aquellas observaciones que pasan el filtro. Una buena manera de identificar rápidamente qué observaciones son estas es decirle a jamovi que produzca una 'Tabla de frecuencia' (en la ventana 'Exploración' - 'Descriptivas') para la variable ID (que debe ser una variable nominal; de lo contrario, la tabla de frecuencia no se genera). En Figure 5.10 puedes ver que los valores de ID para las observaciones donde el margen era superior a 300 son 14 y 134. Estos son casos u observaciones sospechosas, donde debes volver a la fuente de datos original para averiguar qué está pasando.

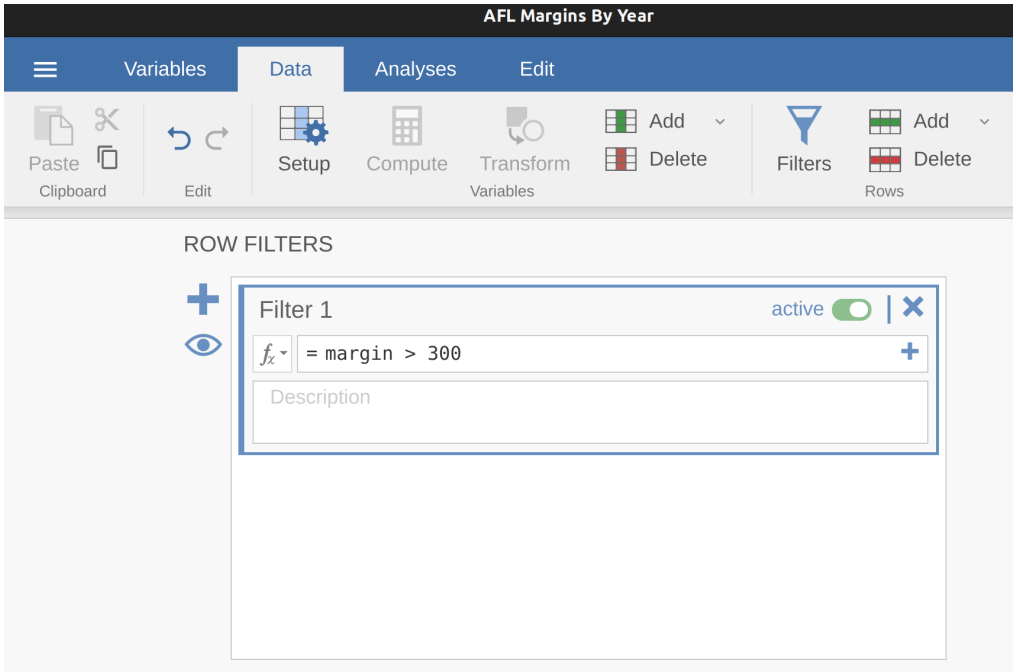


Figure 5.9: La pantalla de filtro jamovi

Suele ocurrir que alguien se equivoca de número. Aunque esto pueda parecer un ejemplo tonto, debo subrayar que este tipo de cosas ocurren realmente a menudo. Los conjuntos de datos del mundo real suelen estar plagados de errores estúpidos, especialmente cuando alguien ha tenido que teclear algo en un ordenador en algún momento. De hecho, esta fase en el análisis de datos tiene un nombre y, en la práctica, puede ocupar una gran parte de nuestro tiempo: limpieza de datos. Consiste en buscar errores tipográficos ("erratas"), datos faltantes y todo tipo de errores molestos en los archivos de datos brutos.

En el caso de los valores menos extremos, incluso si se marcan en un gráfico de caja como valores atípicos, la decisión de incluirlos o excluirlos en cualquier análisis depende en gran medida de por qué crees que los datos son como son y para qué quieres utilizarlos. En este caso hay que actuar con buen criterio. Si el valor atípico te parece legítimo, consérvalo. En cualquier caso, volveré al tema nuevamente en Section 12.10 en Chapter 12.

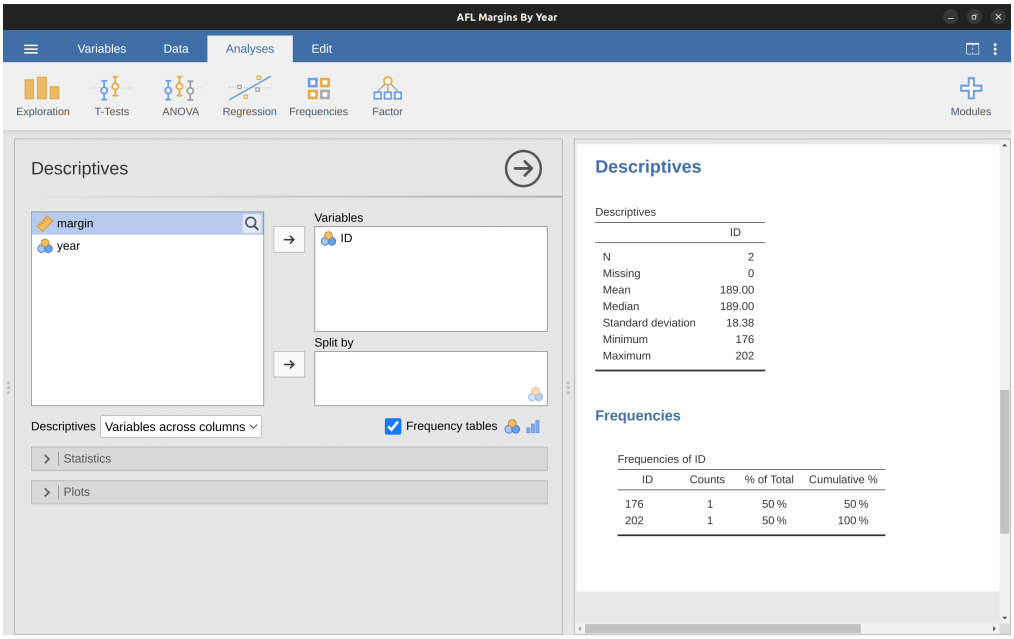


Figure 5.10: tabla de frecuencias para ID que muestra los números de ID de los dos valores atípicos sospechosos, 14 y 134

5.3 Gráficos de barras

Otra forma de gráfico que a menudo se desea trazar es el **gráfico de barras**. Utilicemos el conjunto de datos afl.finalists con la variable afl.finalists que introduje en Section 4.1.6. Lo que quiero hacer es dibujar un gráfico de barras que muestre la cantidad de finales en las que ha jugado cada equipo durante el tiempo que abarca el conjunto de datos afl.finalists. Hay muchos equipos, pero estoy particularmente interesada en solo cuatro: Brisbane, Carlton, Fremantle y Richmond. Así que el primer paso es configurar un filtro para que solo esos cuatro equipos se incluyan en el gráfico de barras. Esto es sencillo en jamovi y puedes hacerlo usando la función 'Filtros' que usamos anteriormente. Abre la pantalla 'Filtros' y escribe lo siguiente:

afl.finalistas == 'Brisbane' o afl.finalistas == 'Carlton' o afl.finalistas == 'Fremantle' o afl.finalistas == 'Richmond' [2]

Cuando hayas hecho esto, verás, en la vista 'Datos', que jamovi ha filtrado todos los valores excepto los que hemos especificado. A continuación, abre la ventana 'Exploración' - 'Descriptivas' y haz clic en la casilla de verificación 'Gráfico de barras' (recuerda mover la variable 'afl.finalists' al cuadro 'Variables' para que jamovi sepa qué variable usar). Luego deberías obtener un gráfico de barras, algo como el que se muestra en Figure 5.11.

[2]jamovi usa el símbolo "==" aquí para significar "coincidencias".

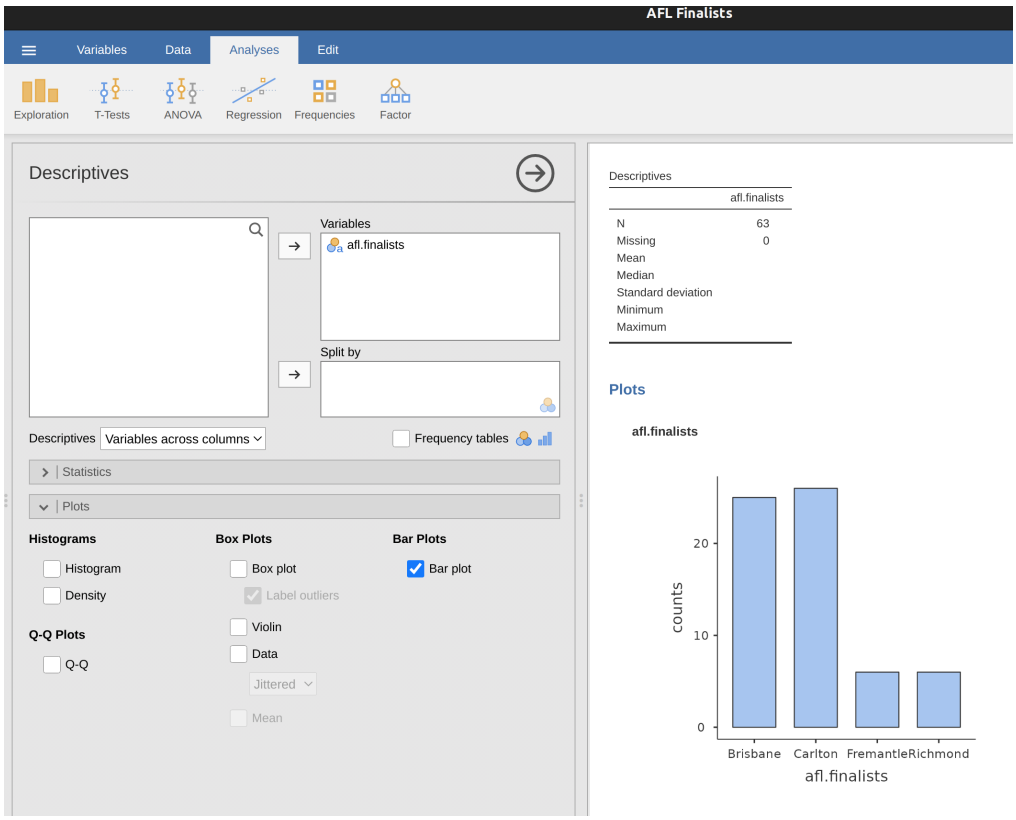


Figure 5.11: filtrar para incluir solo cuatro equipos de la AFL y dibujar un gráfico de barras en jamovi

5.4 Guardar archivos de imagen usando jamovi

Espera, estarás pensando. ¿De qué sirve poder hacer dibujos bonitos en jamovi si no puedo guardarlos y enviárselos a mis amigos para alardear de lo increíbles que son mis datos? ¿Cómo guardo la imagen? Muy sencillo. Haz clic con el botón derecho del ratón en la imagen del gráfico y expórtala a un archivo, ya sea como 'png', 'eps', 'svg' o 'pdf'. Todos estos formatos producen bonitas imágenes que luego puedes enviar o incluir en tus tareas o trabajos.

5.5 Resumen

Tal vez soy una persona de mente simple, pero me encantan las fotos. Cada vez que escribo un nuevo artículo científico, una de las primeras cosas que hago es sentarme y pensar en cuáles serán las imágenes. En mi cabeza, un artículo no es más que una secuencia de imágenes unidas por una historia. Todo lo demás es solo un escaparate. Lo que realmente estoy tratando de decir aquí es que el sistema visual humano es una herramienta de análisis de datos muy poderosa. Dale el tipo correcto de información y proporcionará al lector humano una gran cantidad de conocimiento muy rápidamente. No en vano tenemos el dicho "una imagen vale más que mil palabras". Con eso en mente, creo que este es uno de los capítulos más importantes del libro. Los temas tratados fueron:

- *Gráficos comunes.* Gran parte del capítulo se centró en los gráficos estándar que a los estadísticos les gusta producir: Histogramas, Diagramas de caja y Gráficos de barras
- Guardar archivos de imagen usando jamovi. Es importante destacar que también cubrimos cómo exportar sus imágenes.

Una última cosa a señalar. Si bien jamovi produce algunos gráficos predeterminados realmente buenos, actualmente no es posible editarlos. Para gráficos más avanzados y capacidad de trazado, los paquetes disponibles en R son mucho más potentes. Uno de los sistemas de gráficos más populares lo proporciona el paquete ggplot2 (ver https://ggplot2.tidyverse.org/), que se basa libremente en "La gramática de los gráficos" (Wilkinson et al., 2006). No es para novatos. Necesitas tener un conocimiento bastante bueno de R antes de poder comenzar a usarlo, e incluso entonces lleva un tiempo dominarlo. Pero cuando esté listo, vale la pena tomarse el tiempo para aprender por ti misma, porque es un sistema mucho más poderoso y más limpio.

Chapter 6

Cuestiones prácticas

> *El jardín de la vida nunca parece limitarse a las parcelas que los filósofos han trazado para su conveniencia. Tal vez algunos tractores más bastarían.*
> – Roger Zelazny[1]

Este es un capítulo un tanto extraño, incluso para mis estándares. Mi objetivo en este capítulo es hablar sobre las realidades de trabajar con datos un poco más honestamente de lo que verás en cualquier otra parte del libro. El problema con los conjuntos de datos del mundo real es que están *desordenados*. Muy a menudo, el archivo de datos con el que comienzas no tiene las variables almacenadas en el formato correcto para el análisis que quieres realizar. A veces puede puede que falten muchos valores en el conjunto de datos. A veces, solo quieres analizar un subconjunto de los datos. Etcétera. En otras palabras, hay un montón de **manipulación de datos** que necesitas hacer solo para obtener las variables en el formato que necesitas. El propósito de este capítulo es proporcionar una introducción básica a estos temas prácticos. Aunque el capítulo está motivado por los tipos de problemas prácticos que surgen cuando se manipulan datos reales, seguiré con la práctica que he adoptado durante la mayor parte del libro y me basaré en conjuntos de datos muy pequeños que ilustran el problema subyacente. Como este capítulo es esencialmente una colección de técnicas y no cuenta una sola historia coherente, puede ser útil empezar con una lista de temas:

- Tabulación y tabulación cruzada de datos
- Expresiones lógicas en jamovi
- Transformar y recodificar una variable
- Otras funciones y operaciones matemáticas
- Extracción de un subconjunto de datos

Como puedes ver, la lista de temas que cubre el capítulo es bastante amplia y hay mucho contenido. Aunque este es uno de los capítulos más largos y difíciles del libro, en realidad solo estoy arañando la superficie de varios temas bastante diferentes e importantes. Mi consejo, como siempre, es que leas el capítulo una vez e intentes seguirlo todo lo que puedas. No te preocupe demasiado si no puedes entenderlo todo de una vez, especialmente las últimas secciones. El resto del libro depende muy poco de este capítulo, así que puedes conformarte con entender lo básico. Sin embargo, lo más probable es que

[1] La cita proviene de Home is the Hangman, publicado en 1975.

más adelante tengas que volver a este capítulo para entender algunos de los conceptos a los que me refiero aquí.

6.1 Tabulación y tabulación cruzada de datos

Una tarea muy común al analizar datos es la construcción de tablas de frecuencia, o tabulación cruzada de una variable contra otra. Esto se puede conseguir en jamovi y te mostraré cómo en esta sección.

6.1.1 Creación de tablas para variables individuales

Comencemos con un ejemplo simple. Como padre de un niño pequeño, naturalmente paso mucho tiempo viendo programas de televisión como In the *Night Garden*. En el archivo nightgarden.csv, transcribí una breve sección del diálogo. El archivo contiene dos variables de interés, locutor y enunciado. Abre este conjunto de datos en jamovi y echa un vistazo a los datos en la vista de 'hoja de cálculo'. Verás que los datos tienen este aspecto:

variable 'locutor': upsy-daisy upsy-daisy upsy-daisy upsy-daisy tombliboo tombliboo makka-pakka makka-pakka makka-pakka makka-pakka variable de 'pronunciación': pip pip onk onk ee oo pip pip onk onk

¡Mirando esto queda muy claro lo que le pasó a mi cordura! Con estos como mis datos, una tarea que podría necesitar hacer es construir un recuento de frecuencia de la cantidad de palabras que habla cada personaje durante el programa. La pantalla 'Descriptivas' de jamovi tiene una casilla de verificación llamada 'Tablas de frecuencia' que hace exactamente esto, consulta Table 6.1.

Table 6.1: tabla de frecuencias para la variable locutor

levels	**Counts**	**% of Total**	**Cumulative %**
makka-pakka	4	40%	40%
tombliboo	2	20%	60%
upsy-daisy	4	40%	100%

El resultado aquí nos dice en la primera línea que lo que estamos viendo es una tabulación de la variable locutor. En la columna 'Niveles' enumera los diferentes locutores que existen en los datos, y en la columna 'Recuentos' te dice cuántas veces aparece ese locutor en los datos. En otras palabras, es una tabla de frecuencias.

En jamovi, la casilla de verificación 'Tablas de frecuencia' solo producirá una tabla para variables individuales. Para una tabla de dos variables, por ejemplo, combinar locutor y enunciado para que podamos ver cuántas veces cada locutor dijo un enunciado en particular, necesitamos una tabulación cruzada o una tabla de contingencia. En jamovi, puedes hacer esto seleccionando el análisis 'Frecuencias' - 'Tablas de contingencia' - 'Muestras independientes' y moviendo la variable del locutor al cuadro 'Filas' y la variable de expresiones al cuadro 'Columnas'. Entonces deberías tener una tabla de contingencia como la que se muestra en Figure 6.1.

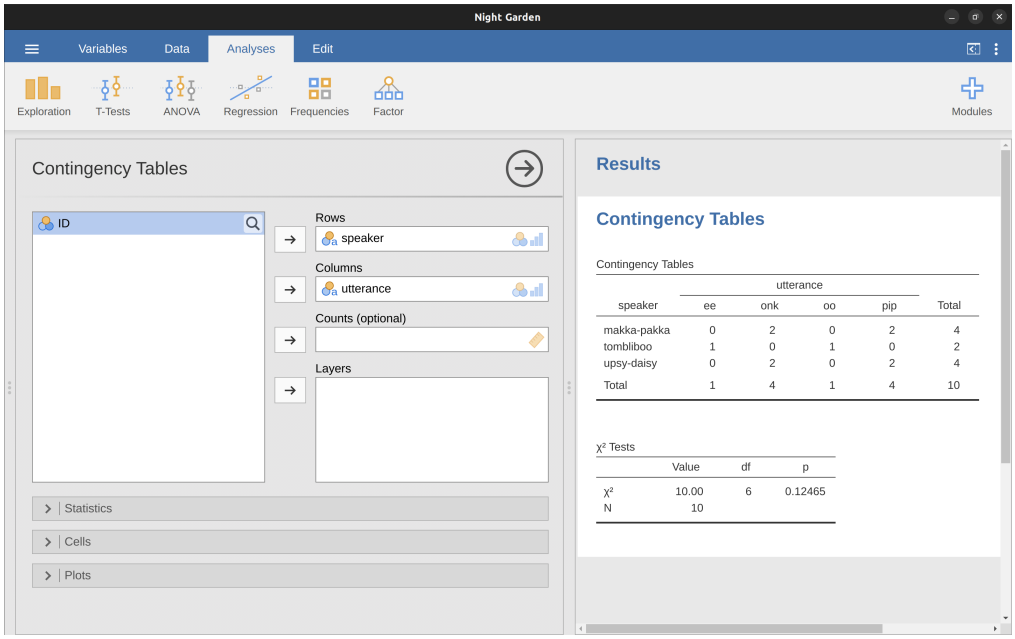


Figure 6.1: tabla de contingencia para las variables locutor y enunciados

No te preocupes por la tabla "χ^2 Tests" que se genera. Veremos esto más adelante en Chapter 10. Al interpretar la tabla de contingencia recuerda que estos son recuentos, por lo que el hecho de que la primera fila y la segunda columna de números correspondan a un valor de 2 indica que Makka-Pakka (fila 1) dice "onk" (columna 2) dos veces en este conjunto de datos.

6.1.2 Añadir porcentajes a una tabla de contingencia

La tabla de contingencia que se muestra en Figure 6.1 muestra una tabla de frecuencias brutas. Es decir, un recuento del número total de casos para diferentes combinaciones de niveles de las variables especificadas. Sin embargo, a menudo quieres que tus datos se organicen en términos de porcentajes y recuentos. Puedes encontrar las casillas de verificación para diferentes porcentajes en la opción 'Celdas' en la ventana 'Tablas de contingencia'. Primero, haz clic en la casilla de verificación 'Fila' y la tabla de contingencia en la ventana de resultados cambiará a la de Figure 6.2.

Lo que estamos viendo aquí es el porcentaje de expresiones hechas por cada personaje. En otras palabras, el 50% de las declaraciones de Makka-Pakka son "pip" y el otro 50% son "onk". Comparemos esto con la tabla que obtenemos cuando calculamos los porcentajes de columna (desmarca 'Fila' y marca 'Columna' en la ventana de opciones de Celdas), ver Figure 6.3. En esta versión, lo que vemos es el porcentaje de caracteres asociados con cada enunciado. Por ejemplo, cada vez que se hace la expresión "ee" (en este conjunto de datos), el 100 % de las veces es un Tombliboo quien lo dice.

Contingency Tables

Contingency Tables

speaker		utterance ee	onk	oo	pip	Total
makka-pakka	Observed	0	2	0	2	4
	% within row	0 %	50 %	0 %	50 %	100 %
tombliboo	Observed	1	0	1	0	2
	% within row	50 %	0 %	50 %	0 %	100 %
upsy-daisy	Observed	0	2	0	2	4
	% within row	0 %	50 %	0 %	50 %	100 %
Total	Observed	1	4	1	4	10
	% within row	10 %	40 %	10 %	40 %	100 %

Figure 6.2: Tabla de contingencia para las variables locutor y enunciados, con porcentajes de fila

Contingency Tables

Contingency Tables

speaker		utterance ee	onk	oo	pip	Total
makka-pakka	Observed	0	2	0	2	4
	% within column	0 %	50 %	0 %	50 %	40 %
tombliboo	Observed	1	0	1	0	2
	% within column	100 %	0 %	100 %	0 %	20 %
upsy-daisy	Observed	0	2	0	2	4
	% within column	0 %	50 %	0 %	50 %	40 %
Total	Observed	1	4	1	4	10
	% within column	100 %	100 %	100 %	100 %	100 %

Figure 6.3: Tabla de contingencia para las variables locutor y enunciados, con porcentajes de columna

6.2 Expresiones lógicas en jamovi

Un concepto clave en el que se basan muchas transformaciones de datos en jamovi es la idea de un **valor lógico**. Un valor lógico es una afirmación sobre si algo es verdadero o falso. Esto se implementa en jamovi de una manera bastante sencilla. Hay dos valores lógicos, a saber, VERDADERO y FALSO. A pesar de la simplicidad, los valores lógicos son muy útiles. Veamos cómo funcionan.

6.2.1 Evaluar verdades matemáticas

En el libro clásico de George Orwell de 1984, uno de los lemas utilizados por el Partido totalitario era "dos más dos es igual a cinco". La idea es que la dominación política de la libertad humana se completa cuando es posible subvertir incluso la más básica de las verdades. Es un pensamiento aterrador, especialmente cuando el protagonista Winston Smith finalmente se derrumba bajo la tortura y acepta la propuesta. "El hombre es infinitamente maleable", dice el libro. Estoy bastante segura de que esto no es cierto para los humanos[2] y definitivamente no es cierto para jamovi. jamovi no es infinitamente maleable, tiene opiniones bastante firmes sobre el tema de lo que es y no es cierto, al menos en lo que respecta a las matemáticas básicas. Si le pido que calcule $2 + 2$[3], siempre da la misma respuesta, ¡y no es un 5!

Por supuesto, hasta ahora jamovi solo está haciendo los cálculos. No le he pedido que afirme explícitamente que $2 + 2 = 4$ es una afirmación verdadera. Si quiero que jamovi haga un juicio explícito, puedo usar un comando como este: $2 + 2 == 4$

Lo que he hecho aquí es usar el **operador de igualdad**, ==, para obligar a jamovi a hacer un juicio de "verdadero o falso".[4] el eslogan del Partido, así que escribe esto en el cuadro Calcular nueva variable 'fórmula':

$$2 + 2 == 5$$

¿Y qué obtienes? Debería ser un conjunto completo de valores 'falsos' en la columna de la hoja de cálculo para tu variable recién calculada. ¡Yupi! ¡Libertad y ponis para todos! O algo así. De todos modos, valió la pena echar un vistazo a lo que sucede si intentas obligar a jamovi a creer que dos más dos son cinco haciendo una afirmación como $2 + 2 = 5$. Sé que si hago esto en otro programa, digamos R, nos da un mensaje de error. Pero espera, si haces esto en jamovi obtienes un conjunto completo de valores 'falsos'. ¿Entonces qué está pasando? Bueno, parece que jamovi está siendo bastante inteligente y se da cuenta de que está probando si es VERDADERO o FALSO que $2 + 2 = 5$, independientemente de si usas el **operador de igualdad** correcto, ==, o el signo igual "=".

[2] Ofrezco mis intentos de adolescente de ser "cool" como evidencia de que algunas cosas simplemente no se pueden hacer.

[3] puedes hacer esto en la pantalla Calcular nueva variable, ¡aunque calcular 2 + 2 para cada celda de una nueva variable no es muy útil!

[4] Ten en cuenta que este es un operador muy diferente al operador de igualdad =. Un error tipográfico común que se comete cuando se intentan escribir comandos lógicos en jamovi (u otros idiomas, ya que la distinción "= versus ==" es importante en muchos programas informáticos y estadísticos) es escribir accidentalmente = cuando realmente quieres decir ==. Ten especial cuidado con esto, he estado programando en varios lenguajes desde que era adolescente y todavía me equivoco mucho. Mmm. Creo que veo por qué no era guay cuando era adolescente. Y por qué todavía sigo sin molar.

6.2.2 Operaciones lógicas

Ya hemos visto cómo funcionan las operaciones lógicas. Pero hasta ahora solo hemos visto el ejemplo más sencillo posible. Probablemente no te sorprenderá descubrir que podemos combinar operaciones lógicas con otras operaciones y funciones de una manera más complicada, como esta: $3 \times 3 + 4 \times 4 == 5 \times 5$ o esto $SQRT(25) == 5$

No solo eso, sino que, como ilustra Table 6.2, existen otros operadores lógicos que puedes usar y que corresponden a algunos conceptos matemáticos básicos. Esperamos que todos estos se expliquen por sí mismos. Por ejemplo, el operador **menor que** $<$ verifica si el número de la izquierda es menor que el número de la derecha. Si es menor, entonces jamovi devuelve una respuesta de VERDADERO, pero si los dos números son iguales, o si el de la derecha es mayor, entonces jamovi devuelve una respuesta de FALSO.

Por el contrario, el operador **menor que o igual a** $<=$ hará exactamente lo que dice. Devuelve un valor de VERDADERO si el número del lado izquierdo es menor o igual que el número del lado derecho. En este punto, espero que sea bastante obvio lo que hacen el operador **mayor que** $>$ y el operador **mayor que o igual a** $>=$.

El siguiente en la lista de operadores lógicos es el operador **distinto de** $!=$ que, como todos los demás, hace lo que dice que hace. Devuelve un valor de VERDADERO cuando las cosas en cualquier lado no son idénticas entre sí. Por lo tanto, dado que $2 + 2$ no es igual a 5, obtendríamos 'verdadero' como el valor de nuestra variable recién calculada. Pruébalo y verás:

$$2 + 2 != 5$$

Aún no hemos terminado. Hay tres operaciones lógicas más que vale la pena conocer, enumeradas en Table 6.3. Estos son el operador **no** !, el operador **y** and, y el operador **o** or. Al igual que los otros operadores lógicos, su comportamiento es más o menos el que cabría dados sus nombres. Por ejemplo, si te pido que evalúes la afirmación de que "o bien $2 + 2 = 4$ o $2 + 2 = 5$", dirías que es verdad. Dado que es una declaración de "o esto o lo otro", lo que necesitamos es que una de las dos partes sea verdadera. Eso es lo que hace el operador or:[5]

$$(2 + 2 == 4) \text{ o } (2 + 2 == 5)$$

Por otro lado, si te pido que evalúes la afirmación de que "ambos $2 + 2 = 4$ y $2 + 2 = 5$", dirías que es falso. Dado que se trata de una afirmación y necesitamos que ambas partes sean verdaderas. Y eso es lo que hace el operador and:

$$(2 + 2 == 4) \text{ y } (2 + 2 == 5)$$

Finalmente, está el operador not, que es simple pero molesto de describir en inglés. Si te pido que evalúes mi afirmación de que "no es cierto que $2 + 2 = 5$", entonces dirías

[5] He aquí una peculiaridad en jamovi. Cuando tenemos expresiones lógicas simples como las que ya hemos visto, por ejemplo, 2 + 2 == 5, jamovi indica claramente 'falso' (o 'verdadero') en la columna correspondiente de la hoja de cálculo. En realidad, jamovi almacena 'falso' como 0 y 'verdadero' como 1. Cuando tenemos expresiones lógicas más complejas, como (2+2 == 4) o (2+2 == 5), jamovi simplemente muestra 0 o 1, dependiendo de si la expresión lógica se evalúa como falsa o verdadera.

Table 6.2: algunos operadores lógicos

operation	operator	example input	answer
less than		2	TRUE
less than or equal to	<	2 < = 2	TRUE
greater than	>	2 > 3	FALSE
greater than or equal to	> =	2 > = 2	TRUE
equal to	= =	2 = = 3	FALSE
not equal to	!=	2 != 3	TRUE

Table 6.3: algunos operadores lógicos más

operation	operator	example input	answer
not	NOT	NOT(1==1)	FALSE
or	or	(1==1) or (2==3)	TRUE
and	and	(1==1) and (2==3)	FALSE

que mi afirmación es verdadera, porque en realidad mi afirmación es que "$2 + 2 = 5$ es falso". Y tengo razón. Si escribimos esto en jamovi usamos esto:

$$NO(2 + 2 == 5)$$

En otras palabras, dado que $2 + 2 == 5$ es una afirmación FALSA, debe darse el caso de que $NO(2 + 2 == 5)$ sea VERDADERA. Esencialmente, lo que realmente hemos hecho es afirmar que "no falso" es lo mismo que "verdadero". Obviamente, esto no es del todo correcto en la vida real. Pero jamovi vive en un mundo mucho más blanco o negro. Para jamovi todo es verdadero o falso. No se permiten matices de gris.

Por supuesto, en nuestro ejemplo de $2+2 = 5$, realmente no necesitábamos usar el operador "no" *NOT* y el operador "igual a" == como dos operadores separados. Podríamos haber usado el operador "no es igual a" $! =$ así:

$$2 + 2 != 5$$

6.2.3 Aplicando operaciones lógicas al texto

También quiero señalar brevemente que puedes aplicar estos operadores lógicos tanto a texto como a datos lógicos. Pero hay que tener un poco más de cuidado a la hora de entendercómo jamovi interpreta las diferentes operaciones. En esta sección hablaré de cómo se aplica el operador igual a == al texto, ya que es el más importante. Obviamente,

el operador no igual a != da exactamente las respuestas opuestas a ==, por lo que implícitamente también estoy hablando de eso, pero no daré comandos específicos que muestren el uso de ! =.

Bien, veamos cómo funciona. En cierto sentido, es muy sencillo. Por ejemplo, puedo preguntarle a jamovi si la palabra "gato" es lo mismo que la palabra "perro", así:

"gato" == "perro" Esto es bastante obvio, y es bueno saber que incluso jamovi puede darse cuenta. De manera similar, jamovi reconoce que un "gato" es un "gato": "gato" == "gato" Nuevamente, eso es exactamente lo que esperaríamos. Sin embargo, lo que debes tener en cuenta es que jamovi no es nada tolerante en lo que respecta a la gramática y el espaciado. Si dos cadenas difieren de alguna manera, jamovi dirá que no son iguales entre sí, como con lo siguiente: " cat" == "cat" "cat" == "CAT" "cat" == "ca t"

También puedes usar otros operadores lógicos. Por ejemplo, jamovi también te permite usar los operadores > y < para determinar cuál de las dos 'cadenas' de texto viene primero, alfabéticamente hablando. Más o menos. En realidad, es un poco más complicado que eso, pero empecemos con un ejemplo sencillo:

"gato" < "perro"

En jamovi, este ejemplo se evalúa como 'verdadero'. Esto se debe a que "gato" viene antes que "perro" en orden alfabético, por lo que jamovi considera que la afirmación es verdadera. Sin embargo, si le pedimos a jamovi que nos diga si "gato" viene antes que "zorro", evaluará la expresión como falsa. Hasta aquí todo bien. Pero los datos de texto son un poco más complicados de lo que sugiere el diccionario. ¿Qué pasa con "gato" y "GATO"? ¿Cuál de estos viene primero? Pruébalo y descúbrelo:

"GATO" < "gato"

De hecho, esto se evalúa como 'verdadero'. En otras palabras, jamovi asume que las letras mayúsculas van antes que las minúsculas. Me parece bien. Es probable que no te sorprenda. Lo que podría sorprenderte es que jamovi asume que todas las letras mayúsculas van antes que las minúsculas. Es decir, mientras que "ardilla" < "zorro" es una afirmación verdadera, y el equivalente en mayúsculas "ARDILLA" < "ZORRO" también es cierto, no es cierto decir que "ardilla" < "ZORRO" ", como ilustra el siguiente extracto. Prueba esto:

"ardilla" < "ZORRO"

Esto se evalúa como 'falso', y puede parecer un poco contraintuitivo. Con eso en mente, puede ser útil echar un vistazo rápido a Table 6.4 que enumera varios caracteres de texto en el orden en que jamovi los procesa.

6.3 Transformar y recodificar una variable

En el análisis de datos del mundo real, no es infrecuente encontrarse con que una de las variables no es exactamente equivalente a la variable que realmente quieres. Por ejemplo, a menudo es conveniente tomar una variable de valor continuo (p. ej., la edad) y dividirla en un número más pequeño de categorías (p. ej., más joven, mediana, mayor). En otras ocasiones, es posible que necesites convertir una variable numérica en una variable numérica diferente (p. ej., puede que quieras analizar el valor absoluto de la

Table 6.4: caracteres de texto en el orden en que jamovi los procesa

!	”	#	$	%	&	’	(
)	*	+	,	-	.	/	0
1	2	3	4	5	6	7	8
9	:	;	<	=	>	?	@
A	B	C	D	E	F	G	H
I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	X
Y	Z	[	\	]	^	_	‘
a	b	c	d	e	g	h	i
j	k	l	m	n	o	p	q
r	s	t	u	v	w	x	y
z	{	|	}				

variable original). En esta sección, describiré algunas formas clave de hacer estas cosas en jamovi.

6.3.1 Crear una variable transformada

El primer truco a discutir es la idea de **transformar** una variable. Literalmente, cualquier cosa que hagas a una variable es una transformación, pero en la práctica lo que generalmente significa es que aplicas una función matemática relativamente simple a la variable original para crear una nueva variable que (a) describa mejor lo que realmente te interesa, o (b) esté más de acuerdo con los supuestos de las pruebas estadísticas que quieres hacer. Como hasta ahora no he hablado de las pruebas estadísticas ni de sus supuestos, te mostraré un ejemplo basado en el primer caso.

Supongamos que he realizado un breve estudio en el que hago una sola pregunta a 10 personas: En una escala de 1 (totalmente en desacuerdo) a 7 (totalmente de acuerdo), ¿en qué medida está de acuerdo con la afirmación de que “los dinosaurios son increíbles”?

Ahora vamos a cargar y ver los datos. El archivo de datos likert.omv contiene una única variable con las respuestas brutas en escala Likert para estas 10 personas. Sin embargo, si se piensa en ello, esta no es la mejor manera de representar estas respuestas. Debido a la forma bastante simétrica en que configuramos la escala de respuesta, en cierto sentido el punto medio de la escala debería haber sido codificado como 0 (sin opinión), y los dos puntos finales deberían ser ‘3 (totalmente de acuerdo) y ´ 3 (totalmente en desacuerdo). Al recodificar los datos de esta manera, refleja un poco más cómo pensamos realmente sobre las respuestas. La recodificación aquí es bastante sencilla, simplemente restamos 4 de las puntuaciones brutas. En jamovi puedes hacer esto calculando una nueva variable: haz clic en el botón ‘Datos’ - ‘Calcular’ y verás que se ha agregado una nueva variable a la hoja de cálculo. Llamemos a esta nueva variable likert.centred (tecléalo) y luego añade lo siguiente en el cuadro de fórmulas, como en Figure 6.4: ‘likert.raw - 4’

Una de las razones por las que puede ser útil tener los datos en este formato es que hay muchas situaciones en las que se puede preferir analizar la fuerza de la opinión por

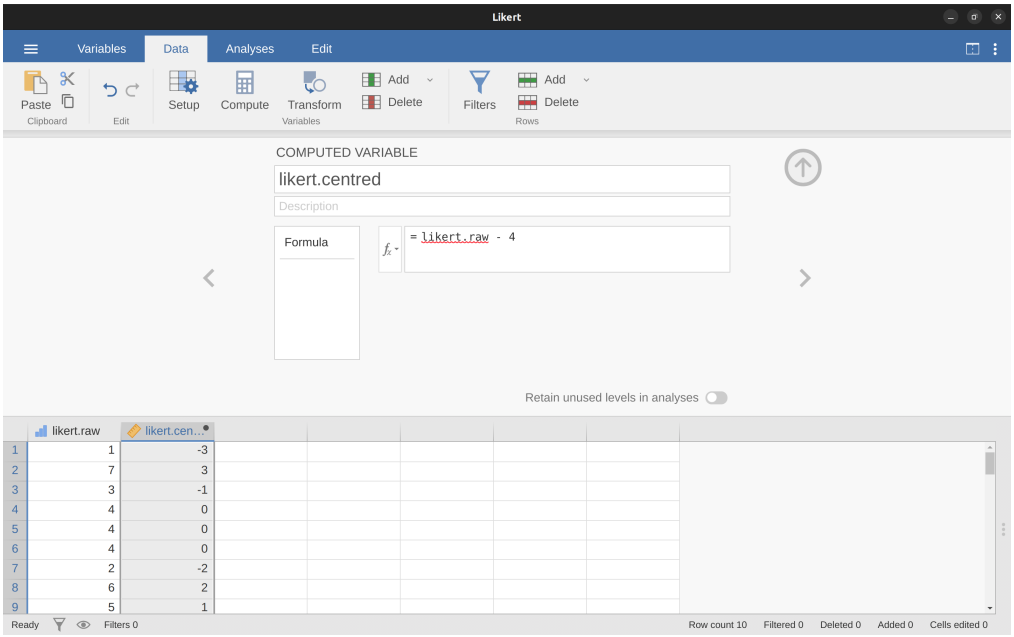


Figure 6.4: Crear una nueva variable calculada en jamovi

separado de la dirección de la opinión. Podemos hacer dos transformaciones diferentes en esta variable Likert centrada para distinguir entre estos dos conceptos diferentes. Primero, para calcular la variable opinion.strength (fuerza de la opinión) coge el valor absoluto de los datos centrados (usando la función 'ABS').[6] En jamovi se crea una nueva variable utilizando el botón 'Calcular'. Llama a la variable opinion.strength y esta vez haz clic en el botón fx situado al lado de la casilla 'Fórmula'. Esto muestra las diferentes 'Funciones' y 'Variables' que puedes añadir a la casilla 'Fórmula', así que haz doble clic en 'ABS' y luego doble clic en "likert.centred" y verás que la casilla 'Fórmula' se rellena con ABS(likert.centred) y se ha creado una nueva variable en la vista de hoja de cálculo, como en Figure 6.5.

En segundo lugar, para calcular una variable que contiene solo la dirección de la opinión e ignora la fuerza, queremos calcular el 'signo' de la variable. En jamovi podemos usar la función IF para hacerlo. Crea otra nueva variable con el botón 'Calcular', llámala opinion.sign, y luego escribe lo siguiente en el cuadro de función:

IF(likert.centred == 0, 0, likert.centred / opinion.strength) Cuando termines, verás que todos los números negativos de la variable likert.centred se convierten en -1, todos los números positivos se convierten en 1 y cero se queda como 0, así:

-1 1 -1 0 0 0 -1 1 1 1

Analicemos qué está haciendo este comando 'IF'. En jamovi hay tres partes en una declaración 'IF', escrita como 'IF (expression, value, else)'. La primera parte, 'expression', puede ser un enunciado lógico o matemático. En nuestro ejemplo, hemos especi-

[6] El valor absoluto de un número es su distancia al cero, independientemente de si su signo es negativo o positivo.

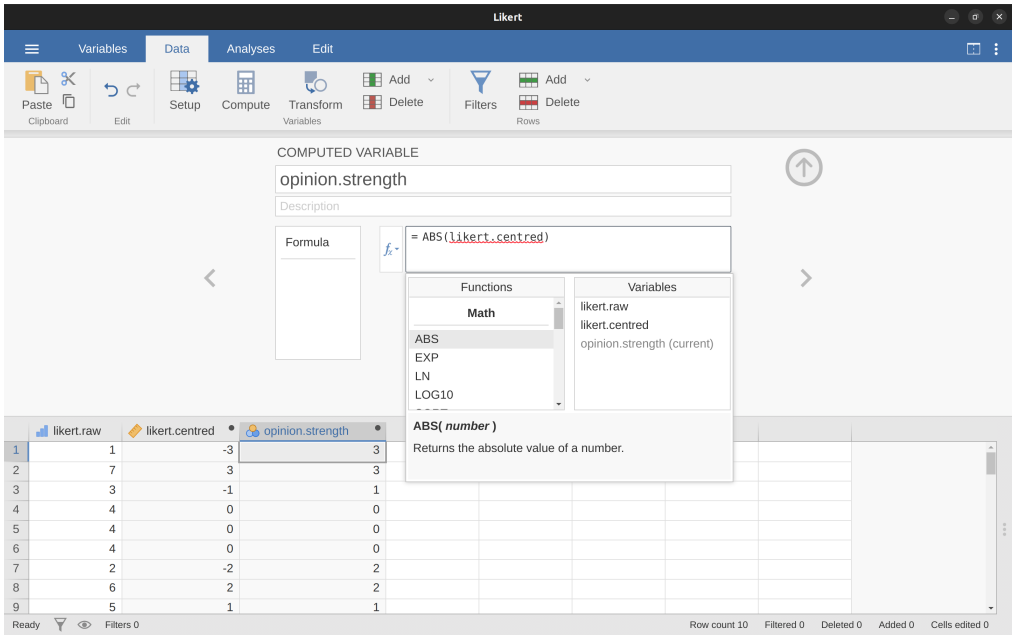


Figure 6.5: Usar el botón f_x para seleccionar funciones y variables

ficado 'likert.centred == 0', que es VERDADERO para valores donde likert.centred es cero. La siguiente parte, 'value', es el nuevo valor donde la expresión en la primera parte es VERDADERA. En nuestro ejemplo, hemos dicho que para todos aquellos valores donde likert.centred es cero, mantenlos en cero. En la siguiente parte, 'else', podemos incluir un enunciado lógico o matemático que se usará si la parte uno se evalúa como FALSO, es decir, donde likert.centred no es cero. En nuestro ejemplo hemos dividido likert.centred por opinion.strength para dar '-1' o '+1' dependiendo del signo del valor original en likert.centred.[7]

Y ya está. Ahora tenemos tres nuevas variables brillantes, todas las cuales son transformaciones útiles de los datos originales de likert.raw.

6.3.2 Descomponer una variable en un número menor de niveles discretos o categorías

Una tarea práctica que surge con bastante frecuencia es el problema de descomponer una variable en un número menor de niveles o categorías discretos. Por ejemplo, supongamos que estoy interesada en observar la distribución por edades de las personas en una reunión social:

60,58,24,26,34,42,31,30,33,2,9

En algunas situaciones, puede ser muy útil agruparlos en un número reducido de categorías. Por ejemplo, podríamos agrupar los datos en tres grandes categorías: jóvenes

[7] la razón por la que tenemos que usar el comando 'IF' y mantener cero como cero es que no puedes simplemente usar likert.centred / opinion.strength para calcular el signo de likert.centred, porque dividir matemáticamente cero por cero no funciona. Pruébalo y verás

(0-20), adultos (21-40) y mayores (41-60). Se trata de una clasificación bastante tosca, y las etiquetas que he adjuntado solo tienen sentido en el contexto de este conjunto de datos (por ejemplo, visto de manera más general, una persona de 42 años no se consideraría "mayor"). Podemos dividir esta variable con bastante facilidad usando la función jamovi 'IF' que ya hemos usado. Esta vez tenemos que especificar declaraciones 'IR' anidadas, lo que significa simplemente que SI la primera expresión lógica es VERDADERA, inserte un primer valor, pero SI una segunda expresión lógica es VERDADERA, inserte un segundo valor, pero SI una tercera expresión lógica es VERDADERA , luego inserte un tercer valor. Esto se puede escribir como:

IF(Edad >= 0 y Edad <= 20, 1, IF(Edad >= 21 y Edad <= 40, 2, IF(Edad >= 41 y Edad <= 60, 3)))

Ten en cuenta que se utilizan tres paréntesis izquierdos durante el anidamiento, por lo que todo el enunciado debe terminar con tres paréntesis derechos; de lo contrario, obtendrás un mensaje de error. La captura de pantalla jamovi para esta manipulación de datos, junto con una tabla de frecuencias adjunta, se muestra en Figure 6.6.

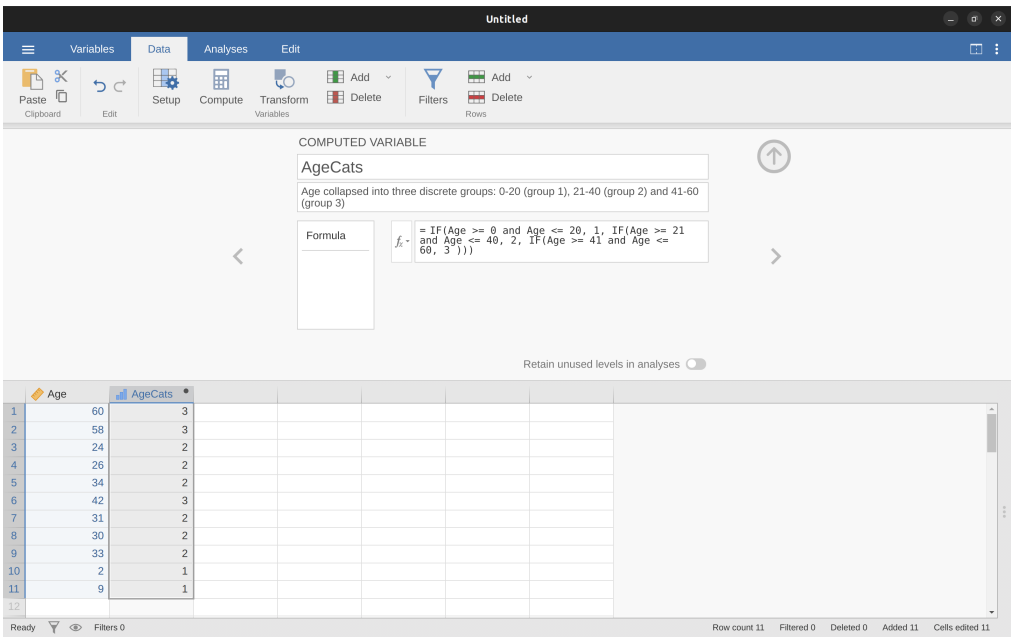


Figure 6.6: descomponer una variable en un número menor de niveles discretos usando la función 'IF' de jamovi

Es importante dedicar tiempo a averiguar si las categorías resultantes tienen algún sentido para tu proyecto de investigación. Si no tienen ningún sentido para ti como categorías significativas, es probable que cualquier análisis de datos que utilice esas categorías no tenga ningún sentido. De forma más general, en la práctica, he observado que la gente tiene un fuerte deseo de dividir sus datos (continuos y desordenados) en unas pocas categorías (discretas y simples), y luego ejecutar análisis utilizando los datos categorizados en lugar de los datos originales.[8] No me atrevería a decir que se trata

[8] si has leído más en el libro y estás releyendo esta sección, entonces un buen ejemplo de esto sería

de una idea intrínsecamente mala, pero a veces tiene algunos inconvenientes bastante graves, por lo que te aconsejaría cierta cautela si estás pensando en hacerlo.

6.3.3 Crear una transformación que puedea aplicarse a múltiples variables

A veces se desea aplicar la misma transformación a más de una variable, por ejemplo, cuando tienes varios elementos del cuestionario que deben volver a calcularse o codificarse de la misma manera. Y una de las características interesantes de jamovi es que puede crear una transformación, usando el botón 'Datos' - 'Transformar', que luego se puede guardar y aplicar a múltiples variables. Volvamos al primer ejemplo anterior, usando el archivo de datos likert.omv que contiene una sola variable con respuestas de escala Likert brutas para 10 personas. Para crear una transformación que puedas guardar y luego aplicar en múltiples variables (suponiendo que tengas más variables como esta en tu archivo de datos), primero en el editor de hojas de cálculo, selecciona (es decir, haz clic) la variable que quieres usar para crear inicialmente la transformación. En nuestro ejemplo, esto es likert.raw. A continuación, haz clic en el botón 'Transformar' en la cinta 'Datos' de jamovi y verás algo como Figure 6.7.

Asigna un nombre a tu nueva variable, llamémosla opinion.strength y luego haz clic en el cuadro de selección 'usar transformación' y selecciona 'Crear nueva transformación...'. Aquí es donde crearás y nombrarás la transformación que se puede volver a aplicar a tantas variables como quieras. La transformación se nombra automáticamente para nosotros como 'Transformar 1' (imaginativo, eh. Puedes cambiarlo si quieres). Luego escribe la expresión "ABS($source - 4)" en el cuadro de texto de la función, como en Figure 6.8, presiona Entrar o Retorno en tu teclado y listo, has creado una nueva transformación y la has aplicado a la variable likert.raw. Bien, eh. Ten en cuenta que en lugar de usar la etiqueta de la variable en la expresión, hemos usado '$source'. Esto es para que podamos usar la misma transformación con tantas variables diferentes como queramos; jamovi requiere que uses '$source' para referirte a la variable de origen que estás transformando. Tu transformación también se ha guardado y se puede reutilizar en cualquier momento que quieras (siempre que guardes el conjunto de datos como un archivo '.omv'; de lo contrario, ¡lo perderás!).

También puedes crear una transformación con el segundo ejemplo que vimos, la distribución por edades de las personas en una reunión social. ¡Adelante, sabes que quieres hacerlo! Recuerda que dividimos esta variable en tres grupos: joven, adulto y mayor. Esta vez vamos a conseguir lo mismo, pero usando el botón jamovi 'Transformar' - 'Agregar condición de recodificación'. Con este conjunto de datos (vuelve a él o créalo de nuevo si no lo guardaste) configura una nueva transformación de variable. Llama a la variable transformada AgeCats y la transformación que crearás Agegroupings. Luego haga clic en el gran signo "+" al lado del cuadro de función. Este es el botón 'Agregar condición' y he pegado una gran flecha roja en Figure 6.9 para que puedas ver exactamente dónde está. Vuelve a crear la transformación que se muestra en Figure 6.9 y cuando hayas terminado, verás que aparecen los nuevos valores en la ventana de la hoja

alguien que elige hacer un ANOVA usando AgeCats como la variable de agrupación, en lugar de ejecutar una regresión utilizando la edad como predictor. A veces hay buenas razones para hacer esto. Por ejemplo, si la relación entre la edad y tu variable de resultado es altamente no lineal y no te sientes cómoda intentando ejecutar una regresión no lineal. Sin embargo, a menos que realmente tengas una buena razón para hacerlo, es mejor no hacerlo. Tiende a introducir todo tipo de problemas (p. ej., los datos probablemente violarán la suposición de normalidad) y puedes perder mucho poder estadístico.

de cálculo. Además, la transformación de grupos de edad se ha guardado y se puede volver a aplicar en el momento que quieras. Ok, sé que es poco probable que tengas más de una variable 'Edad', pero ahora ya entiendes cómo configurar transformaciones en jamovi, así que puedes seguir esta idea con otros tipos de variables. Un escenario típico para esto es cuando tienes una escala de cuestionario con, digamos, 20 ítems (variables) y cada ítem se calificó originalmente de 1 a 6 pero, por alguna razón o peculiaridad de los datos, decides volver a codificar todos los ítems como 1 a 3. Puedes hacerlo fácilmente en jamovi creando y luego volviendo a aplicar tu transformación para cada variable que quieras recodificar.

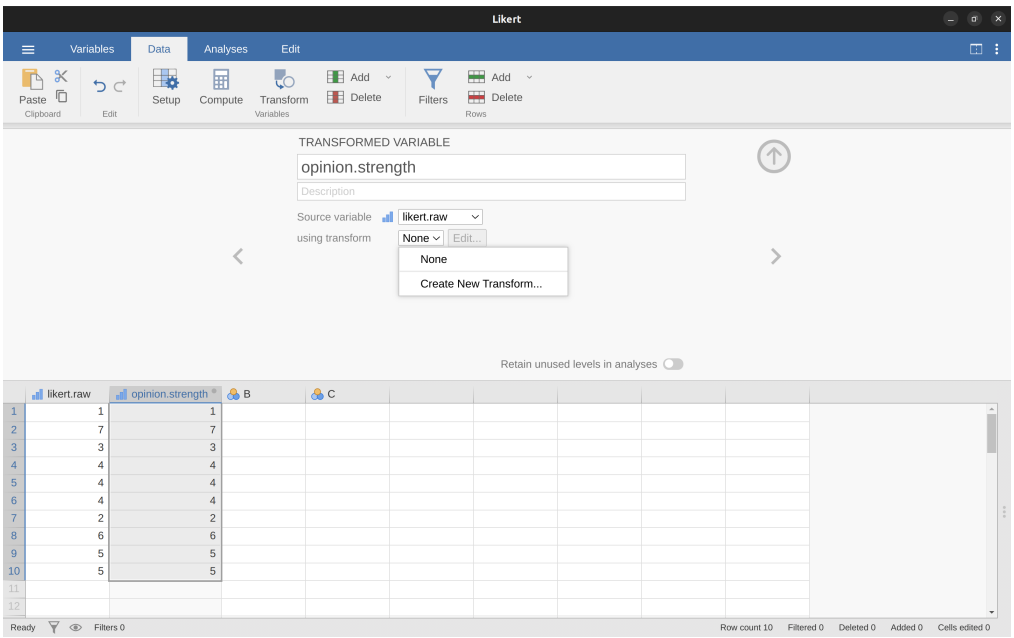


Figure 6.7: Creando una nueva transformación de variable usando el comando jamovi 'Transformar'

6.4 Otras funciones y operaciones matemáticas

En la sección sobre Transformar y recodificar una variable, analicé las ideas detrás de las transformaciones de variables y mostré que muchas de las transformaciones que podrías querer aplicar a tus datos se basan en funciones y operaciones matemáticas bastante simples. En esta sección quiero volver a esa discusión y mencionar otras funciones matemáticas y operaciones aritméticas que en realidad son bastante útiles para muchos análisis de datos del mundo real. Table 6.5 ofrece una breve descripción general de las diversas funciones matemáticas de las que quiero hablar aquí o más adelante. [9] Obviamente, esto ni siquiera se acerca a la catalogación de la gama de posibilidades disponibles, pero sí cubre una gama de funciones que se utilizan regularmente en el análisis de datos y que están disponibles en jamovi.

[9] Dejaremos la función box-cox para más adelante

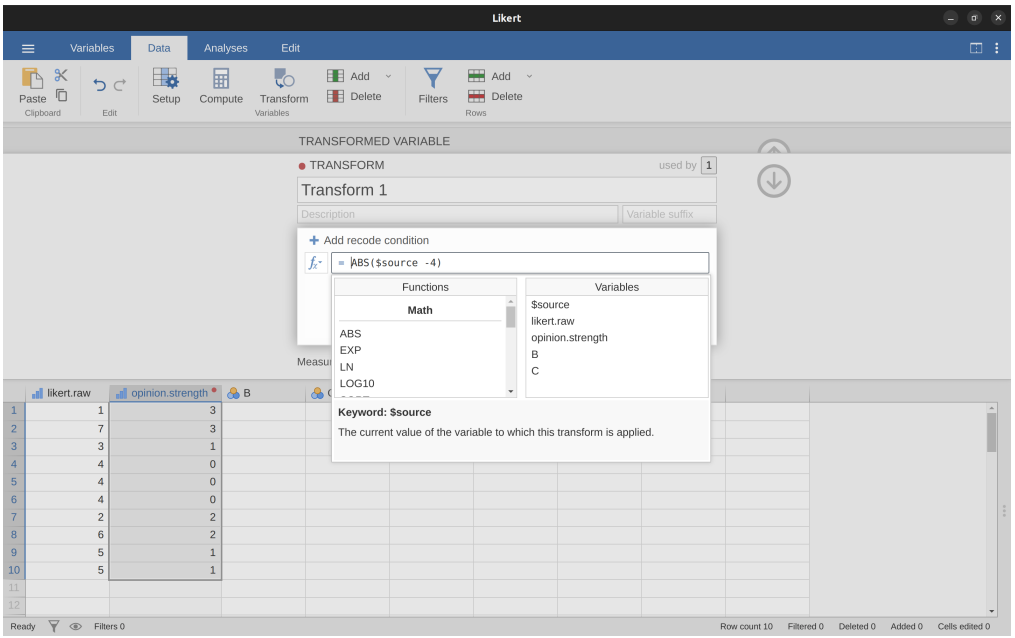


Figure 6.8: especificando una transformación en jamovi, para guardarla como la imaginativamente llamada 'Transform 1'

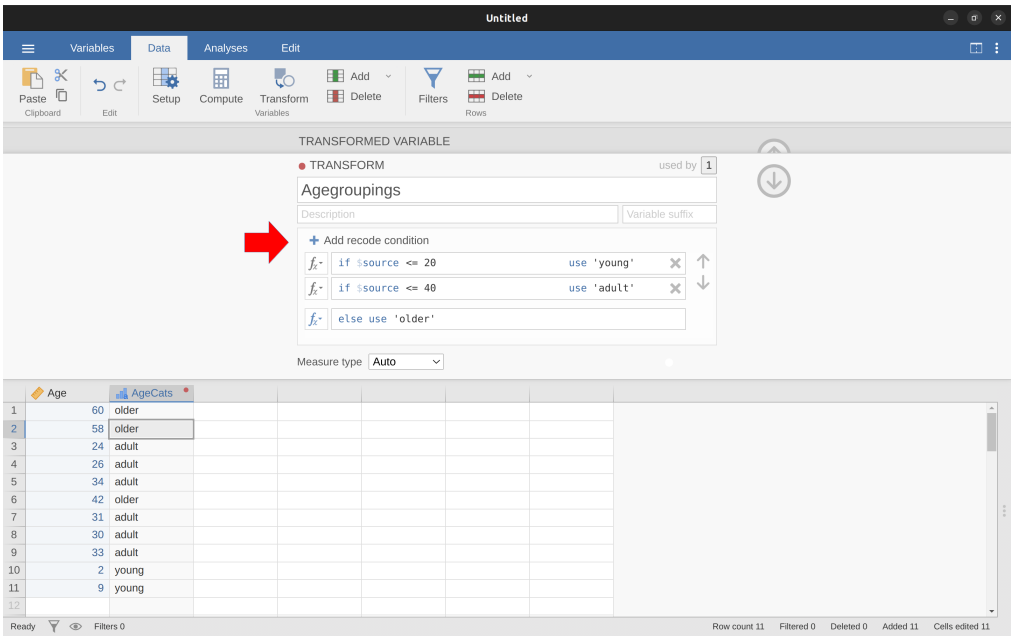


Figure 6.9: transformación jamovi en tres categorías de edad, usando el botón 'Agregar condición'

Table 6.5: algunos operadores matemáticos

	function	**example input**	**(answer)**
square root	SQRT(x)	SQRT(25)	5
absolute value	ABS(x)	ABS(-23)	23
logarithm (base 10)	LOG10(x)	LOG10(1000)	3
logarithm (base e)	LN(x)	LN(1000)	6.91
exponentiation	EXP(x)	EXP(6.908)	1e+03
box-cox	BOXCOX(x, lamda)	BOXCOX(6.908, 3)	110

6.4.1 Logaritmos y exponenciales

Como mencioné anteriormente, jamovi tiene una gama útil de funciones matemáticas incorporadas y no tendría mucho sentido intentar describirlas o enumerarlas todas. Me he centrado mayoritariamente en aquellas funciones que son estrictamente necesarias para este libro. Sin embargo, quiero hacer una excepción con logaritmos y exponenciales. Aunque no se necesitan en ninguna otra parte de este libro, están *en todas partes* en la estadística en general. Y no solo eso, hay *muchas* situaciones en las que es conveniente analizar el logaritmo de una variable (es decir, hacer una "transformación logarítmica" de la variable). Sospecho que muchos (quizás la mayoría) de los lectores de este libro se habrán encontrado con logaritmos y exponenciales antes, pero por experiencias pasadas sé que hay una proporción sustancial de estudiantes que asisten a una clase de estadística en ciencias sociales que no han tocado los logaritmos desde el instituto, y que agradecerían un repaso.

Para entender logaritmos y exponenciales, lo más fácil es calcularlos y ver cómo se relacionan con otros cálculos simples. Hay tres funciones jamovi en particular de las que quiero hablar, a saber, LN(), LOG10() y EXP(). Para empezar, consideremos LOG10(), que se conoce como el "logaritmo en base 10". El truco para entender un **logaritmo** es entender que es básicamente lo "opuesto" a una potencia. Específicamente, el logaritmo en base 10 está estrechamente relacionado con las potencias de 10. Comencemos notando que 10 al cubo es 1000. Matemáticamente, escribiríamos esto:

$$10^3 = 1000$$

El truco para entender un logaritmo es reconocer que la afirmación de que "10 elevado a 3 es igual a 1000" es equivalente a la afirmación de que "el logaritmo (en base 10) de 1000 es igual a 3". Matemáticamente, lo escribimos de la siguiente manera,

$$\log_{10}(1000) = 3$$

Bien, puesto que la función LOG10() está relacionada con las potencias de 10, es de esperar que haya otros logaritmos (en bases distintas de 10) que también estén rela-

cionados con otras potencias. Y, por supuesto, es cierto: en realidad el número 10 no tiene nada de especial desde el punto de vista matemático. Nos resulta útil porque los números decimales se construyen alrededor del número 10, pero el malvado mundo de las matemáticas se burla de nuestros números decimales. Lamentablemente, al universo no le importa cómo escribimos los números. La consecuencia de esta indiferencia cósmica es que no tiene nada de especial calcular logaritmos en base 10. Podrías, por ejemplo, calcular tus logaritmos en base 2. Alternativamente, un tercer tipo de logaritmo, y que vemos mucho más en estadística que la base 10 o la base 2, se llama **logaritmo natural** y corresponde al logaritmo en base e. Como es posible que algún día te encuentres con él, mejor te explico qué es e. El número e, conocido como **número de Euler**, es uno de esos molestos números "irracionales" cuya expansión decimal es infinitamente larga, y se considera uno de los números más importantes de las matemáticas. Los primeros dígitos de e son:

$$e \approx 2.718282$$

Hay bastantes situaciones en estadística que requieren que calculemos potencias de e, aunque ninguna de ellas aparece en este libro. Elevar e a la potencia x se denomina **exponencial** de x, por lo que es muy común ver e^x escrito como $\exp(x)$. Así que no es de extrañar que jamovi tenga una función que calcula exponenciales, llamada EXP(). Como el número e aparece tan a menudo en estadística, el logaritmo natural (es decir, el logaritmo en base e) también tiende a aparecer. Los matemáticos suelen escribirlo como $\log_e(x)$ o $\ln(x)$. De hecho, jamovi funciona de la misma manera: la función LN() corresponde al logaritmo natural.

Y con eso, creo que ya hemos tenido suficientes exponenciales y logaritmos para este libro.

6.5 Extracción de un subconjunto de datos

Un tipo de tratamiento de datos muy importante es poder extraer un subconjunto concreto de datos. Por ejemplo, es posible que solo te interese analizar los datos de una condición experimental, o puedes querer observar de cerca los datos de personas mayores de 50 años. Para hacer esto, el primer paso es hacer que jamovi filtre el subconjunto de datos correspondientes a las observaciones que te interesan.

Esta sección vuelve al conjunto de datos nightgarden.csv. Si estás leyendo todo este capítulo de una sola vez, entonces ya deberías tener este conjunto de datos cargado en una ventana jamovi. Para esta sección, vamos a centrarnos en las dos variables locutor y enunciado (consulta Tabulación y tabulación cruzada de datos si has olvidado cómo son estas variables). Supongamos que lo que queremos es extraer solo las frases pronunciadas por Makka-Pakka. Para ello, tenemos que especificuniversity holidaysar un filtro en jamovi. En primer lugar, abre una ventana de filtro haciendo clic en 'Filtros' en la barra de herramientas principal de 'Datos' de jamovi. A continuauniversity holidaysción, en el cuadro de texto 'Filtro 1', junto al signo '=', escribe lo siguiente:

locutor == 'makka-pakka'

Cuando lo hayas hecho, verás que se ha añadido una nueva columna a la ventana de la hoja de cálculo (véase Figure 6.10), etiquetada como 'Filtro 1', con los casos en los

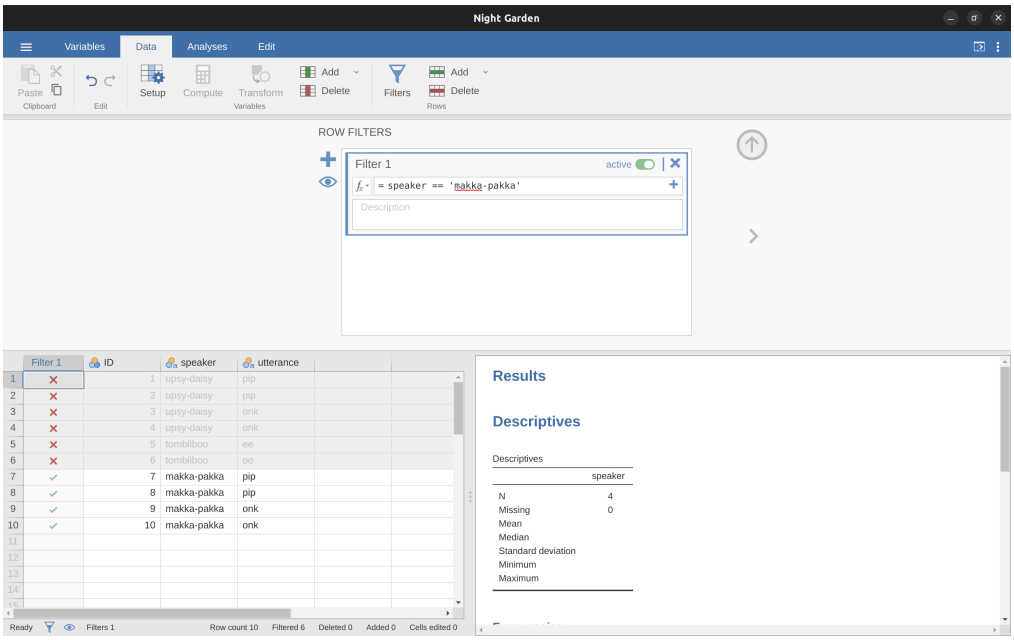


Figure 6.10: Creando un subconjunto de datos nightgarden usando la opción 'Filtros' de jamovi

que el locutor no es 'makka-pakka' en gris (es decir, filtrado) y, por el contrario, los casos en los que el locutor es 'makka-pakka' tienen una marca de verificación verde que indica que están filtrados. Puedes comprobarlo ejecutando 'Exploración' - 'Descriptivos' - 'Tablas de frecuencia' para la variable locutor y ver qué muestra. Pruébalo.

A partir de este sencillo ejemplo, también puedes crear filtros más complejos utilizando expresiones lógicas en jamovi. Por ejemplo, supongamos que quisieras mantener solo aquellos casos en los que el enunciado es "pip" o "oo". En este caso, en el cuadro de texto 'Filtro 1', junto al signo '=', escribirías lo siguiente:

enunciado == 'pip' o enunciado == 'oo'

6.6 Resumen

Obviamente, este capítulo no tiene ninguna coherencia. No es más que un conjunto de temas y trucos que puede ser útil conocer, así que lo mejor que puedo hacer es repetir esta lista:

- Tabulación y tabulación cruzada de datos
- Expresiones lógicas en jamovi
- Transformar y recodificar una variable
- Otras funciones y operaciones matemáticas
- [Extracción de un subconjunto de los datos]

Part IV

Teoría estadística

Sobre los límites del razonamiento lógico

Preludio {.unnumbered}

La Parte IV del libro es, con mucho, la más teórica, ya que se centra sobre la teoría de la inferencia estadística. Durante los próximos tres capítulos mi El objetivo es brindarle una Introducción a la probabilidad teoría, muestreo y estimación en el capítulo sobre Estimación de cantidades desconocidas de una muestra y estadística Prueba de hipótesis. Sin embargo, antes de comenzar, quiero para decir algo sobre el panorama general. La inferencia estadística es principalmente sobre el aprendizaje de los datos. El objetivo ya no es simplemente describir nuestros datos, sino utilizar los datos para sacar conclusiones sobre el mundo. Para motivar la discusión quiero pasar un poco de tiempo hablando sobre un rompecabezas filosófico conocido como el acertijo de la inducción, porque habla de un problema que aparecerá una y otra vez a lo largo el libro: la inferencia estadística se basa en suposiciones. esto suena como una cosa mala. En la vida cotidiana, la gente dice cosas como y las clases de psicología a menudo hablan de suposiciones. y sesgos como cosas malas que debemos tratar de evitar. de amargo experiencia personal he aprendido a nunca decir tales cosas alrededor filósofos!

> *Todo el arte de la guerra consiste en llegar a lo que está del otro lado de la colina, o, en otras palabras, en aprender lo que no sabemos de lo que hacemos.* - Arthur Wellesley, primer duque de Wellington

Me dijeron que la cita anterior surgió como consecuencia de un carruaje cabalga por el campo.[10] Él y su compañero, J. W. Croker, estaban jugando un juego de adivinanzas, cada uno tratando de predecir qué estaría al otro lado de cada colina. En todos los casos resultó que Wellesley tenía razón y Croker estaba equivocado. Muchos años después cuando Cuando se le preguntó a Wellesley sobre el juego, explicó que. De hecho, la guerra no es especial a este respecto. Toda la vida es una adivinanza juego de una forma u otra, y sobrellevar el día a día requiere que hagamos buenas conjeturas. Así que digamos que W se refiere a una victoria de Wellesley y C se refiere a una victoria de Croker. Después de tres colinas, nuestro conjunto de datos parece como esto:

WWW

Nuestra conversación es así:

> tu: tres seguidos no es un poco de un jugador. Es informativo y no veo

[10] http://www.bartleby.com/344/400.html

> razón para preferir Wellesleys. Puedo organizar los datos en bloques de tres para que puedas ver qué lote corresponde a las observaciones que teníamos disponibles en cada paso en nuestro pequeño juego secundario. Después de ver este nuevo lote, nuestra conversación continúa:
>
> tú: Seis victorias seguidas para Duke Wellesley. Esto está empezando a sentirse un un poco sospechoso Va a ganar el siguiente también.
>
> yo: Supongo que no veo ninguna razón lógica por la que eso significa que está bien con mi elección.

Por segunda vez tuviste razón, y por segunda vez yo me equivoqué. Wellesley gana las siguientes tres colinas, extendiendo su récord de victorias contra Croker a 9-0. El conjunto de datos disponible para nosotros ahora es este: *WWW WWW WWW* Y nuestra conversación es así:

> tu: Bien, esto es bastante obvio. Wellesley es mucho mejor en este juego. Ambos coincidimos en que habría dicho que todas eran igualmente probables. yo Asume que tú también lo habrías hecho, ¿verdad? Quiero decir, que no tienes idea?
>
> tu:supongo que si
>
> yo: pues entonces el observaciones que hemos encontrado hasta ahora, ares cambiado ¿después? Al comienzo de nuestro juego, te has encontrado con discriminado entre estas dos posibilidades. Por lo tanto, estos dos posibilidades siguen siendo igualmente plausibles y no veo ninguna razón lógica para prefieren uno sobre el otro. Así que sí, aunque estoy de acuerdo contigo en que Wellesleyt pensar en un buena razón para pensar que todavía está dispuesto a correr el riesgo. Su racha ganadora continúa durante las próximas tres colinas. El puntaje en el juego Wellesley-Croker ahora es 12-0, y el puntaje en nuestro juego ahora es 3-0. A medida que nos acercamos a la cuarta ronda de nuestro juego, nuestro conjunto de datos es este: *WWW WWW WWW WWW* y la conversación continúa:
>
> tu: ah si! Tres victorias más para Wellesley y otra victoria para mí. ¡Admítelo, tenía razón sobre él! Supongo que mojado sé qué pensar. me siento como húmedo ya descartado, *WWW WWW WWW WWW C* y *WWW WWW WWW WWW W*. ¿Son estos dos? igualmente sensato dado que nuestras observaciones aportan la evidencia lógica de que la racha continuará?
>
> tu: creo que eres el experto en estadisticas y estás perdiendo. Estoy ganando. Tal vez tú debería cambiar de estrategia.
>
> yo: Hmm, ese es un buen punto y no me temo que no he observado que es una serie de tres victorias para ti. Sus datos se verían así: *YYY*. Lógicamente, yo no pareces mucho evidencia, y no veo ninguna razón para pensar que su estrategia está funcionando nada mejor que el mío. Si no fuera mejor en ¿nuestro?
>
> tú: Bien, ahora creo que ves la evidencia lógica de eso.

Aprender sin hacer suposiciones es un mito

Hay muchas maneras diferentes en las que podríamos diseccionar este diálogo, pero dado que este es un libro de estadísticas dirigido a psicólogos y no una introducción a la filosofía y psicología del razonamiento, Lo que he descrito anteriormente se refiere a veces a como el enigma de la inducción. Parece totalmente razonable pensar que un El récord de victorias de 12-0 de Wellesley es una evidencia bastante sólida de que lo hará. ganar el juego 13, pero no es fácil proporcionar una lógica adecuada justificación de esta creencia. Por el contrario, a pesar de la obviedad de la respuesta, no tiene ninguna justificación lógica de.

El enigma de la inducción está más asociado con el trabajo filosófico de David Hume y, más recientemente, de Nelson Goodman, pero puedes encontrar ejemplos del problema que surge en campos tan diversos como la literatura (Lewis Carroll) y aprendizaje automático (el teorema). Realmente hay algo raro en tratar de. El punto crítico es que las suposiciones y los sesgos son inevitables si quieres aprender algo sobre el mundo. No hay escape de esto, y es tan cierto para la estadística inferencia como lo es para el razonamiento humano. En el diálogo yo apuntaba en sus inferencias perfectamente sensibles como un ser humano, pero el común sentido de razonamiento en el que confiaste no es diferente a lo que un hubiera hecho un estadístico. Tu mitad del diálogo se basó en una suposición implícita de que existe alguna diferencia en habilidad entre Wellesley y Croker, y lo que estabas haciendo era intentar para averiguar cuál sería esa diferencia en el nivel de habilidad. My rechaza esa suposición por completo. Todo lo que estaba dispuesto a aceptar es que hay secuencias de victorias y derrotas y eso no lo sabia qué secuencias se observarían. A lo largo del diálogo mantuve insistiendo en que todos los conjuntos de datos lógicamente posibles eran igualmente plausibles al comienzo del juego Wellesely-Croker, y la única forma en que puedo alguna vez revisé mis creencias fue eliminar esas posibilidades que eran objetivamente inconsistente con las observaciones.

Eso suena perfectamente sensato en sus propios términos. De hecho, incluso suena como el sello distintivo del buen razonamiento deductivo. Como Sherlock Holmes, mi enfoque era descartar lo que es imposible con la esperanza de que lo que quedaría es la verdad. Sin embargo, como vimos, descartando lo imposible nunca me llevó a hacer una predicción. En sus propios términos todo lo que dije en mi mitad del diálogo fue completamente correcta. Una incapacidad para hacer cualquier predicciones es la consecuencia lógica de hacer. En al final perdí nuestro juego porque hiciste algunas suposiciones y esas las suposiciones resultaron ser correctas. La habilidad es una cosa real, y porque creíste en la existencia de la habilidad fuiste capaz de aprender que Wellesley tenía más que Croker. ¿Había confiado usted en un menos sensato suposición para impulsar su aprendizaje, es posible que no haya ganado el juego.

En última instancia, hay dos cosas que debes quitar de esto. Primero, como a menudo señalaré las suposiciones que sustentan una técnica estadística particular, y cómo se puede verificar si esos las suposiciones son sensatas.

Chapter 7

Introducción a la probabilidad

> *[Dios] nos ha concedido sólo el crepúsculo… de la Probabilidad.*
> – John Locke

Hasta este punto del libro, hemos tratado algunas de las ideas clave del diseño experimental y hemos hablado un poco de cómo resumir un conjunto de datos. Para mucha gente, esto es todo lo que hay en estadística: recopilar todos los números, calcular las medias, hacer dibujos y ponerlos en un informe en algún sitio. Es como coleccionar sellos pero con números. Sin embargo, la estadística abarca mucho más que eso. De hecho, la estadística descriptiva es una de las partes más pequeñas de la estadística y una de las menos potentes. La parte más importante y útil de la estadística es que proporciona información que permite hacer inferencias sobre los datos.

Una vez que empiezas a pensar en la estadística en estos términos, que la estadística está ahí para ayudarnos a sacar conclusiones de los datos, empiezas a ver ejemplos de ello en todas partes. Por ejemplo, aquí hay un pequeño extracto de un artículo de periódico del Sydney Morning Herald (30 de octubre de 2010):

> "Tengo un trabajo difícil", dijo la Primera Ministra en respuesta a una encuesta que reveló que su gobierno es ahora la administración laborista más impopular de la historia, con un voto en las primarias de solo el 23 por ciento.

Este tipo de comentario es totalmente anodino en los periódicos o en la vida cotidiana, pero pensemos en lo que implica. Una empresa de sondeos ha realizado una encuesta, por lo general bastante grande porque se lo puede permitir. Me da pereza buscar la encuesta original, así que imaginemos que han llamado al azar a 1000 votantes de New South Wales (NSW), y 230 (23%) de ellos afirman que tienen la intención de votar por el Partido Laborista Australiano (ALP). En las elecciones federales de 2010, la Comisión Electoral Australiana informó de 4.610.795 votantes inscritos en NSW, por lo que desconocemos las opiniones de los 4.609.795 votantes restantes (alrededor del 99,98 % de los votantes). Incluso suponiendo que nadie mintiera a la empresa encuestadora, lo único que podemos afirmar con un 100 % de seguridad es que el verdadero voto del ALP en las primarias se sitúa entre 230/4610795 (alrededor del 0,005%) y 4610025/4610795 (alrededor del 99,83%). Entonces, ¿en qué se basan la empresa de encuestas, el periódico y los lectores para llegar a la conclusión de que el voto primario del ALP es solo del

23%?

La respuesta a la pregunta es bastante obvia. Si llamo a 1000 personas al azar y 230 de ellas dicen que tienen intención de votar al ALP, parece muy poco probable que estas sean las únicas 230 personas de todo el público votante que realmente tienen la intención de votar por ALP. En otras palabras, asumimos que los datos recopilados por la empresa encuestadora son bastante representativos de la población en general. Pero, ¿hasta qué punto? ¿Nos sorprendería descubrir que el verdadero voto ALP en las primarias es en realidad el 24%? 29%? 37%? En este punto, la intuición cotidiana empieza a fallar un poco. Nadie se sorprendería del 24 % y todo el mundo se sorprendería del 37 %, pero es un poco difícil decir si el 29 % es plausible. Necesitamos herramientas más potentes que mirar los números y adivinar.

La estadística inferencial nos proporciona las herramientas que necesitamos para responder este a tipo de preguntas y, dado que este tipo de preguntas constituyen el núcleo de la empresa científica, ocupan la mayor parte de los cursos introductorios sobre estadística y métodos de investigación. Sin embargo, la teoría de la inferencia estadística se basa en la **teoría de la probabilidad**. Y es a la teoría de la probabilidad a la que debemos referirnos ahora. Esta discusión de la teoría de la probabilidad es básicamente un detalle de fondo. No hay mucha estadística en sí en este capítulo, y no es necesario comprender este material con tanta profundidad como los otros capítulos de esta parte del libro. Sin embargo, dado que la teoría de la probabilidad sustenta gran parte de la estadística, merece la pena cubrir algunos de los aspectos básicos.

7.1 ¿En qué se diferencian la probabilidad y la estadística?

Antes de empezar a hablar de la teoría de la probabilidad, conviene dedicar un momento a reflexionar sobre la relación entre probabilidad y estadística. Ambas disciplinas están estrechamente relacionadas pero no son idénticas. La teoría de la probabilidad es "la doctrina de las probabilidades". Es una rama de las matemáticas que nos dice con qué frecuencia ocurrirán diferentes tipos de sucesos. Por ejemplo, todas estas preguntas pueden responderse usando la teoría de la probabilidad:

- ¿Qué probabilidad hay de que una moneda salga cara 10 veces seguidas?
- Si tiro un dado de seis caras dos veces, ¿qué probabilidad hay de que saque dos seises?
- ¿Qué probabilidad hay de que cinco cartas extraídas de una baraja perfectamente barajada sean todas corazones?
- ¿Qué probabilidad hay de que gane la lotería?

Fíjate que todas estas preguntas tienen algo en común. En cada caso, se conoce la "verdad del mundo" y mi pregunta se refiere a "qué tipo de sucesos" ocurrirán. En la primera pregunta, sé que la moneda es justa, por lo que hay un 50% de probabilidades de que salga cara. En la segunda pregunta, sé que la probabilidad de sacar un 6 en un solo dado es de 1 entre 6. En la tercera pregunta, sé que la baraja se barajó correctamente. Y en la cuarta pregunta sé que la lotería sigue unas reglas específicas. Entiendes la idea. El punto crítico es que las preguntas probabilísticas empiezan con un **modelo** conocido del mundo, y usamos ese modelo para hacer algunos cálculos. El modelo subyacente puede ser bastante simple. Por ejemplo, en el ejemplo del lanzamiento de una moneda,

podemos escribir el modelo así:

$$P(cara) = 0.5$$

que se puede leer como "la probabilidad de que salga cara es 0,5". Como veremos más adelante, del mismo modo que los porcentajes son números que van del 0% al 100%, las probabilidades son números que van del 0 al 1. Cuando utilizo este modelo de probabilidad para responder a la primera pregunta, en realidad no sé exactamente lo que va a pasar. Puede que salga, como dice la pregunta. Pero tal vez obtenga tres caras. Esa es la clave. En la teoría de la probabilidad se conoce el modelo, pero no los datos.

Eso es probabilidad. ¿Y la estadística? Las preguntas estadísticas funcionan al revés. En estadística no sabemos la verdad sobre el mundo. Lo único que tenemos son los datos y es a partir de ellos que queremos saber la verdad sobre el mundo. Las preguntas estadísticas tienden a parecerse más a estas:

- Si mi amigo lanza una moneda 10 veces y sale 10 caras, ¿me está gastando una broma?
- Si cinco cartas de la parte superior de la baraja son corazones, ¿qué probabilidad hay de que la baraja se haya barajado?
- Si el cónyuge del comisario de lotería gana la lotería, ¿qué probabilidad hay de que la lotería estuviera amañada?

Esta vez lo único que tenemos son datos. Lo que sé es que vi a mi amigo lanzar la moneda 10 veces y que salió cara todas las veces. Y lo que quiero deducir es si debo o no concluir que lo que acabo de ver era realmente una moneda justa lanzada 10 veces seguidas, o si debo sospechar que mi amigo me está gastando una broma. Los datos que tengo son los siguientes:

HHHHHHHHHHHH

y lo que intento es averiguar en qué "modelo del mundo" debo confiar. Si la moneda es justa, entonces el modelo que debo adoptar es el que dice que la probabilidad de que salga cara es 0.5, es decir P(cara) = 0,5. Si la moneda no es justa, debo concluir que la probabilidad de cara no es 0,5, lo que escribiríamos como $P(cara) \neq 0,5$. En otras palabras, el problema de la inferencia estadística consiste en averiguar cuál de estos modelos de probabilidad es correcto. Evidentemente, la pregunta estadística no es la misma que la pregunta de probabilidad, pero están profundamente conectadas entre sí. Debido a esto, una buena introducción a la teoría estadística comenzará con una discusión de lo que es la probabilidad y cómo funciona.

7.2 ¿Qué significa probabilidad?

Empecemos con la primera de estas preguntas. ¿Qué es "probabilidad"? Puede parecerte sorprendente, pero aunque los estadísticos y los matemáticos (en su mayoría) están de acuerdo en cuáles son las reglas de la probabilidad, hay mucho menos consenso sobre el significado real de la palabra. Parece extraño porque todos nos sentimos muy cómodos usando palabras como "azar", "posible" y "probable", y no parece que sea una pregunta muy difícil de responder. Pero si alguna vez has tenido esa experiencia en la vida real, es posible que se salgas de la conversación con la sensación de que no lo

has entendido del todo y que (como muchos conceptos cotidianos) resulta que no sabes realmente de qué se trata.

Así que voy a intentarlo. Supongamos que quiero apostar en un partido de fútbol entre dos equipos de robots, el Arduino Arsenal y el C Milan. Después de pensarlo, decido que hay un 80% de probabilidad de que el Arduino Arsenal gane. ¿Qué quiero decir con eso? Aquí hay tres posibilidades:

- Son equipos de robots, así que puedo hacer que jueguen una y otra vez, y si lo hiciera, el Arduino Arsenal ganaría 8 de cada 10 juegos en promedio.
- Para cualquier partido, estaría de acuerdo en que apostar en este partido solo es "justo" si una apuesta de $1 al C Milan da un beneficio de $5 (es decir, recupero mi $1 más una recompensa de $4 por acertar), al igual que una apuesta de $4 al Arduino Arsenal (es decir, mi apuesta de $4 más una recompensa de $1).
- Mi "creencia" o "confianza" subjetiva en una victoria del Arduino Arsenal es cuatro veces mayor que mi creencia en una victoria del C Milan.

Cada una de ellas parece sensata. Sin embargo, no son idénticas y no todos los estadísticos las respaldarían todas. La razón es que existen diferentes ideologías estadísticas (sí, de verdad) y dependiendo de a cuál te suscribas, podrías decir que algunas de esas afirmaciones no tienen sentido o son irrelevantes. En esta sección presento brevemente los dos enfoques principales que existen en la literatura. No son ni mucho menos los únicos enfoques, pero son los dos grandes.

7.2.1 La vista frecuentista

El primero de los dos enfoques principales de la probabilidad, y el más dominante en estadística, se conoce como la **visión frecuentista** y define la probabilidad como una **frecuencia a largo plazo**. Supongamos que intentamos lanzar una moneda al aire una y otra vez. Por definición, se trata de una moneda que tiene $P(C) = 0,5$. ¿Qué podríamos observar? Una posibilidad es que los primeros 20 lanzamientos tengan este aspecto:

T,H,H,H,H,T,T,H,H,H,H,T,H,H,T,T,T,T,T,H

En este caso, 11 de estos 20 lanzamientos (55%) han salido cara. Supongamos ahora que llevo la cuenta del número de caras (que llamaré N_H) que he visto, en los primeros N lanzamientos, y calculo la proporción de caras $\frac{N_H}{N}$ cada vez. La Table 7.1 muestra lo que obtendría (literalmente lancé monedas para producir esto):

Observa que al principio de la secuencia, la *proporción* de caras fluctúa enormemente, empezando por .00 y subiendo hasta .80. Más tarde, se tiene la impresión de que se atenúa un poco, y cada vez más los valores se acercan a la respuesta "correcta" de .50. Esta es la definición frecuentista de probabilidad en pocas palabras. Lanzar una moneda una y otra vez, y a medida que N crece (se acerca al infinito, denotado $N \rightarrow \infty$) la proporción de caras convergerá al 50%. Hay algunos tecnicismos sutiles que preocupan a los matemáticos, pero cualitativamente hablando, así es como los frecuentistas definen la probabilidad. Desafortunadamente, no tengo un número infinito de monedas o la paciencia infinita necesaria para lanzar una moneda un número infinito de veces. Sin embargo, tengo un ordenador y los ordenadores destacan en tareas repetitivas sin sentido. Así que le pedí a mi ordenador que simulara lanzar una moneda 1000 veces y luego hice un dibujo de lo que ocurre con la proporción $\frac{N_H}{N}$ a medida que aumenta N. De hecho, lo hice cuatro veces para asegurarme de que no fuera una casualidad. Los resultados se

Table 7.1: Lanzamiento de monedas y proporción de caras

number of flips	number of heads	proportion
1	0	0.00
2	1	0.50
3	2	0.67
4	3	0.75
5	4	0.80
6	4	0.67
7	4	0.57
8	5	0.63
9	6	0.67
10	7	0.70
11	8	0.73
12	8	0.67
13	9	0.69
14	10	0.71
15	10	0.67
16	10	0.63
17	10	0.59
18	10	0.56
19	10	0.53
20	11	0.55

muestran en Figure 7.1. Como puedes ver, la proporción de caras observadas deja de fluctuar y se estabiliza. Cuando lo hace, el número en el que finalmente se asienta es la verdadera probabilidad de caras.

La definición frecuentista de probabilidad tiene algunas características deseables. En primer lugar, es objetiva. La probabilidad de un suceso está *necesariamente* fundamentada en el mundo. Las afirmaciones sobre la probabilidad sólo tienen sentido si se refieren a (una secuencia de) sucesos que ocurren en el universo físico.[1] En segundo lugar, no es ambigua. Si dos personas observan la misma secuencia de acontecimientos e intentan calcular la probabilidad de un suceso, inevitablemente obtendrán la misma respuesta.

Sin embargo, también tiene características indeseables. En primer lugar, las secuencias infinitas no existen en el mundo físico. Supongamos que cogemos una moneda del bolsillo y empezamos a lanzarla. Cada vez que cae, impacta contra el suelo. Cada impacto desgasta un poco la moneda. Al final, la moneda se destruye. Por tanto, cabe preguntarse si realmente tiene sentido pretender que una secuencia "infinita" de lanzamientos de monedas es siquiera un concepto significativo u objetivo. No podemos decir que una "secuencia infinita" de sucesos sea algo real en el universo físico, porque el universo físico no permite nada infinito. Y lo que es más grave, la definición frecuentista

[1] Esto no significa que los frecuentistas no puedan hacer afirmaciones hipotéticas, por supuesto. Lo que ocurre es que si se quiere hacer una afirmación sobre la probabilidad, debe ser posible volver a describir esa afirmación en términos de una secuencia de sucesos potencialmente observables, junto con las frecuencias relativas de los distintos resultados que aparecen dentro de esa secuencia.

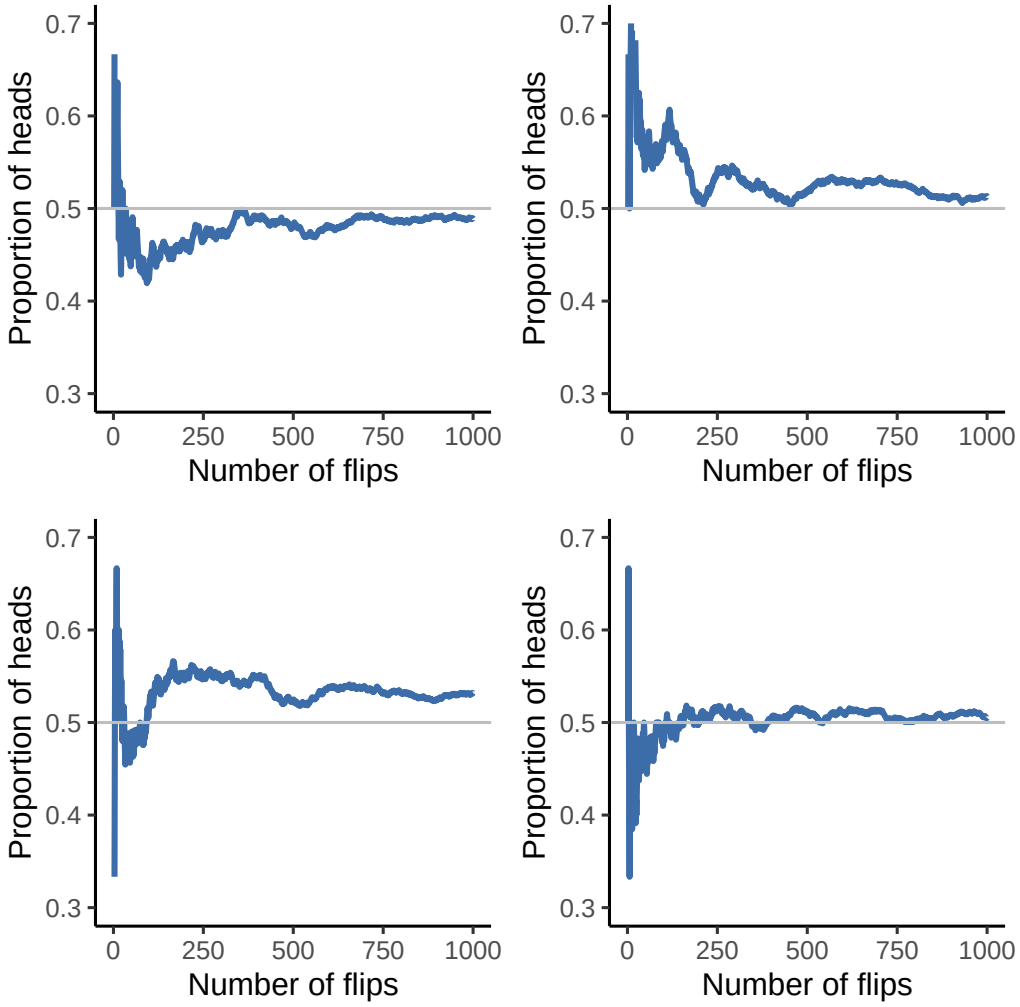


Figure 7.1: Una ilustración de cómo funciona la probabilidad frecuentista. Si lanzas una moneda al aire una y otra vez, la proporción de caras que has visto se estabiliza y converge a la probabilidad real de 0.5. Cada panel muestra cuatro experimentos simulados diferentes. En cada caso simulamos que lanzamos una moneda 1000 veces y llevamos la cuenta de la proporción de lanzamientos que salieron cara a medida que avanzábamos. Aunque en realidad ninguna de estas secuencias terminó con un valor exacto de .5, si hubiéramos ampliado el experimento a un número infinito de lanzamientos de la moneda, hubiera llegado a ese valor

tiene un alcance limitado. Hay muchas cosas a las que los seres humanos asignan probabilidades en el lenguaje cotidiano, pero que no pueden (ni siquiera en teoría) asignarse a una secuencia hipotética de sucesos. Por ejemplo, si un meteorólogo sale en la televisión y dice "la probabilidad de que llueva en Adelaide el 2 de noviembre de 2048 es del 60%", los seres humanos lo aceptamos de buen grado. Pero no está claro cómo definir esto en términos frecuentistas. Solo hay una ciudad, Adelaide, y solo un 2 de noviembre de 2048. Aquí no hay una secuencia infinita de eventos, solo algo único. La probabilidad frecuentista nos *prohibe* hacer afirmaciones probabilísticas sobre un único suceso. Desde la perspectiva frecuentista, mañana lloverá o no lloverá. No existe una "probabilidad" asociada a un único suceso no repetible. Ahora bien, hay que decir que los frecuentistas pueden utilizar algunos trucos muy ingeniosos para evitar esto. Una posibilidad es que lo que quiere decir el meteorólogo sea algo así como "Hay una categoría de días para los que predigo un 60% de probabilidad de lluvia, y si nos fijamos solo en los días para los que hago esta predicción, entonces el 60% de esos días lloverá de verdad". Es muy extraño y contraintuitivo pensar de este modo, pero los frecuentistas a veces lo hacen. Y aparecerá más adelante en este libro (por ejemplo, en Section 8.5).

7.2.2 La vista bayesiana

El punto de vista bayesiano de la probabilidad suele denominarse subjetivista y, aunque ha sido un punto de vista minoritario entre los estadísticos, ha ido ganando terreno en las últimas décadas. Hay muchos tipos de bayesianismo, por lo que es difícil decir exactamente cuál es "la" visión bayesiana. La forma más común de pensar en la probabilidad subjetiva es definir la probabilidad de un acontecimiento como el **grado de creencia** que un agente inteligente y racional asigna a la verdad de ese suceso. Desde esa perspectiva, las probabilidades no existen en el mundo sino en los pensamientos y suposiciones de las personas y otros seres inteligentes.

Sin embargo, para que este enfoque funcione necesitamos alguna forma de operacionalizar el "grado de creencia". Una forma de hacerlo es formalizarlo en términos de "juego racional", aunque hay muchas otras formas. Supongamos que creo que hay un 60% de probabilidades de que llueva mañana. Si alguien me ofrece una apuesta en la que si llueve mañana gano $5, pero si no llueve pierdo $5 está claro que, desde mi perspectiva, es una apuesta bastante buena. En cambio, si creo que la probabilidad de que llueva es solo del 40%, entonces es una mala apuesta. Así que podemos operacionalizar la noción de "probabilidad subjetiva" en términos de qué apuestas estoy dispuesta a aceptar.

¿Cuáles son las ventajas y desventajas del enfoque bayesiano? La principal ventaja es que permite asignar probabilidades a cualquier suceso que se desee. No es necesario limitarse a los sucesos repetibles. La principal desventaja (para mucha gente) es que no podemos ser puramente objetivos. Especificar una probabilidad requiere que especifiquemos una entidad que tenga el grado de creencia relevante. Esta entidad puede ser un ser humano, un extraterrestre, un robot o incluso un estadístico. Pero tiene que haber un agente inteligente que crea en las cosas. Para mucha gente esto es incómodo, parece que hace que la probabilidad sea arbitraria. Aunque el enfoque bayesiano exige que el agente en cuestión sea racional (es decir, que obedezca a las reglas de la probabilidad), permite que cada uno tenga sus propias creencias. Yo puedo creer que la moneda es justa y tú no, aunque ambos seamos racionales. La visión frecuentista no permite que dos observadores atribuyan diferentes probabilidades a un mismo suceso. Cuando eso

ocurre, al menos uno de ellos debe estar equivocado. La visión bayesiana no impide que esto ocurra. Dos observadores con diferentes conocimientos previos pueden legítimamente tener diferentes creencias sobre el mismo suceso. En resumen, mientras que la visión frecuentista a veces se considera demasiado estrecha (prohibe muchas cosas a las que queremos asignar probabilidades), la visión bayesiana a veces se considera demasiado amplia (permite demasiadas diferencias entre los observadores).

7.2.3 ¿Cual es la diferencia? ¿Y quién tiene razón?

Ahora que has visto cada uno de estos dos puntos de vista de forma independiente, es útil que te asegures que puedes compararlos. Vuelve al hipotético partido de fútbol de robots del principio de la sección. ¿Qué crees que dirían un frecuentista y un bayesiano sobre estas tres afirmaciones? ¿Qué afirmación diría un frecuentista que es la definición correcta de probabilidad? ¿Por cuál optaría un bayesiano? ¿Algunas de estas afirmaciones carecerían de sentido para un frecuentista o un bayesiano? Si has entendido las dos perspectivas, deberías tener alguna idea de cómo responder a estas preguntas.

Bien, suponiendo que entiendas la diferencia, te estarás preguntando cuál de ellos *tiene razón*. Sinceramente, no sé si hay una respuesta correcta. Hasta donde yo sé, no hay nada matemáticamente incorrecto en la forma en que los frecuentistas piensan sobre las secuencias de acontecimientos, y no hay nada matemáticamente incorrecto en la forma en que los bayesianos definen las creencias de un agente racional. De hecho, cuando profundizas en los detalles, bayesianos y frecuentistas coinciden en muchas cosas. Muchos métodos frecuentistas conducen a decisiones que los bayesianos están de acuerdo en que tomaría un agente racional. Muchos métodos bayesianos tienen muy buenas propiedades frecuentistas.

En general, soy pragmática, así que usaré cualquier método estadístico en el que confíe. Resulta que prefiero los métodos bayesianos por razones que explicaré al final del libro. Pero no me opongo fundamentalmente a los métodos frecuentistas. No todo el mundo está tan relajado. Por ejemplo, consideremos a Sir Ronald Fisher, una de las figuras más destacadas de la estadística del siglo XX y un opositor vehemente a todo lo bayesiano, cuyo artículo sobre los fundamentos matemáticos de la estadística se refería a la probabilidad bayesiana como "una jungla impenetrable [que] detiene el progreso hacia la precisión de los conceptos estadísticos" (Fisher, 1922b, p. 311). O el psicólogo Paul Meehl, quien sugiere que confiar en los métodos frecuentistas podría convertirte en "un libertino intelectual potente pero estéril que deja en su alegre camino una larga cola de doncellas violadas pero ninguna descendencia científica viable" (Meehl, 1967, p. 114). La historia de la estadística, como se puede deducir, no está exenta de entretenimiento.

En cualquier caso, aunque personalmente prefiero la visión bayesiana, la mayoría de los análisis estadísticos se basan en el enfoque frecuentista. Mi razonamiento es pragmático. El objetivo de este libro es cubrir aproximadamente el mismo territorio que una clase típica de estadística de grado en psicología, y si quieres entender las herramientas estadísticas utilizadas por la mayoría de los psicólogos y psicólogas, necesitarás una buena comprensión de los métodos frecuentistas. Te prometo que no es un esfuerzo en vano. Incluso si al final quieres pasarte a la perspectiva bayesiana, deberías leer al menos un libro sobre la visión frecuentista "ortodoxa". Además, no no voy a ignorar por completo la perspectiva bayesiana. De vez en cuando añadiré algún comentario desde un punto de vista bayesiano, y volveré a tratar el tema con más profundidad en Chapter 16.

7.3 Teoría básica de la probabilidad

A pesar de las discusiones ideológicas entre bayesianos y frecuentistas, resulta que la mayoría de la gente está de acuerdo en las reglas que deben seguir las probabilidades. Hay muchas maneras diferentes de llegar a estas reglas. El método más utilizado se basa en el trabajo de Andrey Kolmogorov, uno de los grandes matemáticos soviéticos del siglo XX. No entraré en muchos detalles, pero intentaré darte una idea de cómo funciona. Y para ello voy a tener que hablar de mis pantalones.

7.3.1 Introducción a las distribuciones de probabilidad

Una de las verdades más inquietantes de mi vida es que solo tengo 5 pares de pantalones. Tres vaqueros, la mitad inferior de un traje y un pantalón de chándal. Y lo que es más triste, les he puesto nombres: los llamo X_1, X_2, X_3, X_4 y X_5. De verdad, por eso me llaman Mister Imaginative. Ahora, en un día cualquiera, elijo exactamente un pantalón para ponerme. Ni siquiera yo soy tan estúpido como para intentar llevar dos pares de pantalones, y gracias a años de entrenamiento ya nunca salgo a la calle sin usar pantalones. Si tuviera que describir esta situación usando el lenguaje de la teoría de la probabilidad, me referiría a cada par de pantalones (es decir, cada X) como un suceso elemental. La característica clave de los **sucesos elementales** es que cada vez que hacemos una observación (p. ej., cada vez que me pongo unos pantalones), el resultado será uno y solo uno de estos sucesos. Como he dicho, estos días siempre llevo exactamente un pantalón, así que mis pantalones satisfacen esta restricción. Del mismo modo, el conjunto de todos los sucesos posibles se denomina **espacio muestral**. Es cierto que algunas personas lo llamarían "armario", pero eso es porque se niegan a pensar en mis pantalones en términos probabilísticos. Qué triste.

Bien, ahora que tenemos un espacio muestral (un armario), que se construye a partir de muchos sucesos elementales posibles (pantalones), lo que queremos hacer es asignar una **probabilidad** a uno de estos sucesos elementales. Para un suceso X, la probabilidad de ese suceso $P(X)$ es un número comprendido entre 0 y 1. Cuanto mayor sea el valor de $P(X)$, más probable es que ocurra el suceso. Así, por ejemplo, si $P(X) = 0$ significa que el suceso X es imposible (es decir, nunca me pongo esos pantalones). Por otro lado, si $P(X) = 1$ significa que el suceso X seguramente ocurrirá (es decir, siempre llevo esos pantalones). Para valores de probabilidad intermedios significa que a veces llevo esos pantalones. Por ejemplo, si $P(X) = 0.5$ significa que llevo esos pantalones la mitad de las veces.

Llegados a este punto, casi hemos terminado. Lo último que debemos reconocer es que "siempre pasa algo". Cada vez que me pongo unos pantalones, realmente acabo llevando pantalones (loco, ¿no?). Lo que significa esta afirmación un tanto trillada, en términos probabilísticos, es que las probabilidades de los sucesos elementales tienen que sumar 1. Esto se conoce como la **ley de probabilidad total**, aunque a ninguna de nosotras nos importe realmente. Y lo que es más importante, si se cumplen estos requisitos, lo que tenemos es una **distribución de probabilidad**. Por ejemplo, la Table 7.2 muestra un ejemplo de una distribución de probabilidad.

Cada uno de los sucesos tiene una probabilidad comprendida entre 0 y 1, y si sumamos las probabilidades de todos los sucesos, suman 1. Impresionante. Incluso podemos dibujar un bonito gráfico de barras (ver Section 5.3) para visualizar esta distribución, como se muestra en la Figure 7.2. Y, llegados a este punto, todos hemos conseguido algo.

Table 7.2: una distribución de probabilidad para el uso de pantalones

Which trousers?	**Label**	**Probability**
Blue jeans	X_1	$P(X_1) = .5$
Grey jeans	X_2	$P(X_2) = .3$
Black jeans	X_3	$P(X_3) = .1$
Black suit	X_4	$P(X_4) = 0$
Blue tracksuit	X_5	$P(X_5) = .1$

Aprendiste lo que es una distribución de probabilidad y yo, por fin, encontré una manera de crear un gráfico que se centre por completo en mis pantalones. ¡Todo el mundo gana! Lo único que tengo que decir es que la teoría de la probabilidad permite hablar tanto de **sucesos no elementales** como de los elementales. La forma más fácil de ilustrar el concepto es con un ejemplo. En el ejemplo de los pantalones, es perfectamente legítimo referirse a la probabilidad de que yo lleve vaqueros. En este escenario, el suceso "Dani lleva vaqueros" se dice que ha ocurrido siempre que el suceso elemental que realmente ocurrió sea uno de los apropiados. En este caso "vaqueros azules", "vaqueros negros" o "vaqueros grises". En términos matemáticos definimos el suceso "vaqueros" E como el conjunto de sucesos elementales $(X1, X2, X3)$. Si se produce alguno de estos sucesos elementales, también se dice que se ha producido E. Habiendo decidido escribir la definición del E de esta manera, es bastante sencillo establecer cuál es la probabilidad P(E) y, puesto que las probabilidades de los vaqueros azules, grises y negros respectivamente son .5, .3 y $ .1$, la probabilidad de que lleve vaqueros es igual a .9. es: simplemente lo sumamos todo. En este caso concreto,

$$P(E) = P(X_1) + P(X_2) + P(X_3)$$

Llegados a este punto, puede que estés pensando que todo esto es terriblemente obvio y sencillo y estarías en lo cierto. En realidad, lo único que hemos hecho es envolver unas cuantas intuiciones de sentido común con algunas matemáticas básicas. Sin embargo, a partir de estos sencillos principios es posible construir algunas herramientas matemáticas extremadamente potentes. No voy a entrar en detalles en este libro, pero lo que sí voy a hacer es enumerar, en la Table 7.3, algunas de las otras reglas que satisfacen las probabilidades. Estas reglas se pueden derivar de los supuestos básicos que he descrito anteriormente, pero como en realidad no usamos estas reglas para nada en este libro, no lo haré aquí.

Table 7.3: algunas reglas que cumplen las probabilidades

English	**Notation**	**Formula**
not A	$P(\neg A)$	$1 - P(A)$
A or B	$P(A \cup B)$	$P(A)+P(B)-P(A \cap B)$
A and B	$P(A \cap B)$	$P(A\|B)P(B)$

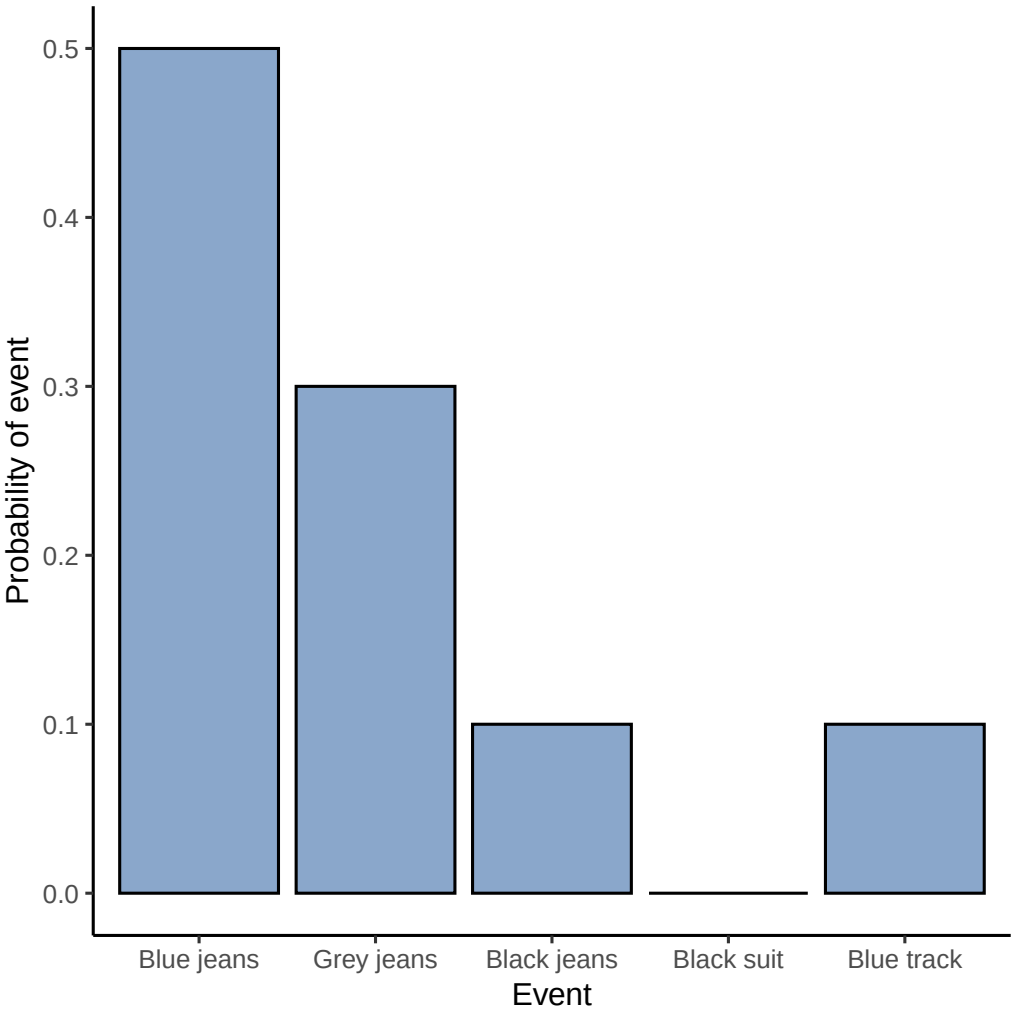


Figure 7.2: una representación visual de la distribución de probabilidad de los 'pantalones'. Hay cinco 'sucesos elementales', correspondientes a los cinco pantalones que tengo. Cada suceso tiene alguna probabilidad de ocurrir - esta probabilidad es un número entre 0 y 1. La suma de estas probabilidades es 1

7.4 La distribución binomial

Como puedes imaginar, las distribuciones de probabilidad varían enormemente y existe una gran variedad de distribuciones. Sin embargo, no todas tienen la misma importancia. De hecho, la mayor parte del contenido de este libro se basa en una de cinco distribuciones: la distribución binomial, la distribución normal, la distribución t, la distribución χ^2 ("ji-cuadrado") y la distribución F . Por ello, lo que haré en las próximas secciones será una breve introducción a estas cinco distribuciones, prestando especial atención a la binomial y la normal. Empezaré por la distribución binomial ya que es la más sencilla de las cinco.

7.4.1 Introducción a la distribución binomial

La teoría de la probabilidad se originó en el intento de describir cómo funcionan los juegos de azar, por lo que parece apropiado que nuestra discusión sobre la **distribución binomial** incluya una discusión sobre tirar dados y lanzar monedas. Imaginemos un sencillo "experimento". En mi mano tengo 20 dados idénticos de seis caras. En una cara de cada dado hay una imagen de una calavera, las otras cinco caras están en blanco. Si tiro los 20 dados, ¿cuál es la probabilidad de que obtenga exactamente 4 calaveras? Suponiendo que los dados sean justos, sabemos que la probabilidad de que salga una calavera es de 1 entre 6. Dicho de otro modo, la probabilidad de que salga una calavera con un solo dado es de aproximadamente 0,167. Esta información es suficiente para responder a nuestra pregunta, así que veamos cómo se hace.

Como de costumbre, tendremos que introducir algunos nombres y alguna notación. Dejaremos que N denote el número de lanzamientos de dados en nuestro experimento, cantidad que suele denominarse **parámetro de tamaño** de nuestra distribución binomial. También usaremos θ para referirnos a la probabilidad de que un solo dado salga calavera, cantidad que generalmente se denomina **probabilidad de éxito** de la binomial.[2] Finalmente, usaremos X para referirnos a los resultados de nuestro experimento, es decir, el número de calaveras que obtengo al tirar los dados. Dado que el valor real de X se debe al azar, nos referimos a él como **variable aleatoria**. En cualquier caso, ahora que tenemos toda esta terminología y notación podemos usarla para plantear el problema con un poco más de precisión. La cantidad que queremos calcular es la probabilidad de que $X = 4$ dado que sabemos que $\theta = .167$ y $N = 20$. La "forma" general de lo que me interesa calcular podría escribirse como

$$P(X|\theta, N)$$

y nos interesa el caso especial donde $X = 4, \theta = .167$ y $N = 20$.

[Detalle técnico adicional [3]]

[2] ten en cuenta que el término "éxito" es bastante arbitrario y no implica realmente que el resultado sea algo deseable. Si θ se refiriera a la probabilidad de que un pasajero resulte herido en un accidente de autobús, seguiría llamándola probabilidad de éxito, pero eso no significa que quiera que la gente resulte herida en accidentes de autobús.

[3] Para los lectores que sepan un poco de cálculo, daré una explicación un poco más precisa. Del mismo modo que las probabilidades son números no negativos que deben sumar 1, las densidades de probabilidad son números no negativos que deben integrarse en 1 (donde la integral se toma a lo largo de todos los valores posibles de X). Para calcular la probabilidad de que X se encuentre entre a y b calculamos la integral definida de la función de densidad sobre el intervalo correspondiente, $\int_a^b p(x)dx$. Si no recuerdas o nunca has aprendido cálculo, no te preocupes. No es necesario para este libro.

Sí, sí. Sé lo que estás pensando: notación, notación, notación. Realmente, ¿a quién le importa? Muy pocos lectores de este libro están aquí por la notación, así que probablemente debería continuar y hablar de cómo utilizar la distribución binomial. He incluido la fórmula de la distribución binomial en una nota a pie de página [4], ya que algunos lectores pueden querer jugar con ella por sí mismos, pero como a la mayoría de la gente probablemente no le importe mucho y porque no necesitamos la fórmula en este libro, no hablaré de ella en detalle. En su lugar, solo quiero mostrarte cómo es la distribución binomial.

Para ello, la Figure 7.3 muestra las probabilidades binomiales para todos los valores posibles de X para nuestro experimento de lanzamiento de dados, desde $X = 0$ (sin calaveras) hasta $X = 20$ (todas las calaveras). Ten en cuenta que esto es básicamente un gráfico de barras, y no difiere en nada del gráfico de "probabilidad de los pantalones" que dibujé en la Figure 7.2. En el eje horizontal tenemos todos los sucesos posibles y en el eje vertical podemos leer la probabilidad de cada uno de esos sucesos. Así, la probabilidad de sacar calaveras de 4 de 20 es de aproximadamente $0,20$ (la respuesta real es $0,2022036$, como veremos en un momento). En otras palabras, esperarías que ocurriera alrededor del 20% de las veces que repitieras este experimento.

Para que nos hagamos una idea de cómo cambia la distribución binomial cuando modificamos los valores de *theta* y N, supongamos que, en lugar de tirar los dados, lo que hago es lanzar monedas. Esta vez, mi experimento consiste en lanzar una moneda al aire repetidamente y el resultado que me interesa es el número de caras que observo. En este escenario, la probabilidad de éxito ahora es $\theta = \frac{1}{2}$. Supongamos que lanzo la moneda $N = 20$ veces. En este ejemplo, he cambiado la probabilidad de éxito pero he mantenido el mismo tamaño del experimento. ¿Cómo afecta esto a nuestra distribución binomial? Bueno, como muestra la Figure 7.4, el efecto principal de esto es desplazar toda la distribución, como era de esperar. Bien, ¿y si lanzamos una moneda $N = 100$ veces? Bueno, en ese caso obtenemos la Figure 7.4 (b). La distribución se mantiene aproximadamente en el centro, pero hay un poco más de variabilidad en los posibles resultados.

7.5 La distribución normal

Aunque la distribución binomial es conceptualmente la más sencilla de entender, no es la más importante. Ese honor particular corresponde a la distribución normal, también conocida como "la curva de campana" o "distribución gaussiana". Una **distribución normal** se describe utilizando dos parámetros: la media de la distribución μ y la desviación estándar de la distribución σ.

[Detalle técnico adicional [5]]

Intentemos hacernos una idea de lo que significa que una variable se distribuya normalmente. Para ello, observa la Figure 7.5 que representa una distribución normal con

[4] En la ecuación de la binomial, $X!$ es la función factorial (es decir, multiplicar todos los números enteros de 1 a X):

$$P(X|\theta, N) = \frac{N!}{X!(N-X)!}\theta^X(1-\theta)^{N-X}$$

Si esta ecuación no tiene mucho sentido para ti, no te preocupes.

[5] Al igual que en el caso de la distribución binomial, he incluido la fórmula de la distribución normal en este libro, porque creo que es lo suficientemente importante como para que todo el que aprenda estadística deba al menos echarle un vistazo, pero como éste es un texto introductorio no quiereo

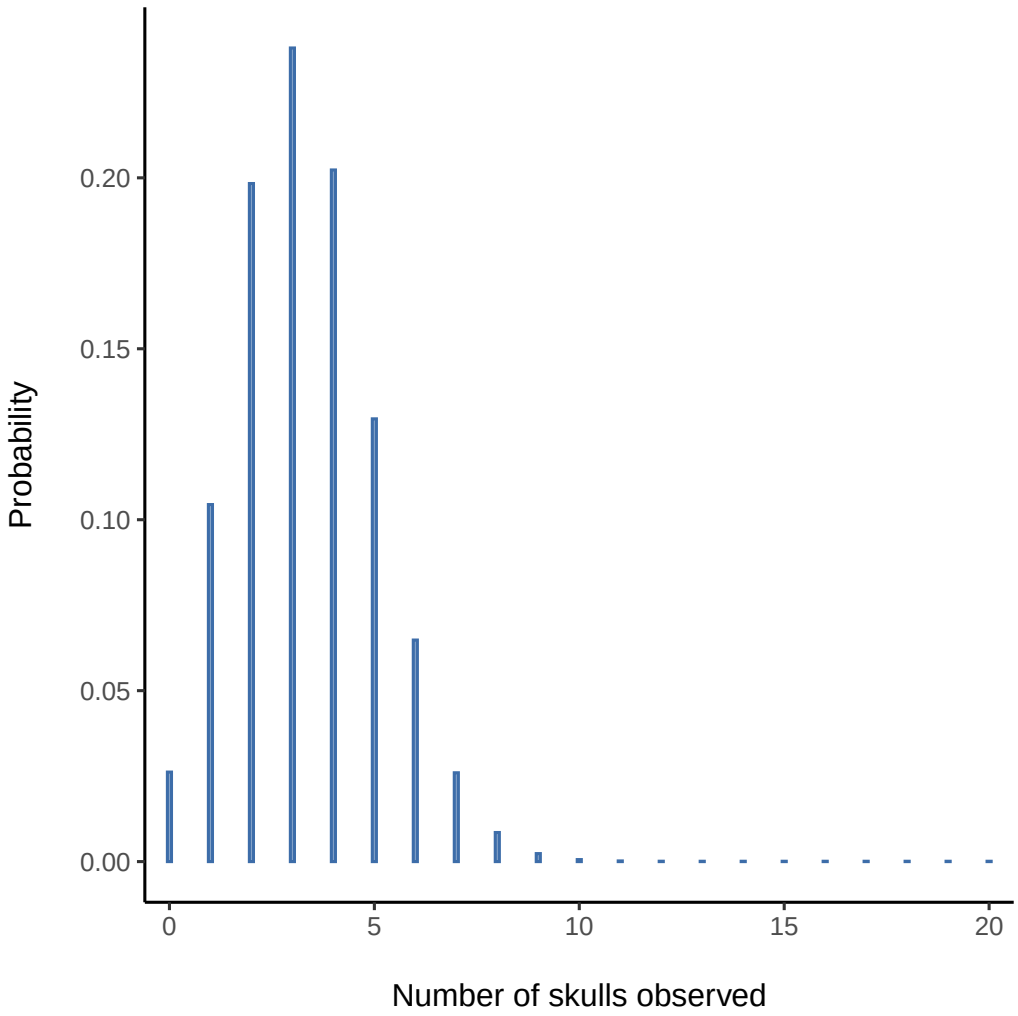


Figure 7.3: La distribución binomial con parámetro de tamaño de $N = 20$ y una probabilidad de éxito subyacente de $\theta = \frac{1}{6}$. Cada barra vertical representa la probabilidad de un resultado específico (es decir, un valor posible de X). Como se trata de una distribución de probabilidad, cada una de las probabilidades debe ser un número entre 0 y 1, y las alturas de las barras también deben sumar 1.

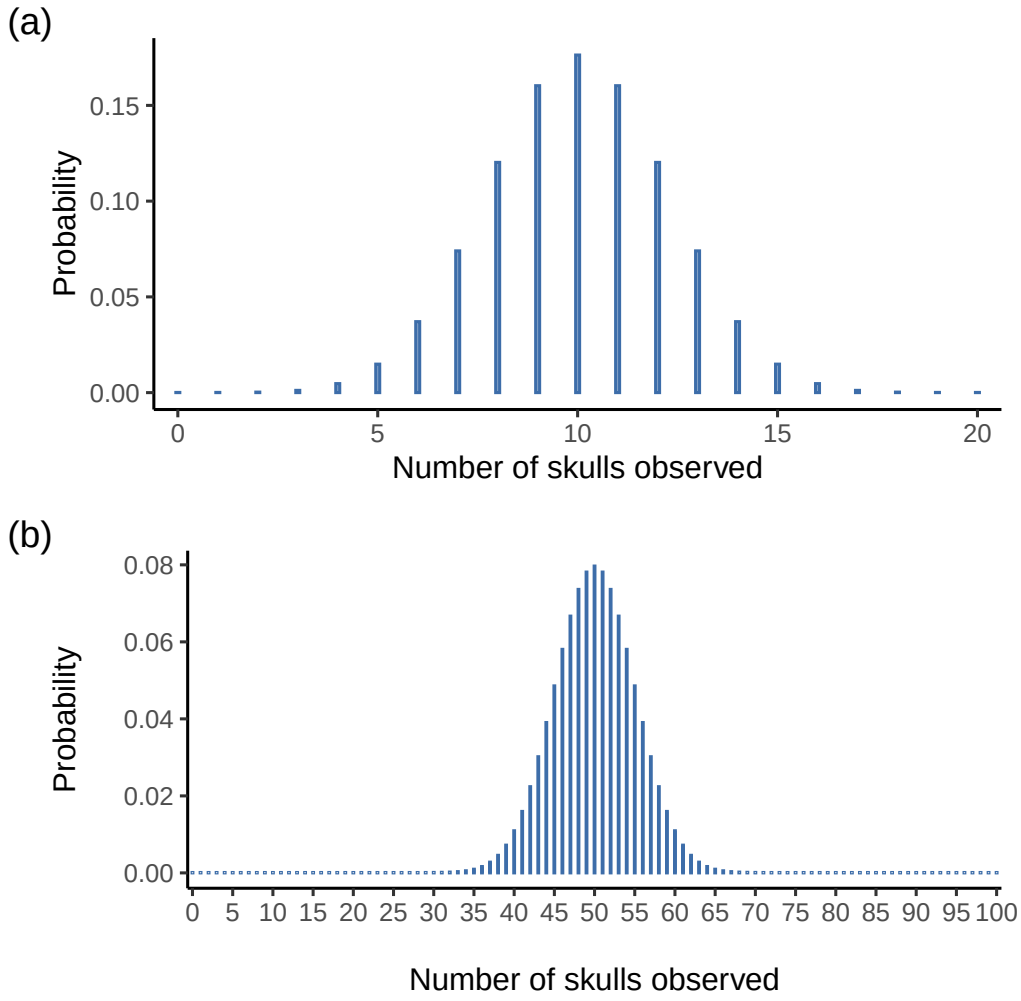


Figure 7.4: dos distribuciones binomiales, en un escenario en el que lanzo una moneda justa, por lo que la probabilidad de éxito subyacente es $\theta = \frac{1}{2}$. En el panel (a), lancé la moneda $N = 20$ veces. En el panel (b) la moneda se lanzó $N = 100$ veces

media $\mu = 0$ y desviación estándar $\sigma = 1$. Puedes ver de dónde viene el nombre "curva de campana"; se parece un poco a una campana. Observa que, a diferencia de los gráficos que he dibujado para ilustrar la distribución binomial, la imagen de la distribución normal de la Figure 7.5 muestra una curva suave en lugar de barras "tipo histograma". No se trata de una elección arbitraria, ya que la distribución normal es continua, mientras que la binomial es discreta. Por ejemplo, en el ejemplo del dado de la sección anterior era posible obtener 3 calaveras o 4 calaveras, pero era imposible obtener 3,9 calaveras. Las figuras que dibujé en el apartado anterior reflejan este hecho. En la Figure 7.3, por ejemplo, hay una barra situada en $X = 3$ y otra en $X = 4$ pero no hay nada en medio. Las cantidades continuas no tienen esta restricción. Por ejemplo, supongamos que hablamos del tiempo. La temperatura en un agradable día de primavera podría ser de 23 grados, 24 grados, 23,9 grados o cualquier valor intermedio, ya que la temperatura es una variable continua. Por tanto, una distribución normal podría ser muy adecuada para describir las temperaturas primaverales[6]

Teniendo esto en cuenta, veamos si podemos intuir cómo funciona la distribución normal. En primer lugar, veamos qué ocurre cuando jugamos con los parámetros de la distribución. Para ello, en la Figure 7.6 se representan distribuciones normales con medias diferentes pero con la misma desviación estándar. Como era de esperar, todas estas distribuciones tienen la misma "anchura". La única diferencia entre ellas es que se han desplazado a la izquierda o a la derecha. Por lo demás, son idénticas. Por el contrario, si aumentamos la desviación estándar manteniendo la media constante, el pico de la distribución se mantiene en el mismo lugar, pero la distribución se ensancha, como se puede ver en la Figure 7.7. Sin embargo, observa que cuando ampliamos la distribución, la altura del pico disminuye. Esto tiene que suceder, del mismo modo que las alturas de las barras que usamos para dibujar una distribución binomial discreta deben sumar 1, el área total bajo la curva de la distribución normal debe ser igual a 1. Antes de continuar, quiero señalar una característica importante de la distribución normal. Independientemente de cuál sea la media real y la desviación estándar, $68,3\%$ del área cae dentro de 1 desviación estándar de la media. Del mismo modo, $95,4\%$ de la distribución cae dentro de 2 desviaciones estándar de la media, y $(99,7\%)$ de la distribución está dentro de 3 desviaciones estándar. Esta idea se ilustra en la Figure 7.8; ver también la Figure 7.9.

7.5.1 Densidad de probabilidad

Hay algo que he intentado ocultar a lo largo de mi discusión sobre la distribución normal, algo que algunos libros de texto introductorios omiten por completo. Puede que tengan razón al hacerlo. Esta "cosa" que estoy ocultando es extraña y contraintuitiva, incluso para los estándares distorsionados que se aplican en estadística. Afortunadamente, no es algo que haya que entender a un nivel profundo para hacer estadística básica. Más bien, es algo que empieza a ser importante más adelante, cuando se va más allá de lo

centrarme en ella, así que la he escondido en esta nota a pie de página:

$$p(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

[6] en la práctica, la distribución normal es tan práctica que la gente tiende a utilizarla incluso cuando la variable no es realmente continua. Siempre que haya suficientes categorías (p. ej., las respuestas de una escala Likert a un cuestionario), es una práctica bastante habitual usar la distribución normal como aproximación. Esto funciona mucho mejor de lo que parece.

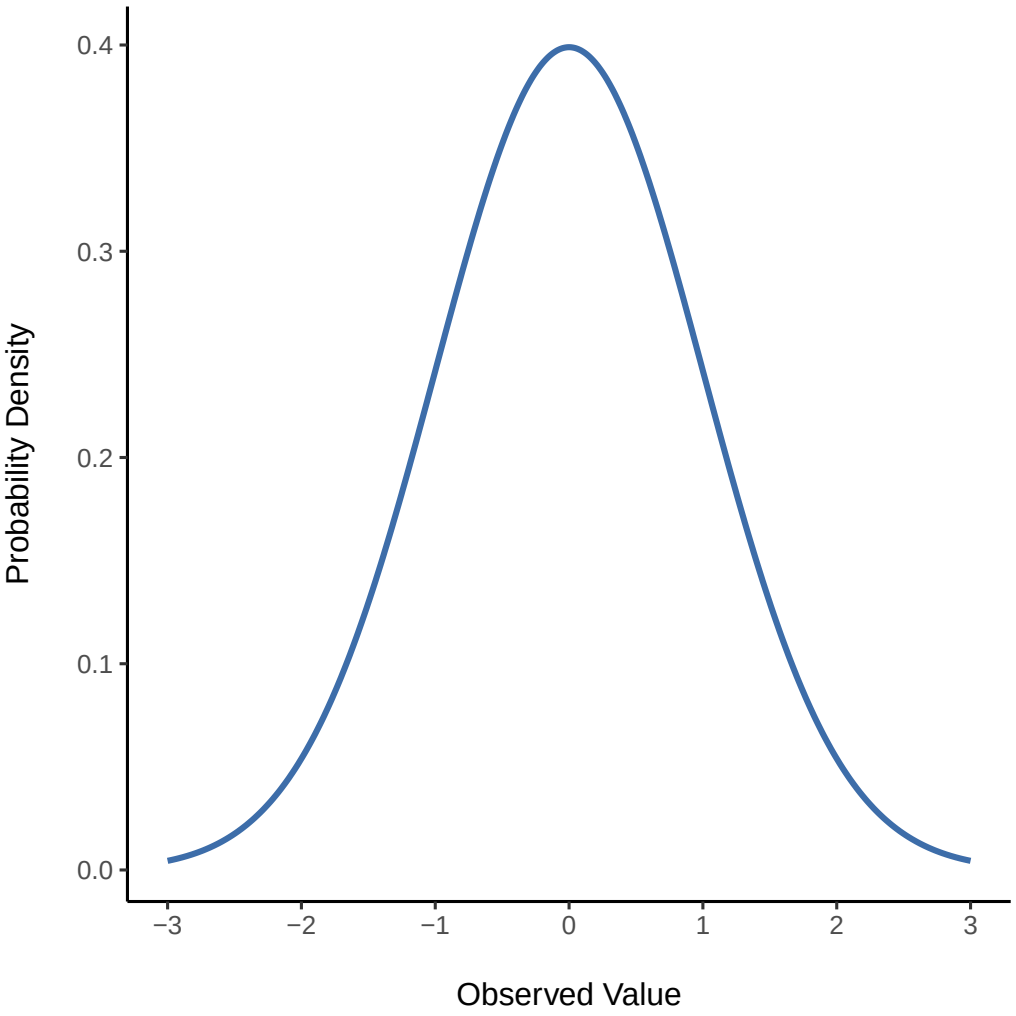


Figure 7.5: La distribución normal con media $\mu = 0$ y desviación estándar $\sigma = 1$. El eje x corresponde al valor de alguna variable, y el eje y nos dice algo sobre la probabilidad de que observemos ese valor. Sin embargo, fíjate que el eje y está etiquetado como *Densidad de probabilidad* y no como *Probabilidad*. Hay una característica sutil y algo frustrante de las distribuciones continuas que hace que el eje y se comporte de forma un poco extraña:- la altura de la curva aquí no es realmente la probabilidad de observar un valor de x en particular. Por otro lado, es cierto que la altura de la curva indica qué valores de x son más probables (¡los más altos!). (ver la sección Densidad de probabilidad para más detalles)

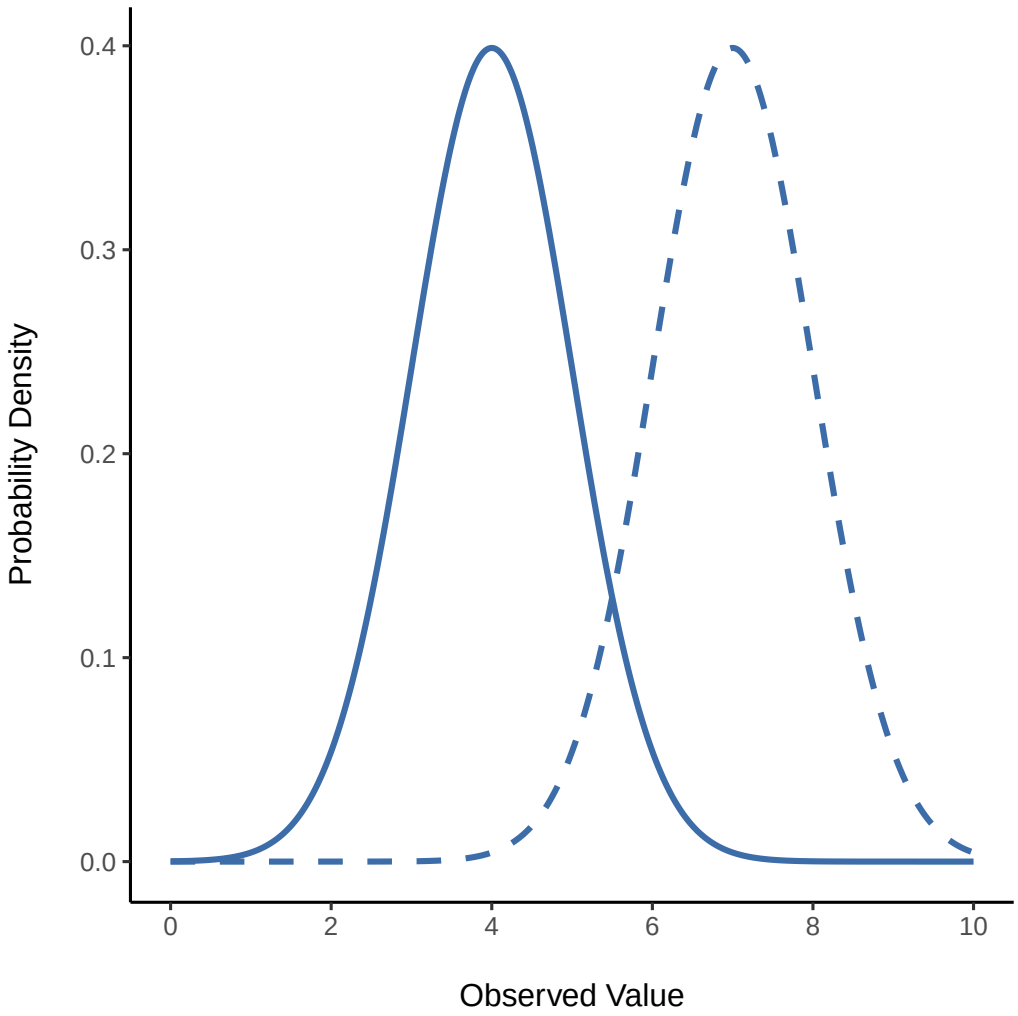


Figure 7.6: ilustración de lo que ocurre cuando se modifica la media de una distribución normal. La línea continua representa una distribución normal con una media de $\mu = 4$. La línea discontinua muestra una distribución normal con una media de $\mu = 7$. En ambos casos, la desviación estándar es $\sigma = 1$. Como es lógico, las dos distribuciones tienen la misma forma, pero la línea discontinua está desplazada hacia la derecha.

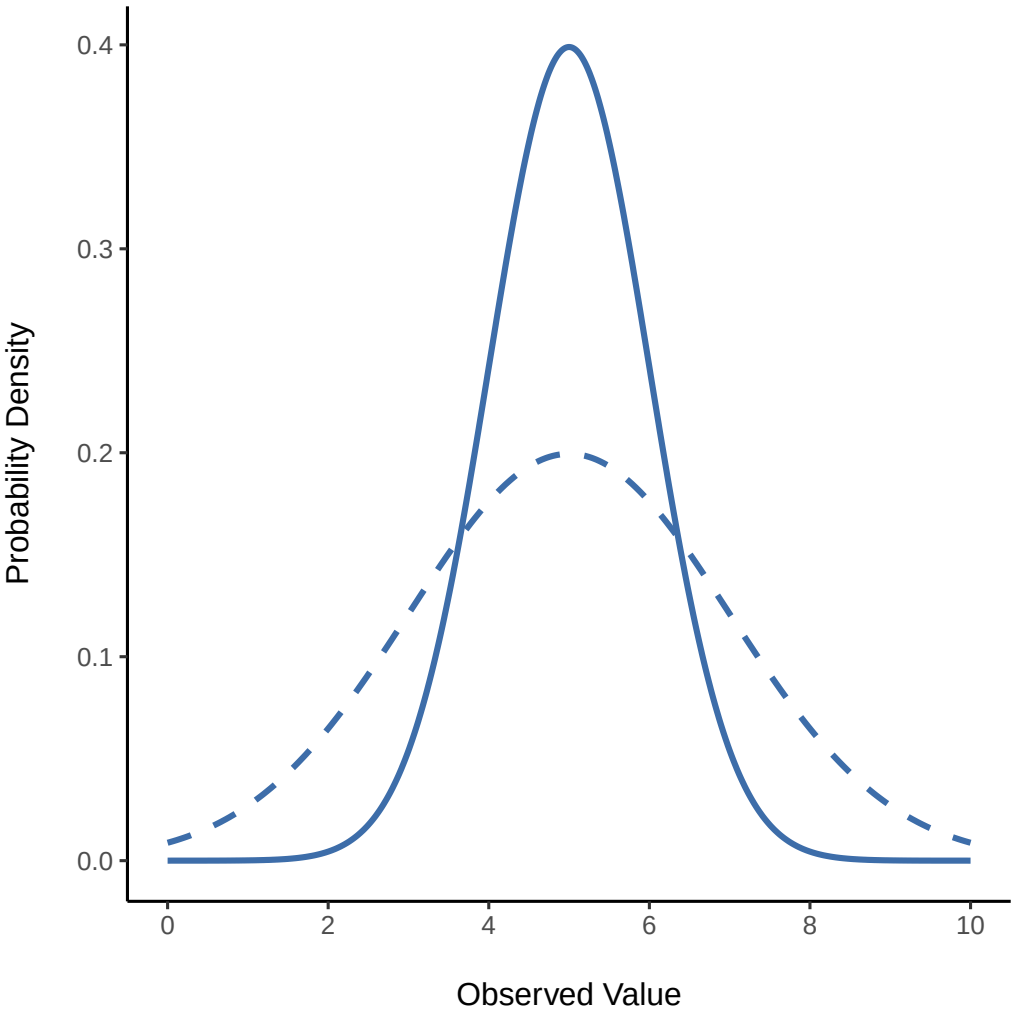


Figure 7.7: ilustración de lo que ocurre cuando se cambia la desviación estándar de una distribución normal. Ambas distribuciones representadas en esta figura tienen una media de $\mu = 5$, pero tienen diferentes desviaciones estándar. La línea continua representa una distribución con desviación estándar $\sigma = 1$ y la línea discontinua muestra una distribución con desviación estándar $\sigma = 2$. Por consiguiente, ambas distribuciones están 'centradas' en el mismo punto, pero la línea discontinua es más ancha que la continua.

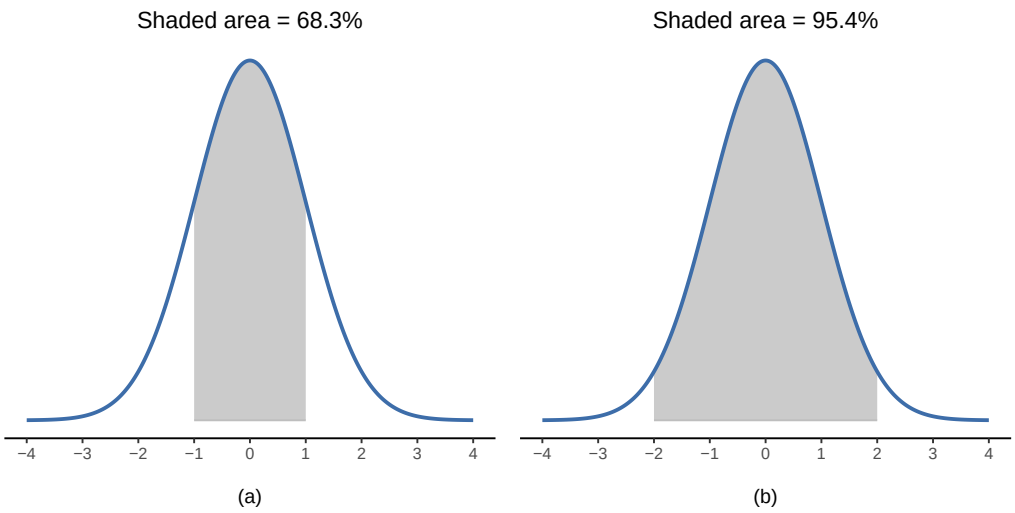


Figure 7.8: El área bajo la curva indica la probabilidad de que una observación se encuentre dentro de un intervalo determinado. Las líneas continuas representan distribuciones normales con media $\mu = 0$ y desviación estándar $\sigma = 1$. Las áreas sombreadas ilustran las 'áreas bajo la curva' de dos casos importantes. En el panel (a), podemos ver que hay un 68,3% de probabilidad de que una observación caiga dentro de una desviación estándar de la media. En el panel (b), vemos que hay un 95,4 % de probabilidad de que una observación se sitúe a dos desviaciones estándar de la media

básico. Así que, si no tiene mucho sentido, no te preocupes demasiado, pero intenta asegurarte de que entiendes lo esencial.

A lo largo de mi exposición sobre la distribución normal ha habido una o dos cosas que no acaban de tener sentido. Quizás te hayas dado cuenta de que el eje y de estas figuras está etiquetado como "Densidad de probabilidad" en lugar de densidad. Tal vez te hayas dado cuenta de que he utilizado $P(X)$ en lugar de $P(X)$ al dar la fórmula de la distribución normal.

Resulta que lo que se presenta aquí no es en realidad una probabilidad, es otra cosa. Para entender qué es ese algo, hay que dedicar un poco de tiempo a pensar qué significa realmente decir que X es una variable continua. Digamos que estamos hablando de la temperatura exterior. El termómetro me dice que hace 23 grados, pero sé que eso no es realmente cierto. No son exactamente 23 grados. Tal vez sean 23.1 grados, pienso. Pero sé que eso tampoco es realmente cierto porque puede que sean $ 23.09 $ grados. Pero sé que… bueno, entiendes la idea. Lo complicado de las cantidades realmente continuas es que nunca se sabe exactamente lo que son.

Ahora piensa en lo que esto implica cuando hablamos de probabilidades. Supongamos que la temperatura máxima de mañana se muestra a partir de una distribución normal con una media 23 y desviación estándar 1. ¿Cuál es la probabilidad de que la temperatura sea exactamente de 23 grados? La respuesta es "cero", o posiblemente "un número tan cercano a cero que bien podría ser cero". ¿A qué se debe esto? Es como intentar lanzar un dardo a una diana infinitamente pequeña. Por muy buena puntería que tengas, nunca acertarás. En la vida real, nunca obtendrás un valor de exactamente

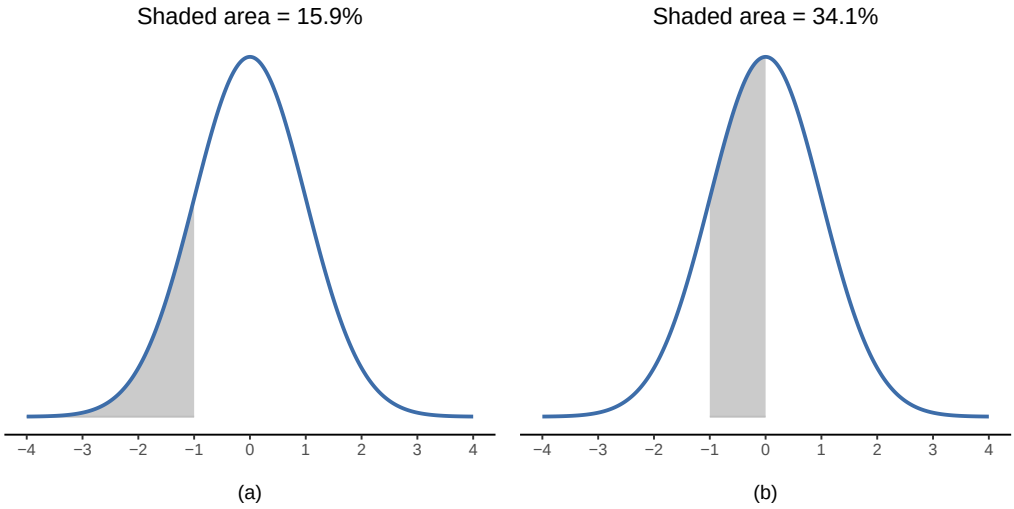


Figure 7.9: Dos ejemplos más de la 'idea del área bajo la curva'. Hay una probabilidad del 15,9 % de que una observación esté una desviación estándar por debajo de la media o menos (panel (a)), y una probabilidad del 34,1 % de que la observación esté entre una desviación estándar por debajo de la media y la media (panel (b)). Fijate que si sumas estas dos cifras, obtienes 15,9 % + 34,1 % = 50 %. Para datos distribuidos normalmente, existe un 50% de probabilidad de que una observación se sitúe por debajo de la media. Y, por supuesto, eso también implica que hay un 50% de probabilidades de que se sitúe por encima de la media.

$ 23 $. Siempre será algo como $ 23,1 $ o $ 22,99998 $ o algo así. En otras palabras, no tiene ningún sentido hablar de la probabilidad de que la temperatura sea exactamente de 23 grados. Sin embargo, en el lenguaje cotidiano, si te dijera que afuera había $23 grados y resultara que hace $22,9998, probablemente no me llamarías mentirosa. Porque en el lenguaje cotidiano "23 grados" suele significar algo así como "algo entre $22,5$ y $23,5$ grados". Y aunque no parece muy significativo preguntar sobre la probabilidad de que la temperatura sea exactamente de 23 grados, sí parece sensato preguntar sobre la probabilidad de que la temperatura esté entre $22,5$ y $23,5$, o entre 20 y 30. , o cualquier otro rango de temperaturas.

El objetivo de esta discusión es dejar claro que, cuando hablamos de distribuciones continuas, no tiene sentido hablar de la probabilidad de un valor concreto. Sin embargo, de lo que sí podemos hablar es de la probabilidad de que el valor se encuentre dentro de un rango concreto de valores. Para averiguar la probabilidad asociada a un rango particular, lo que hay que hacer es calcular el "área bajo la curva". Ya hemos visto este concepto, en la Figure 7.8 las áreas sombreadas muestran probabilidades reales (p. ej., en la Figure 7.8 muestra la probabilidad de observar un valor que se encuentra dentro de 1 desviación estándar de la media).

Vale, eso explica parte de la historia. He explicado un poco acerca de cómo las distribuciones continuas de probabilidad deben ser interpretadas (es decir, el área bajo la curva es la clave). Pero, ¿qué significa realmente la fórmula para ppxq que he descrito antes? Obviamente, $P(x)$ no describe una probabilidad, pero ¿qué es? El nombre de esta cantidad $P(x)$ es **densidad de probabilidad** y, en términos de los gráficos que hemos estado dibujando, corresponde a la altura de la curva. Las densidades en sí mismas no son significativas, pero están "amañadas" para garantizar que el área bajo la curva siempre se pueda interpretar como probabilidades genuinas. Para ser sincera, eso es todo lo que necesitas saber por ahora.[7]

7.6 Otras distribuciones útiles

La distribución normal es la que más utiliza en estadística (por razones que veremos en breve), y la distribución binomial es muy útil para muchos propósitos. Pero el mundo de la estadística está lleno de distribuciones de probabilidad, algunas de las cuales veremos de pasada. En concreto, las tres que aparecerán en este libro son la distribución t, la distribución χ^2 y la distribución F. No daré fórmulas para ninguna de ellas, ni hablaré de ellas con demasiado detalle, pero mostraré algunas imágenes: Figure 7.10, Figure 7.11 y Figure 7.12.

- La distribución t es una distribución continua que se parece mucho a una distribución normal, consulta la Figure 7.10. Observa que las "colas" de la distribución t son "más pesadas" (es decir, se extienden más hacia afuera) que las colas de la distribución normal. Esa es la diferencia importante entre ambas. Esta distribución suele aparecer en situaciones en las que se cree que los datos en realidad siguen

[7] Para los lectores que sepan un poco de cálculo, daré una explicación un poco más precisa. De la misma manera que las probabilidades son números no negativos que deben sumar 1, las densidades de probabilidad son números no negativos que deben integrarse a 1 (donde la integral se toma a través de todos los valores posibles de X). Para calcular la probabilidad de que X se encuentre entre a y b calculamos la integral definida de la función de densidad sobre el intervalo correspondiente, $\int_a^b p(x)dx$. Si no recuerdas o nunca has aprendido cálculo, no te preocupes por esto. No es necesario para este libro.

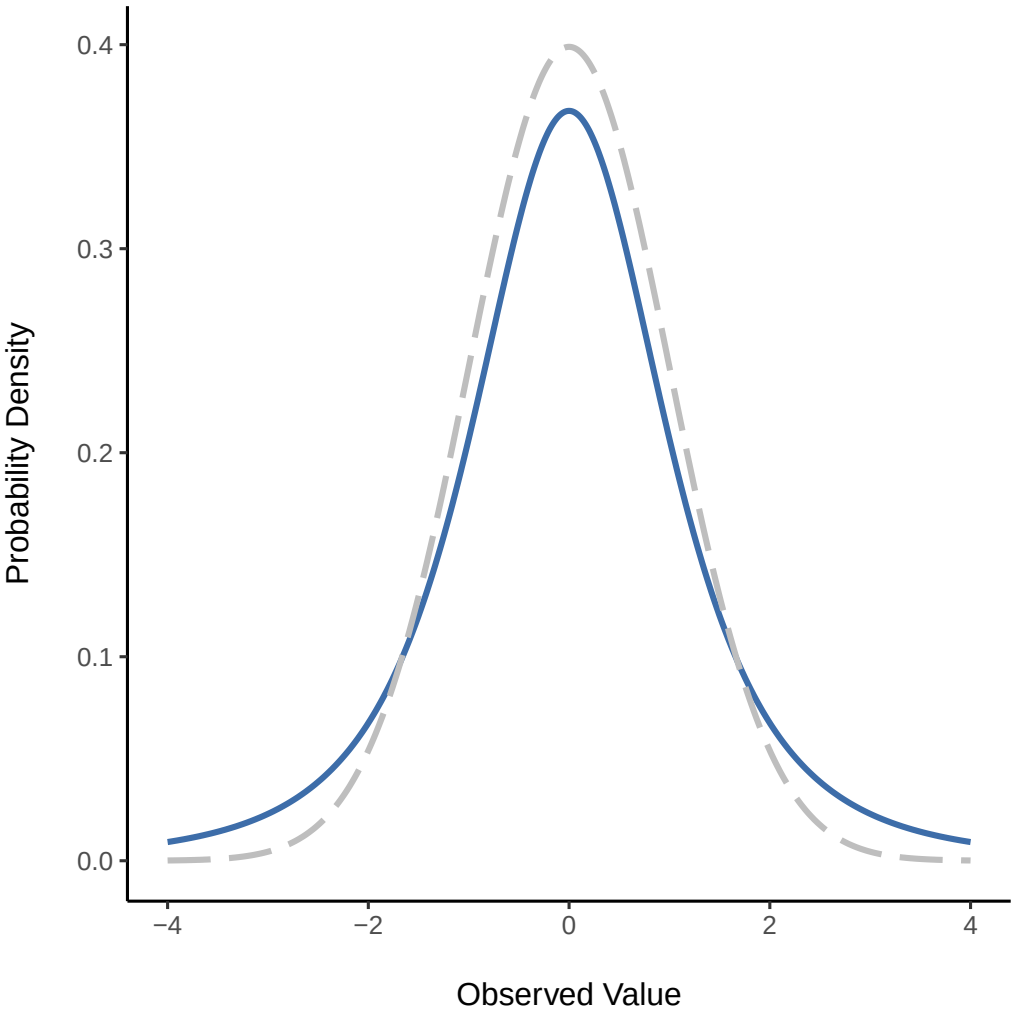


Figure 7.10: Una distribución t con 3 grados de libertad (línea continua). Se parece a una distribución normal, pero no es exactamente lo mismo. Para comparar, he trazado una distribución normal estándar como línea discontinua

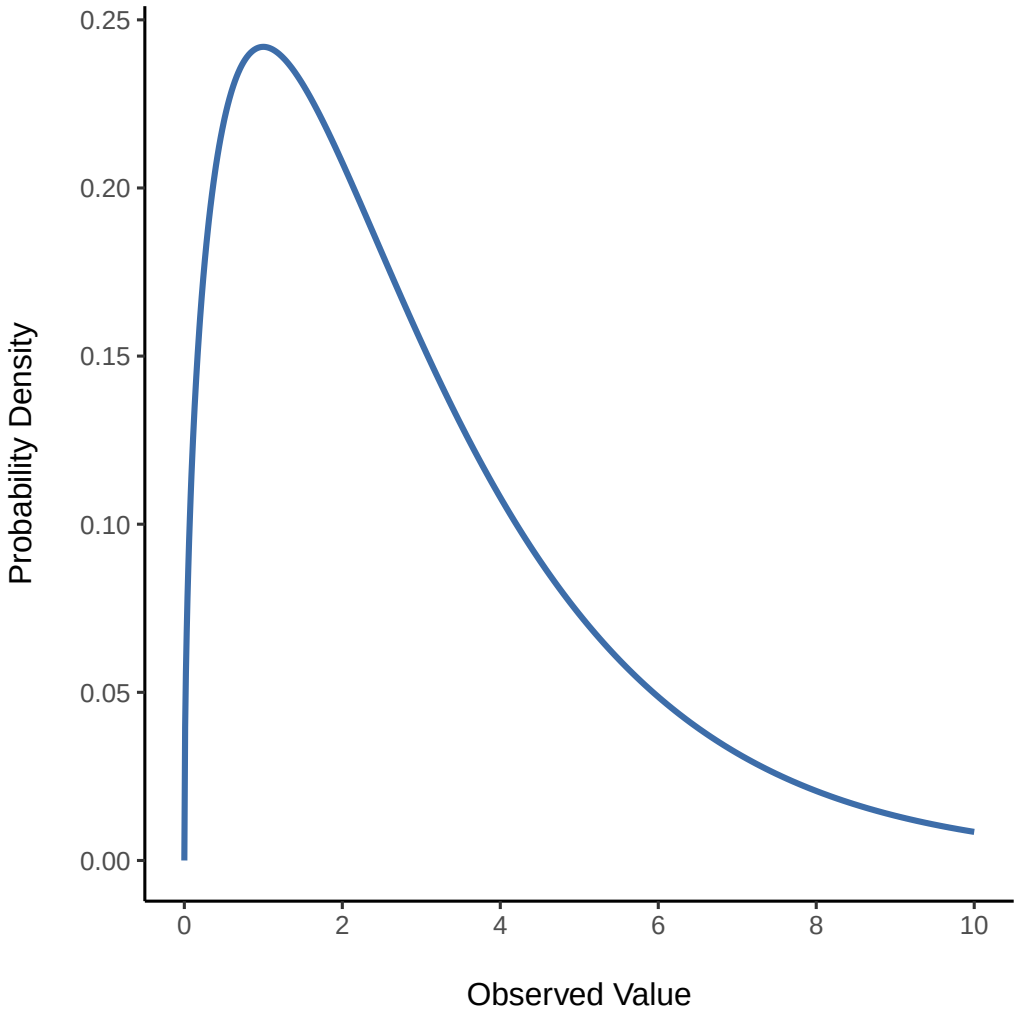


Figure 7.11: distribución χ^2 con 3 grados de libertad. Fíjate que los valores observados siempre deben ser mayores que cero y que la distribución está bastante sesgada. Estas son las características clave de una distribución ji-cuadrado

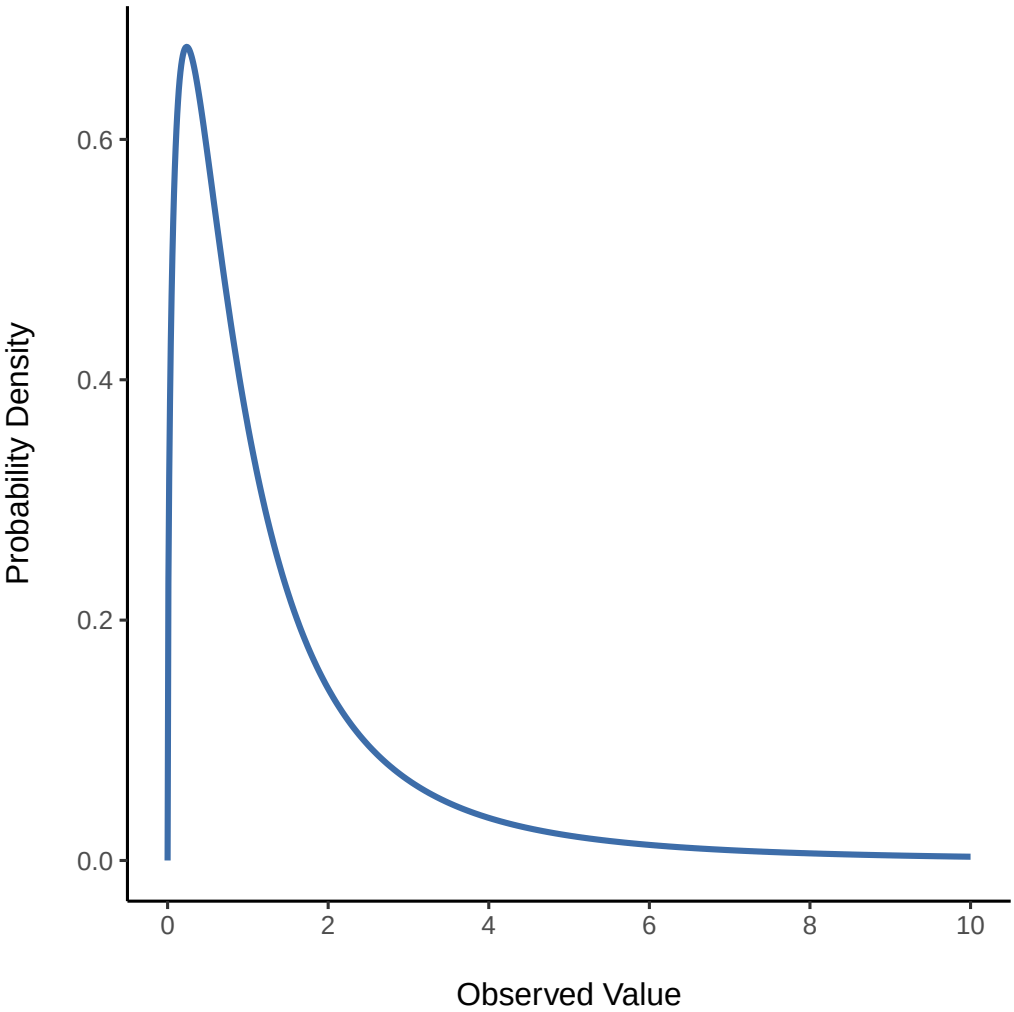


Figure 7.12: Una distribución F con 3 y 5 grados de libertad. Cualitativamente hablando, se parece bastante a una distribución ji-cuadrado, pero no son iguales en general.

una distribución normal, pero se desconoce la media o la desviación estándar. Nos encontraremos con esta distribución nuevamente en Chapter 11.

- La distribución χ^2 es otra distribución que aparece en muchos lugares diferentes. La situación en la que la veremos será cuando hagamos un análisis de datos categóricos en Chapter 10, pero es una de esas cosas que aparecen por todas partes. Cuando se profundiza en las matemáticas (¿y a quién no le gusta hacerlo?), resulta que la razón principal por la que la distribución χ^2 aparece por todas partes es que si tienes un montón de variables que se distribuyen normalmente, se elevan al cuadrado sus valores y luego se suman (un procedimiento conocido como "suma de cuadrados"), esta suma tiene una distribución χ^2. Te sorprendería saber con qué frecuencia este hecho resulta útil. De todos modos, así es como se ve una distribución χ^2: Figure 7.11.
- La distribución F se parece un poco a una distribución χ^2, y surge siempre que se necesita comparar dos distribuciones χ^2 entre sí. Hay que reconocer que esto no suena exactamente como algo que cualquier persona en su sano juicio querría hacer, pero resulta ser muy importante en el análisis de datos del mundo real. ¿Recuerdas cuando dije que χ^2 resulta ser la distribución clave cuando tomamos una "suma de cuadrados"? Bueno, lo que eso significa es que si quieres comparar dos "sumas de cuadrados" diferentes, probablemente estés hablando de algo que tiene una distribución F. Por supuesto, aún no te he dado un ejemplo de algo que involucre una suma de cuadrados, pero lo haré en Chapter 13. Y ahí es donde veremos la distribución F. Ah, y hay una imagen en la Figure 7.12.

Bien, es hora de terminar esta sección. Hemos visto tres distribuciones nuevas: χ^2), t y F. Todas son distribuciones continuas y están estrechamente relacionadas con la distribución normal. Lo principal para nuestros propósitos es que comprendas la idea básica de que estas distribuciones están profundamente relacionadas entre sí y con la distribución normal. Más adelante en este libro nos encontraremos con datos que se distribuyen normalmente, o que al menos se supone que se distribuyen normalmente. Lo que quiero que entiendas ahora es que, si asumes que tus datos se distribuyen normalmente, no deberías sorprenderte al ver las distribuciones χ^2, t y F apareciendo por todas partes cuando empieces a intentar hacer tu análisis de datos.

7.7 Resumen

En este capítulo hemos hablado de la probabilidad. Hemos hablado sobre lo que significa probabilidad y de por qué los estadísticos no se ponen de acuerdo sobre su significado. Hemos hablado de las reglas que deben cumplir las probabilidades. Hemos introducido la idea de distribución de probabilidad y hemos dedicado una buena parte del capítulo a hablar de algunas de las distribuciones de probabilidad más importantes con las que trabajan los estadísticos. El desglose por secciones es el siguiente:

- Teoría de la probabilidad versus estadística: ¿En qué se diferencian la probabilidad y la estadística?
- [La visión frecuentista] versus [La visión bayesiana] de la probabilidad
- Teoría básica de la probabilidad
- La distribución binomial, La distribución normal y Otras distribuciones útiles

Como era de esperar, mi cobertura no es en absoluto exhaustiva. La teoría de la probabilidad es una gran rama de las matemáticas por derecho propio, totalmente independiente

de su aplicación a la estadística y el análisis de datos. Como tal, hay miles de libros escritos sobre el tema y las universidades suelen ofrecer múltiples clases dedicadas por completo a la teoría de la probabilidad. Incluso la tarea "más sencilla" de documentar las distribuciones de probabilidad estándar es un gran tema. He descrito cinco distribuciones de probabilidad estándar en este capítulo, pero en mi estantería tengo un libro de 45 capítulos titulado "Distribuciones estadísticas" (M. Evans et al., 2011) que contiene muchas más. Afortunadamente para ti, muy poco de esto es necesario. Es poco probable que necesites conocer docenas de distribuciones estadísticas cuando salgas a hacer análisis de datos del mundo real, y definitivamente no las necesitarás para este libro, pero nunca está de más saber que hay otras posibilidades por ahí.

Retomando este último punto, hay un sentido en el que todo este capítulo es una especie de digresión. Muchas clases de psicología de grado sobre estadística pasan por alto este contenido muy rápidamente (sé que la mía lo hizo), e incluso las clases más avanzadas a menudo "olvidan" revisar los fundamentos básicos del campo. La mayoría de los psicólogos académicos no conocerían la diferencia entre probabilidad y densidad, y hasta hace poco muy pocos habrían sido conscientes de la diferencia entre probabilidad bayesiana y frecuentista. Sin embargo, creo que es importante comprender estas cosas antes de pasar a las aplicaciones. Por ejemplo, hay muchas reglas sobre lo que está "permitido" decir cuando se hace inferencia estadística y muchas de ellas pueden parecer arbitrarias y extrañas. Sin embargo, empiezan a tener sentido si se entiende que existe esta distinción entre bayesianos y frecuentistas. Del mismo modo, en Chapter 11 vamos a hablar de algo llamado la prueba t, y si realmente quieres comprender la mecánica de la prueba t, te ayudará tener una idea de cómo es realmente una distribución t. Espero que te hagas una idea.

Chapter 8

Estimación de cantidades desconocidas de una muestra

Al principio del último capítulo destaqué la distinción fundamental entre estadística descriptiva y *estadística inferencial*. Como se explica en Chapter 4, la función de la estadística descriptiva es resumir de manera concisa lo que *sabemos*. Por el contrario, el propósito de la estadística inferencial es "aprender lo que no sabemos a partir de lo que sabemos". Ahora que tenemos una base en la teoría de la probabilidad, estamos en una buena posición para pensar en el problema de la inferencia estadística. ¿Qué tipo de cosas nos gustaría aprender? ¿Y cómo las aprendemos? Estas son las preguntas que constituyen el nucleo de la estadística inferencial, y tradicionalmente se dividen en dos "grandes ideas": estimación y prueba de hipótesis. El objetivo de este capítulo es presentar la primera de estas grandes ideas, la teoría de la estimación, pero primero hablaré sobre la teoría del muestreo porque la teoría de la estimación no tiene sentido hasta que se entiende el muestreo. Como consecuencia, este capítulo se divide naturalmente en dos partes: las tres primeras secciones se centran en la teoría del muestreo y las dos últimas secciones hacen uso de la teoría del muestreo para discutir cómo piensan los estadísticos acerca de la estimación.

8.1 Muestras, poblaciones y muestreo

En el preludio de la parte IV, hablé del enigma de la inducción y destaqué el hecho de que todo aprendizaje requiere hacer suposiciones. Aceptando que esto es cierto, nuestra primera tarea consiste en plantear algunas hipótesis bastante generales sobre los datos que tengan sentido. Aquí es donde entra en juego la **teoría del muestreo**. Si la teoría de la probabilidad es la base sobre la que se construye toda la teoría estadística, la teoría del muestreo es el marco alrededor del cual se puede construir el resto de la casa. La teoría del muestreo juega un papel fundamental a la hora de especificar los supuestos en los que se basan las inferencias estadísticas. Y para hablar de "hacer inferencias" de la forma en que los estadísticos piensan en ello, debemos ser un poco más explícitas acerca de lo que estamos haciendo inferencias *a partir de* (la muestra) y sobre lo que estamos haciendo inferencias (la población).

En casi todas las situaciones de interés, lo que tenemos a nuestra disposición como in-

vestigadoras es una **muestra** de datos. Podríamos haber realizado un experimento con un número determinado de participantes, una empresa de sondeos podría haber llamado por teléfono a un número determinado de personas para preguntarles sobre las intenciones de voto, etc. De esta forma, el conjunto de datos de que disponemos es finito e incompleto. Es imposible que todas las personas del mundo participen en nuestro experimento; por ejemplo, una empresa de sondeos no tiene el tiempo ni el dinero para llamar a todos los votantes del país. En nuestro debate anterior sobre estadística descriptiva en Chapter 4, esta muestra era lo único que nos interesaba. Nuestro único objetivo era encontrar formas de describir, resumir y representar gráficamente esa muestra. Esto está a punto de cambiar.

8.1.1 Definir una población

Una muestra es algo concreto. Puede abrir un archivo de datos y allí están los datos de tu muestra. Una **población**, en cambio, es una idea más abstracta. Se refiere al conjunto de todas las personas posibles, o todas las observaciones posibles, sobre las que se quieren sacar conclusiones y, por lo general, es *mucho más grande* que la muestra. En un mundo ideal, el investigador comenzaría el estudio con una idea clara de cuál es la población de interés, ya que el proceso de diseñar un estudio y probar hipótesis con los datos depende de la población sobre la que se quiere hacer afirmaciones.

A veces es fácil establecer la población de interés. Por ejemplo, en el ejemplo de la "empresa de sondeos" que abrió el capítulo, la población estaba compuesta por todos los votantes inscritos en el momento del estudio, millones de personas. La muestra era un conjunto de 1000 personas que pertenecientes todas ellas a dicha población. En la mayoría de los estudios, la situación es mucho menos sencilla. En un experimento psicológico típico, determinar la población de interés es un poco más complicado. Supongamos que realizo un experimento con 100 estudiantes universitarios como participantes. Mi objetivo, como científica cognitiva, es intentar aprender algo sobre el funcionamiento de la mente. Entonces, ¿cuál de los siguientes contaría como "la población"?

- ¿Todos los estudiantes de psicología de la Universidad de Adelaida?
- ¿Los estudiantes de psicología en general, de cualquier parte del mundo?
- ¿Australianos que viven actualmente?
- ¿Australianos de edades similares a las de mi muestra?
- ¿Cualquier persona viva en la actualidad?
- ¿Cualquier ser humano, pasado, presente o futuro?
- ¿Cualquier organismo biológico con un grado de inteligencia suficiente que opere en un medio terrestre?
- ¿Cualquier ser inteligente?

Cada una de ellas define un grupo real de entidades poseedoras de mente, todas las cuales podrían interesarme como científica cognitiva, y no está nada claro cuál debería ser la verdadera población de interés. Como otro ejemplo, consideremos el juego Wellesley-Croker que discutimos en el Preludio de la parte IV. La muestra aquí es una secuencia específica de 12 victorias y 0 derrotas para Wellesley. ¿Cual es la población? De nuevo, no es obvio cuál es la población.

- ¿Todos los resultados hasta que Wellesley y Croker llegaron a su destino?
- ¿Todos los resultados si Wellesley y Croker hubieran jugado al juego durante el resto de sus vidas?

- ¿Todos los resultados si Wellseley y Croker vivieran para siempre y jugaran al juego hasta que el mundo se quedara sin colinas?
- ¿Todos los resultados si creáramos un conjunto infinito de universos paralelos y la pareja Wellesely/Croker adivinara las mismas 12 colinas en cada universo?

8.1.2 Muestras aleatorias simples

Independientemente de cómo definas la población, el punto crítico es que la muestra es un subconjunto de la población y nuestro objetivo es utilizar nuestro conocimiento de la muestra para hacer inferencias sobre las propiedades de la población. La relación entre ambos depende del procedimiento por el que se seleccionó la muestra. Este procedimiento se denomina **método de muestreo** y es importante entender por qué es importante.

Para simplificar, imaginemos que tenemos una bolsa con 10 fichas. Cada ficha lleva impresa una letra única para que podamos distinguir las 10 fichas. Las fichas son de dos colores, blanco y negro. Este conjunto de fichas es la población de interés y se representa gráficamente a la izquierda de Figure 8.1. Como puedes ver en la imagen, hay 4 fichas negras y 6 blancas, pero en la vida real no lo sabríamos a menos que miráramos en la bolsa. Ahora imagina que haces el siguiente "experimento": agitas la bolsa, cierras los ojos y sacas 4 fichas sin volver a meter ninguna en la bolsa. Primero sale la ficha a (negra), luego la c (blanca), después la j (blanca) y finalmente la b (negra). Si quisieras, podrías volver a meter todas las fichas en la bolsa y repetir el experimento, como se muestra en el lado derecho de Figure 8.1. Cada vez se obtienen resultados diferentes, pero el procedimiento es idéntico en todos los casos. El hecho de que el mismo procedimiento pueda llevar a resultados diferentes cada vez lo denominamos *proceso aleatorio*.[1] Sin embargo, como agitamos la bolsa antes de sacar ninguna ficha, parece razonable pensar que todas las fichas tienen la misma probabilidad de ser seleccionadas. Un procedimiento en el que cada miembro de la población tiene la misma probabilidad de ser seleccionado se denomina **muestra aleatoria simple**. El hecho de que no hayamos vuelto a meter las fichas en la bolsa después de sacarlas significa que no se puede observar lo mismo dos veces y, en tales casos, se dice que las observaciones se han muestreado **sin reemplazo**.

Para asegurarte de que comprendes la importancia del procedimiento de muestreo, considera una forma alternativa en la que podría haberse realizado el experimento. Supongamos que mi hijo de 5 años hubiera abierto la bolsa y decidido sacar cuatro fichas negras sin volver a meter ninguna en la bolsa. Este esquema de muestreo sesgado se representa en Figure 8.2. Consideremos ahora el valor probatorio de ver 4 fichas negras y 0 blancas. Está claro que depende del esquema de muestreo, ¿no? Si sabemos que el sistema de muestreo está sesgado para seleccionar solo fichas negras, entonces una muestra compuesta solo por fichas negras no nos dice mucho sobre la población. Por esta razón, a los estadísticos les gusta mucho cuando un conjunto de datos puede considerarse una muestra aleatoria simple, porque facilita *mucho* el análisis de datos.

Merece la pena mencionar un tercer procedimiento. Esta vez cerramos los ojos, agitamos la bolsa y sacamos una ficha. Esta vez, sin embargo, anotamos la observación y luego volvemos a meter la ficha en la bolsa. Volvemos a cerrar los ojos, agitamos la

[1] La definición matemática correcta de aleatoriedad es extraordinariamente técnica y va mucho más allá del alcance de este libro. No seremos técnicas aquí y diremos que un proceso tiene un elemento de aleatoriedad siempre que sea posible repetir el proceso y obtener respuestas diferentes cada vez.

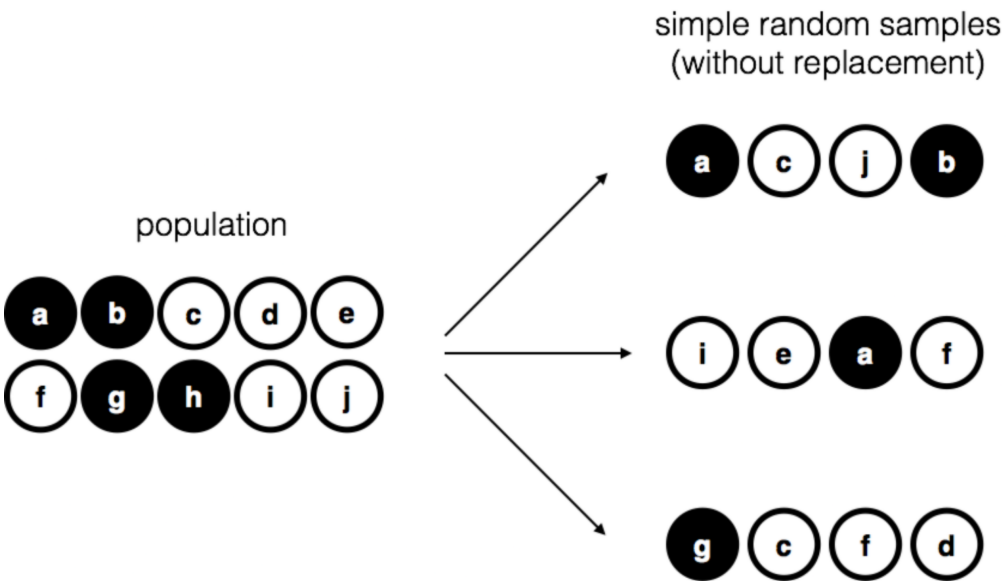


Figure 8.1: muestreo aleatorio simple sin reemplazo de una población finita

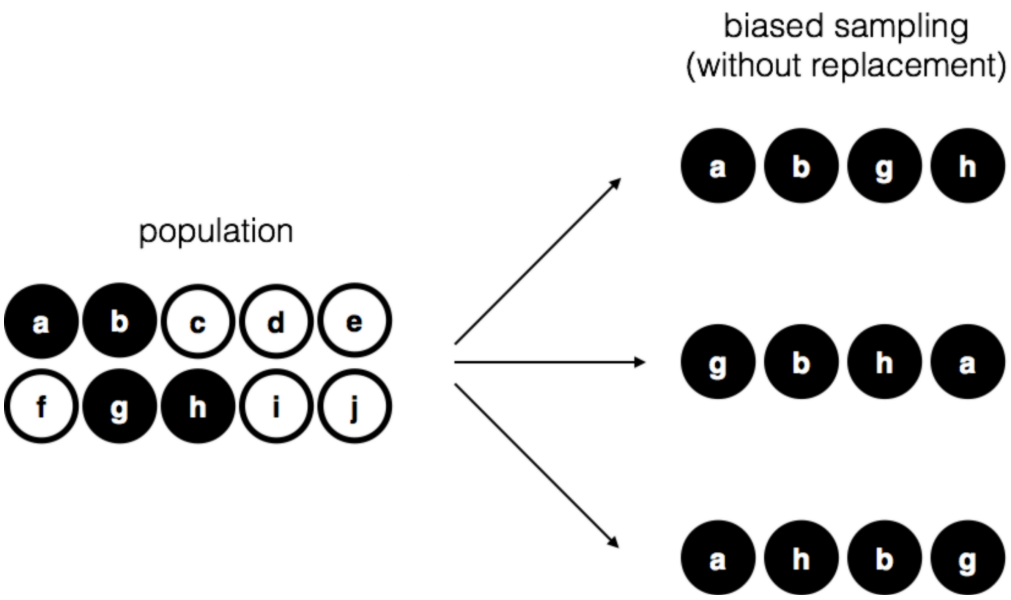


Figure 8.2: Muestreo sesgado sin reemplazo de una población finita

bolsa y sacamos una ficha. Repetimos este procedimiento hasta tener 4 fichas. Los conjuntos de datos generados de esta forma siguen siendo muestras aleatorias simples, pero como volvemos a meter las fichas en la bolsa inmediatamente después de extraerlas, se denomina muestra **con reemplazo**. La diferencia entre esta situación y la primera es que es posible observar al mismo miembro de la población varias veces, como se ilustra en Figure 8.3.

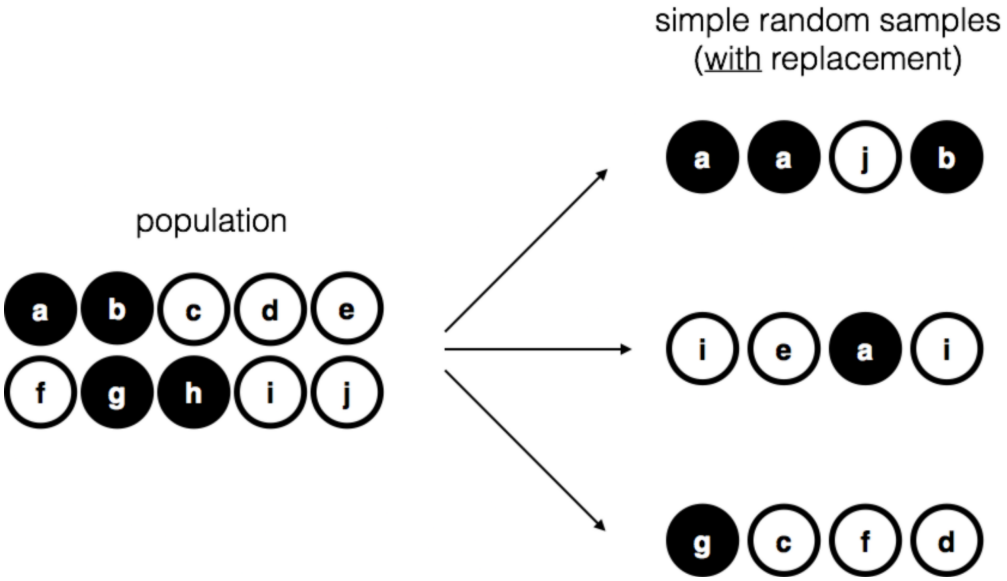


Figure 8.3: Muestreo aleatorio simple *con* reemplazo de una población finita

Según mi experiencia, la mayoría de los experimentos de psicología suelen ser muestreos sin reemplazo, porque no se permite que la misma persona participe en el experimento dos veces. Sin embargo, la mayor parte de la teoría estadística se basa en el supuesto de que los datos proceden de una muestra aleatoria simple **con reemplazo**. En la vida real, esto rara vez importa. Si la población de interés es grande (por ejemplo, tiene más de 10 entidades), la diferencia entre el muestreo con y sin reemplazo es demasiado pequeña como para preocuparse por ella. En cambio, la diferencia entre muestras aleatorias simples y muestras sesgadas no es tan fácil de descartar.

8.1.3 La mayoría de las muestras no son muestras aleatorias simples

Como se puede ver en la lista de posibles poblaciones que he mostrado antes, es casi imposible obtener una muestra aleatoria simple de la mayoría de las poblaciones de interés. Cuando realizo experimentos, consideraría un pequeño milagro que mis participantes fueran una muestra aleatoria de los estudiantes de psicología de la Universidad de Adelaida, aunque esta es, con diferencia, la población más reducida a la que podría querer generalizar. Un análisis exhaustivo de otros tipos de sistemas de muestreo queda fuera del alcance de este libro, pero para que te hagas una idea de lo que hay, enumeraré algunos de los más importantes.

- *Muestreo estratificado.* Supongamos que tu población está (o puede estar) dividida en varias subpoblaciones o estratos diferentes. Por ejemplo, puede que estés realizando un estudio en varios sitios diferentes. En lugar de intentar obtener muestras aleatorias de la población en su conjunto, se intenta recoger una muestra aleatoria separada de cada uno de los estratos. El muestreo estratificado a veces es más fácil de realizar que el muestreo aleatorio simple, sobre todo cuando la población ya está dividida en estratos distintos. También puede ser más eficaz que el muestreo aleatorio simple, especialmente cuando algunas de las subpoblaciones son poco frecuentes. Por ejemplo, cuando se estudia la esquizofrenia, sería mucho mejor dividir la población en dos [2] estratos (esquizofrénicos y no esquizofrénicos) y luego muestrear un número igual de personas de cada grupo. Si seleccionaras personas al azar, tendrías tan pocas personas esquizofrénicas en la muestra que tu estudio sería inútil. Este tipo específico de muestreo estratificado se conoce como sobremuestreo porque intenta deliberadamente sobrerrepresentar a grupos poco frecuentes.
- El *muestreo de bola de nieve* es una técnica especialmente útil cuando se toman muestras de una población "oculta" o de difícil acceso y es especialmente habitual en ciencias sociales. Por ejemplo, supongamos que los investigadores quieren realizar una encuesta de opinión entre personas transgénero. Es posible que el equipo de investigación solo disponga de los datos de contacto de unas pocas personas trans, por lo que la encuesta comienza pidiéndoles que participen (etapa 1). Al final de la encuesta, se pide a los participantes que faciliten los datos de contacto de otras personas que puedan querer participar. En la etapa 2 se encuesta a esos nuevos contactos. El proceso continúa hasta que los investigadores disponen de datos suficientes. La gran ventaja del muestreo de bola de nieve es que obtiene datos en situaciones que de otro modo serían imposibles de obtener. Desde el punto de vista estadístico, la principal desventaja es que la muestra es muy poco aleatoria, y poco aleatoria en aspectos difíciles de abordar. En la vida real, la desventaja es que el procedimiento puede ser poco ético si no se maneja bien, porque las poblaciones ocultas suelen estar oculatas por una razón. He elegido a las personas transgénero como ejemplo para destacar este problema. Si no se tiene cuidado, se puede acabar delatando a personas que no quieren ser delatadas (muy, muy mala forma), e incluso si no se comete ese error, puede resultar intrusivo usar las redes sociales de las personas para estudiarlas. Sin duda, es muy difícil obtener el consentimiento informado de las personas antes de ponerse en contacto con ellas, pero en muchos casos el simple hecho de ponerse en contacto con ellas y decirles "oye, queremos estudiarte" puede resultar hiriente. Las redes sociales son cosas complejas, y sólo porque puedas usarlas para obtener datos no siempre significa que debas hacerlo.
- *Muestreo de conveniencia* es más o menos lo que parece. Las muestras se eligen de forma que convenga a la investigadora y no se seleccionan al azar de la población de interés. El muestreo de bola de nieve es un tipo de muestreo de conveniencia, pero hay muchos otros. Un ejemplo común en psicología son los estudios que se basan en estudiantes universitarios de psicología. Estas muestras son generalmente no aleatorias en dos aspectos. En primer lugar, recurrir a estudiantes universitarios de psicología significa automáticamente que los datos se limitan a

[2] Nada en la vida es tan sencillo. No existe una división obvia de las personas en categorías binarias como "esquizofrénico" y "no esquizofrénico". Pero este no es un texto de psicología clínica, así que os ruego que me perdonéis algunas simplificaciones aquí y allá.

una única subpoblación. En segundo lugar, los estudiantes suelen elegir en qué estudios participarán, por lo que la muestra es un subconjunto de estudiantes de psicología autoseleccionados y no un subconjunto seleccionado al azar. En la vida real, la mayoría de los estudios son muestras de conveniencia de una forma u otra. Esto es a veces una limitación grave, pero no siempre.

8.1.4 ¿CQué importancia tiene no tener una muestra aleatoria simple?

De acuerdo, la recogida de datos en el mundo real no suele consistir en agradables muestras aleatorias simples. ¿Eso importa? Un poco de reflexión te dejará claro que puede importar si tus datos no son una muestra aleatoria simple. Piensa en la diferencia entre Figure 8.1 y Figure 8.2. Sin embargo, no es tan malo como parece. Algunos tipos de muestras sesgadas no plantean ningún problema. Por ejemplo, cuando utiliza una técnica de muestreo estratificado, realmente sabes cuál es el sesgo porque lo has creado deliberadamente, a menudo para *aumentar* la eficacia de tu estudio, y existen técnicas estadísticas que puedes utilizar para ajustar los sesgos que has introducido (no tratadas en este libro). Así que en esas situaciones no es un problema.

Sin embargo, en términos más generales, es importante recordar que el muestreo aleatorio es un medio para alcanzar un fin, y no el fin en sí mismo. Supongamos que has recurrido a una muestra de conveniencia y, como tal, puedes suponer que está sesgada. Un sesgo en tu método de muestreo es solo un problema si te lleva a sacar conclusiones equivocadas. Visto desde esa perspectiva, yo diría que no necesitamos que la muestra se genere aleatoriamente en *todos* los aspectos, solo necesitamos que sea aleatoria con respecto al fenómeno psicológicamente relevante de interés. Supongamos que estoy haciendo un estudio sobre la capacidad de memoria de trabajo. En el estudio 1, puedo tomar muestras aleatorias de todos los seres humanos vivos, con una excepción: solo puedo tomar muestras de personas nacidas un lunes. En el estudio 2, puedo tomar muestras al azar de la población australiana. Quiero generalizar mis resultados a la población de todos los seres humanos vivos. ¿Qué estudio es mejor? La respuesta, obviamente, es el estudio 1. ¿Por qué? Porque no tenemos ninguna razón para pensar que "nacer un lunes" tenga alguna relación interesante con la capacidad de la memoria de trabajo. En cambio, se me ocurren varias razones por las que "ser australiano" podría ser importante. Australia es un país rico e industrializado con un sistema educativo muy bien desarrollado. Las personas que crecen en ese sistema habrán tenido experiencias vitales mucho más parecidas a las de las personas que diseñaron las pruebas de capacidad de memoria de trabajo. Esta experiencia compartida podría traducirse fácilmente en creencias similares sobre cómo "hacer un examen", una suposición compartida sobre cómo funciona la experimentación psicológica, etc. Estas cosas podrían ser realmente importantes. Por ejemplo, el estilo de "hacer exámenes" podría haber enseñado a los participantes australianos a dirigir su atención exclusivamente a materiales de examen bastante abstractos mucho más que a las personas que no han crecido en un entorno similar. Por tanto, esto podría dar lugar a una imagen engañosa de lo que es la capacidad de memoria de trabajo.

Hay dos puntos ocultos en esta discusión. En primer lugar, al diseñar tus propios estudios, es importante pensar en qué población te interesa y esforzarte por muestrear de forma adecuada esa población. En la práctica, una suele verse obligada a conformarse con una "muestra de conveniencia" (por ejemplo, los profesores de psicología recogen

muestras de las estudiantes de psicología porque es la forma menos costosa de recopilar datos, y nuestras arcas no están precisamente rebosantes de oro), pero si es así una debería al menos dedicar algún tiempo a pensar cuáles pueden ser los peligros de esta práctica. En segundo lugar, si vas a criticar el estudio de otra persona porque ha utilizado una muestra de conveniencia en lugar de realizar un laborioso muestreo aleatorio de toda la población humana, al menos ten la cortesía de ofrecer una teoría específica sobre cómo esto podría haber distorsionado los resultados.

8.1.5 Parámetros poblacionales y estadísticas muestrales

Bueno. Dejando a un lado las espinosas cuestiones metodológicas asociadas a la obtención de una muestra aleatoria, consideremos una cuestión ligeramente diferente. Hasta ahora hemos hablado de poblaciones como lo haría un científico. Para una psicóloga, una población podría ser un grupo de personas. Para un ecologista, una población podría ser un grupo de osos. En la mayoría de los casos, las poblaciones que preocupan a los científicos son cosas concretas que existen en el mundo real. Los estadísticos, sin embargo, son un grupo curioso. Por un lado, les interesan los datos del mundo real y la ciencia real igual que a los científicos. Por otro lado, también operan en el ámbito de la abstracción pura, al igual que los matemáticos. En consecuencia, la teoría estadística tiende a ser un poco abstracta en cuanto a la definición de población. Del mismo modo que los investigadores psicológicos operacionalizan nuestras ideas teóricas abstractas en términos de mediciones concretas (Section 2.1), los estadísticos operacionalizan el concepto de "población" en términos de objetos matemáticos con los que saben trabajar. Estos objetos ya los hemos visto en Chapter 7. Se llaman distribuciones de probabilidad.

La idea es bastante sencilla. Digamos que estamos hablando de puntuaciones de CI. Para un psicólogo, la población de interés es un grupo de seres humanos reales que tienen puntuaciones de CI. Un estadístico "simplifica" esto definiendo operativamente la población como la distribución de probabilidad representada en Figure 8.4 (a). Las pruebas de CI están diseñadas para que el CI promedio sea 100, la desviación estándar de las puntuaciones de CI sea 15 y la distribución de las puntuaciones de CI sea normal. Estos valores se denominan **parámetros poblacionales** porque son características de toda la población. Es decir, decimos que la media poblacional μ es 100 y la desviación estándar poblacional es 15.

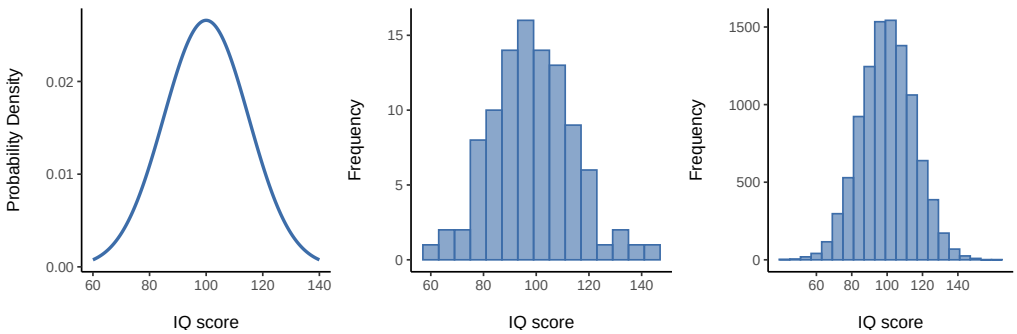


Figure 8.4: La distribución poblacional de puntuaciones de CI (panel (a)) y dos muestras extraídas al azar de ella. En el panel (b) tenemos una muestra de 100 observaciones, y en el panel (c) tenemos una muestra de 10,000 observaciones

Supongamos que hago un experimento. Selecciono 100 personas al azar y les administro una prueba de CI, lo que me da una muestra aleatoria simple de la población. Mi muestra consistiría en una colección de números como esta:

106 101 98 80 74 … 107 72 100

Cada una de estas puntuaciones de CI es una muestra de una distribución normal con media 100 y desviación estándar 15. Así que si trazo un histograma de la muestra, obtengo algo como lo que se muestra en Figure 8.4 (b). Como puedes ver, el histograma tiene aproximadamente la forma correcta, pero es una aproximación muy burda a la distribución real de la población que se muestra en Figure 8.4 (a). Cuando calculo la media de mi muestra, obtengo un número bastante cercano a la media poblacional 100, pero no idéntico. En este caso, resulta que las personas de mi muestra tienen una media de CI de 98,5 y la desviación estándar de sus puntuaciones de CI es 15,9. Estos **estadísticos muestrales** son propiedades de mi conjunto de datos y, aunque son bastante similares a los valores reales de la población, no son iguales. En general, los estadísticos muestrales son las cosas que puedes calcular a partir de tu conjunto de datos y los parámetros poblacionales son las cosas sobre las que quieres aprender. Más adelante en este capítulo hablaré de [Estimar los parámetros poblacionales] utilizando tus estadísticos muestrales y también de [Estimar un intervalo de confianza], pero antes de llegar a eso hay algunas ideas más sobre la teoría del muestreo que debes conocer.

8.2 La ley de los grandes números

En la sección anterior, te mostré los resultados de un experimento ficticio sobre el CI con un tamaño de muestra de N = 100. Los resultados fueron algo alentadores, ya que la media poblacional real es 100 y la media muestral de 98,5 es una aproximación bastante razonable. En muchos estudios científicos ese nivel de precisión es perfectamente aceptable, pero en otras situaciones necesitas ser mucho más precisa. Si queremos que los estadísticos muestrales se acerquen mucho más a los parámetros poblacionales, ¿qué podemos hacer? La respuesta obvia es recopilar más datos. Supongamos que realizamos un experimento mucho mayor, esta vez midiendo el CI de 10.000 personas. Podemos simular los resultados de este experimento usando jamovi. El archivo IQsim.omv es un archivo de datos de jamovi. En este archivo he generado 10,000 números aleatorios muestreados a partir de una distribución normal para una población con media = 100 y sd = 15. Esto se ha hecho calculando una nueva variable usando la función = NORM(100,15). Un histograma y un gráfico de densidad muestran que esta muestra más grande es una aproximación mucho mejor a la verdadera distribución de la población que la más pequeña. Esto se refleja en los estadísticos muestrales. La media del CI de la muestra más grande es de 99,68 y la desviación estándar es 14,90. Estos valores ahora están muy próximos a la población real. Ver Figure 8.5.

Me da un poco de vergüenza decirlo, pero lo que quiero que entiendas es que las muestras grandes suelen dar mejor información. Me da un poco de vergüenza decirlo porque es tan obvio que no hace falta decirlo. De hecho, es un punto tan obvio que cuando Jacob Bernoulli, uno de los fundadores de la teoría de la probabilidad, formalizó esta idea allá por 1713, fue un poco idiota al respecto. Así es como describió el hecho de que todos compartimos esta intuición:

> *Porque hasta el más estúpido de los hombres, por algún instinto de la naturaleza, por sí mismo y sin ninguna instrucción (lo cual es algo notable),*

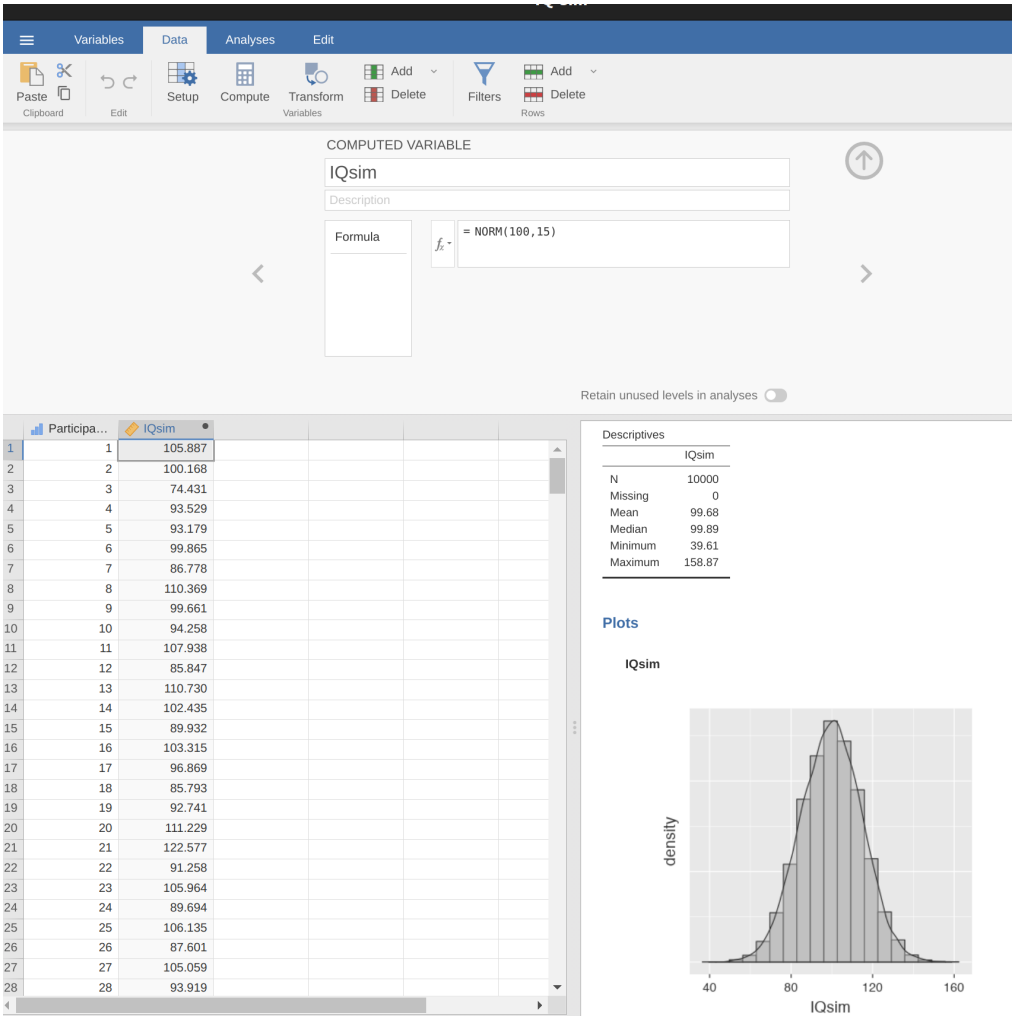


Figure 8.5: Una muestra aleatoria extraída de una distribución normal usando jamovi

> *está convencido de que cuantas más observaciones se han hecho, menor es el peligro de desviarse de propia la meta.* (Stigler, 1986, p. 65).

De acuerdo, el pasaje resulta un poco condescendiente (por no decir sexista), pero su argumento principal es correcto. Es obvio que con más datos se obtienen mejores respuestas. La pregunta es: ¿por qué? No es sorprendente que esta intuición que todos compartimos resulte ser correcta, y los estadísticos se refieren a ella como la **ley de los grandes números**. La ley de los grandes números es una ley matemática que se aplica a muchos estadísticos muestrales diferentes, pero la forma más sencilla de entenderla es como una ley sobre promedios. La media muestral es el ejemplo más obvio de un estadístico que se basa en el promedio (porque eso es lo que es la media… un promedio), así que veámosla. Cuando se aplica a la media muestral, la ley de los grandes números establece que a medida que aumenta la muestra, la media muestral tiende a acercarse a la verdadera media de la población. O, para decirlo con un poco más de precisión, a medida que el tamaño de la muestra "se acerca" al infinito (escrito como $N \longrightarrow \infty$), la media muestral se acerca a la media poblacional $\bar{X} \longrightarrow \mu$) [3]

No pretendo someterte a una prueba de que la ley de los grandes números es cierta, pero es una de las herramientas más importantes de la teoría estadística. La ley de los grandes números es lo que podemos usar para justificar nuestra creencia de que recoger cada vez más datos nos llevará finalmente a la verdad. Para cualquier conjunto de datos concreto, los estadísticos muestrales que calculemos a partir de él serán erróneos, pero la ley de los grandes números nos dice que si seguimos recopilando más datos, esos estadísticos muestrales tenderán a acercarse cada vez más a los verdaderos parámetros poblacionales.

8.3 Distribuciones muestrales y el teorema central del límite

La ley de los grandes números es una herramienta muy poderosa, pero no va a ser suficiente para responder a todas nuestras preguntas. Entre otras cosas, lo único que nos da es una "garantía a largo plazo". A largo plazo, si de alguna manera pudiéramos recopilar una cantidad infinita de datos, la ley de los grandes números nos garantizaría que nuestros estadísticos muestrales serían correctos. Pero, como dijo John Maynard Keynes en economía, una garantía a largo plazo sirve de poco en la vida real.

> *[El] largo plazo es una guía engañosa de la actualidad. A largo plazo, todos estaremos muertos. Los economistas se imponen una tarea demasiado fácil, demasiado inútil, si en las estaciones tempestuosas solo pueden decirnos que cuando la tormenta ha pasado hace tiempo, el océano vuelve a estar plano.* (Keynes, 1923, p. 80).

Como en economía, también en psicología y estadística. No basta con saber que al final llegaremos a la respuesta correcta cuando calculemos la media muestral. Saber que un

[3] Técnicamente, la ley de los grandes números se aplica a cualquier estadístico muestral que pueda describirse como un promedio de cantidades independientes. Esto es cierto para la media muestral. Sin embargo, también es posible escribir muchos otros estadísticos muestrales como promedios de una forma u otra. La varianza de una muestra, por ejemplo, se puede reescribir como un tipo de promedio y, por tanto, está sujeta a la ley de los grandes números. Sin embargo, el valor mínimo de una muestra no se puede escribir como un promedio de nada y, por lo tanto, no se rige por la ley de los grandes números.

conjunto de datos infinitamente grande me dirá el valor exacto de la media poblacional es un consuelo frío cuando mi conjunto de datos real tiene un tamaño de muestra de $N = 100$. En la vida real, por tanto, debemos saber algo sobre el comportamiento de la media muestral cuando se calcula a partir de un conjunto de datos más modesto.

8.3.1 Distribución muestral de la media

Teniendo esto en cuenta, abandonemos la idea de que nuestros estudios tendrán tamaños de muestra de 10.000 y consideremos en su lugar un experimento muy modesto. Esta vez tomaremos una muestra de $N = 5$ personas y mediremos sus puntuaciones de CI. Como antes, puedo simular este experimento en la función jamovi = NORM(100,15), pero esta vez solo necesito 5 ID de participantes, no 10,000. Estos son los cinco números que generó jamovi:

90 82 94 99 110

El CI medio en esta muestra resulta ser exactamente 95. No es sorprendente que sea mucho menos preciso que el experimento anterior. Ahora imaginemos que decido **replicar** el experimento. Es decir, repito el experimento lo más fielmente posible y tomo al azar una muestra de 5 personas nuevas y mido su CI. De nuevo, jamovi me permite simular los resultados de este procedimiento y genera estos cinco números:

78 88 111 111 117

Esta vez, el CI medio en mi muestra es 101. Si repito el experimento 10 veces, obtengo los resultados que se muestran en Table 8.1 y, como se puede ver, la media de la muestra varía de una repetición a otra.

Table 8.1: Diez repeticiones del experimento CI, cada una con un tamaño de muestra de ($N = 5$)

	Person 1	**Person 2**	**Person 3**	**Person 4**	**Person 5**	**Sample Mean**
Rep. 1	90	82	94	99	110	95.0
Rep. 2	78	88	111	111	117	101.0
Rep. 3	111	122	91	98	86	101.6
Rep. 4	98	96	119	99	107	103.8
Rep. 5	105	113	103	103	98	104.4
Rep. 6	81	89	93	85	114	92.4
Rep. 7	100	93	108	98	133	106.4
Rep. 8	107	100	105	117	85	102.8
Rep. 9	86	119	108	73	116	100.4
Rep. 10	95	126	112	120	76	105.8

Supongamos ahora que decido seguir así, replicando este experimento de "cinco puntuaciones de CI" una y otra vez. Cada vez que reproduzco el experimento anoto la media muestral. Con el tiempo, estaría acumulando un nuevo conjunto de datos, en el que cada experimento genera un único punto de datos. Las primeras 10 observaciones de

mi conjunto de datos son las medias muestrales enumeradas en Table 8.1, por lo que mi conjunto de datos comienza así:

95,0 101,0 101,6 103,8 104,4 …

¿Y si continuara así durante 10,000 repeticiones y luego dibujara un histograma? Bueno, eso es exactamente lo que hice, y puedes ver los resultados en Figure 8.6. Como ilustra esta imagen, el promedio de 5 puntuaciones de CI suele estar entre 90 y 110. Pero lo que es más importante, lo que pone de relieve es que si replicamos un experimento una y otra vez, ¡lo que obtenemos al final es una distribución de medias muestrales! (Table 8.1)) Esta distribución tiene un nombre especial en estadística, se llama **distribución muestral de la media**.

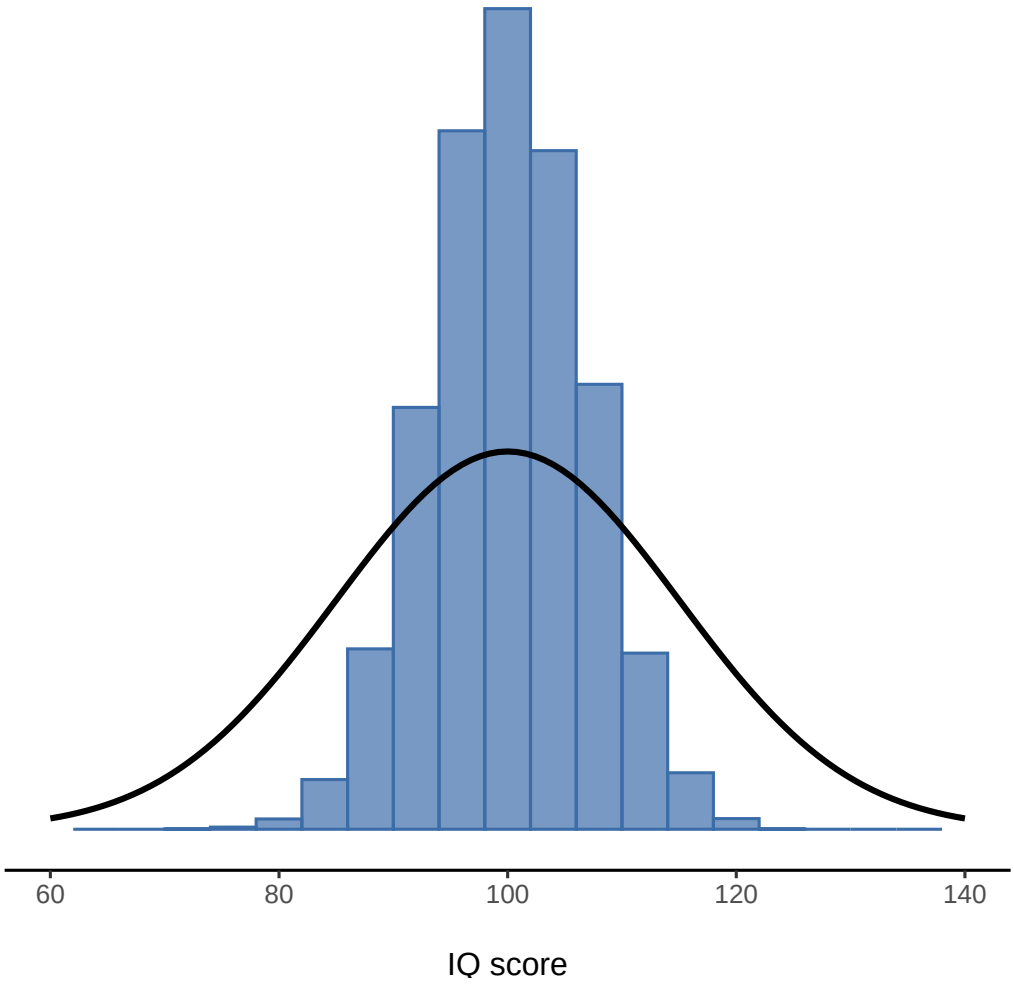


Figure 8.6: La distribución muestral de la media del 'experimento de las cinco puntuaciones de CI'. Si se toma una muestra de 5 personas al azar y se calcula su CI promedio, es casi seguro que se obtendrá un número entre 80 y 120, aunque hay bastantes personas que tienen CI superiores a 120 o inferiores a 80. A modo de comparación, la línea negra muestra la distribución poblacional de las puntuaciones de CI

Las distribuciones muestrales son otra idea teórica importante en estadística, y son cruciales para comprender el comportamiento de las muestras pequeñas. Por ejemplo, cuando realicé el primer experimento de "cinco puntuaciones de CI", la media de la muestra resultó ser 95. Sin embargo, lo que nos dice la distribución muestral en Figure 8.6 es que el experimento de "cinco puntuaciones de CI" no es muy preciso. Si repito el experimento, la distribución muestral me dice que puedo esperar ver una media muestral entre 80 y 120.

8.3.2 ¡Existen distribuciones muestrales para cualquier estadístico muestral!

Una cosa que hay que tener en cuenta cuando se piensa en distribuciones muestrales es que cualquier estadística muestral que quiera calcular tiene una distribución muestral. Por ejemplo, supongamos que cada vez que repito el experimento de las "cinco puntuaciones de CI" escribo la puntuación de CI más alta del experimento. Esto me daría un conjunto de datos que empezaría así:

110 117 122 119 113 …

Hacer esto una y otra vez me daría una distribución muestral muy diferente, a saber, la distribución muestral del máximo. La distribución muestral del máximo de 5 puntuaciones de CI se muestra en Figure 8.7. No es de extrañar que si eliges a 5 personas al azar y luego encuentras a la persona con la puntuación de cociente intelectual más alta, vaya a tener un CI superior al promedio. La mayoría de las veces acabarás con alguien cuyo CI se mida en el rango de 100 a 140.

8.3.3 El teorema central del límite

A estas alturas espero que tengas una idea bastante clara de lo que son las distribuciones muestrales y, en particular, de lo que es la distribución muestral de la media. En esta sección quiero hablar de cómo cambia la distribución muestral de la media en función del tamaño de la muestra. Intuitivamente, ya sabes parte de la respuesta. Si solo tienes unas pocas observaciones, es probable que la media muestral sea bastante inexacta. Si replicas un experimento pequeño y vuelves a calcular la media, obtendrás una respuesta muy diferente. En otras palabras, la distribución muestral es bastante amplia. Si replicas un experimento grande y vuelves a calcular la media muestral, probablemente obtendrás la misma respuesta que obtuviste la última vez, por lo que la distribución muestral será muy estrecha. Puedes ver esto visualmente en Figure 8.8, que muestra que cuanto mayor es el tamaño de la muestra, más estrecha es la distribución de muestreo. Podemos cuantificar este efecto calculando la desviación estándar de la distribución de muestreo, que se denomina **error estándar**. El error estándar de un estadístico se suele denotar SE, y como normalmente nos interesa el error estándar de la media muestral, a menudo usamos el acrónimo SEM. Como se puede ver con solo mirar la imagen, a medida que aumenta el tamaño de la muestra N, el SEM disminuye.

Bien, esa es una parte de la historia. Sin embargo, hay algo que he pasado por alto hasta ahora. Todos mis ejemplos hasta ahora se han basado en los experimentos de "puntuaciones de CI", y como las puntuaciones de CI se distribuyen de forma aproximadamente normal, he supuesto que la distribución de la población es normal. ¿Y si no es normal? ¿Qué ocurre con la distribución muestral de la media? Lo sorprendente es que, sea cual sea la forma de la distribución de la población, a medida que N aumenta,

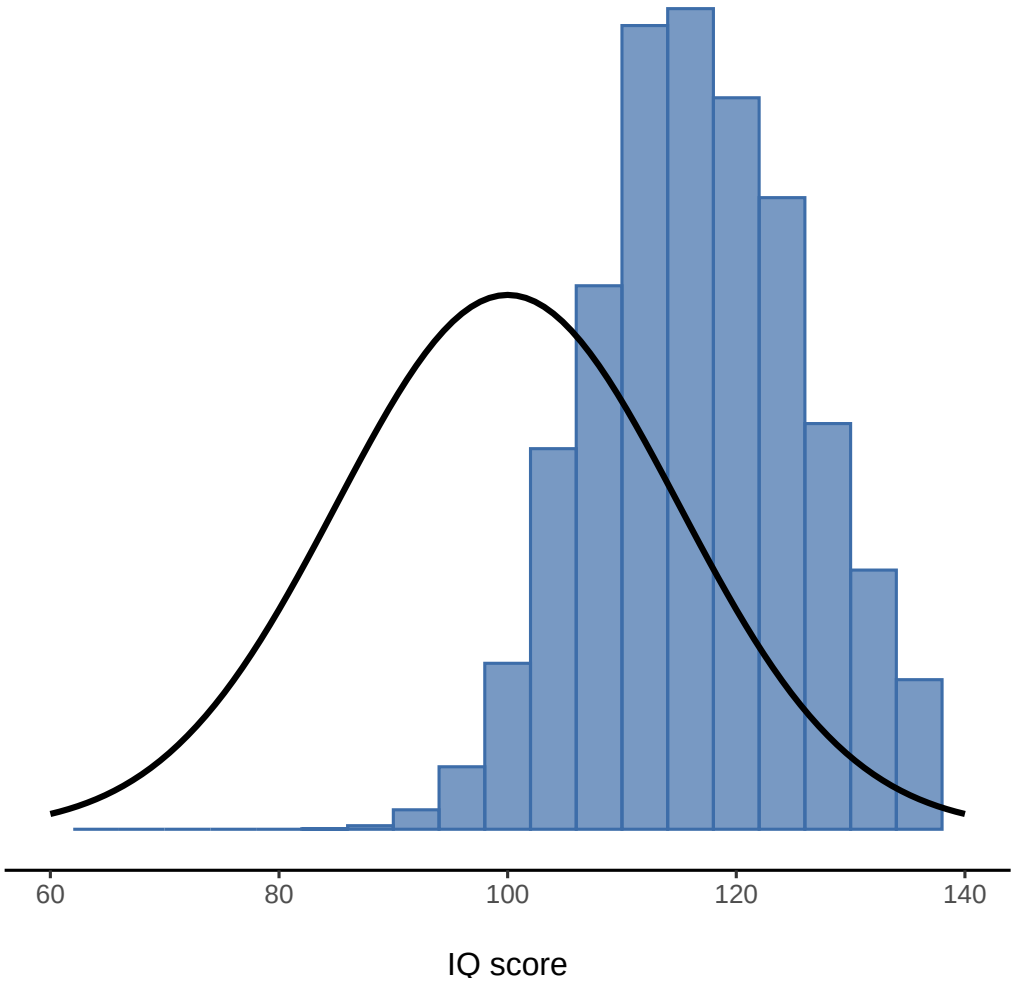


Figure 8.7: La distribución muestral del máximo para el 'experimento de las cinco puntuaciones de CI'. Si tomas una muestra de 5 personas al azar y seleccionas a la que tenga la puntuación de CI más alta, probablemente verás a alguien con un CI entre 100 y 140.

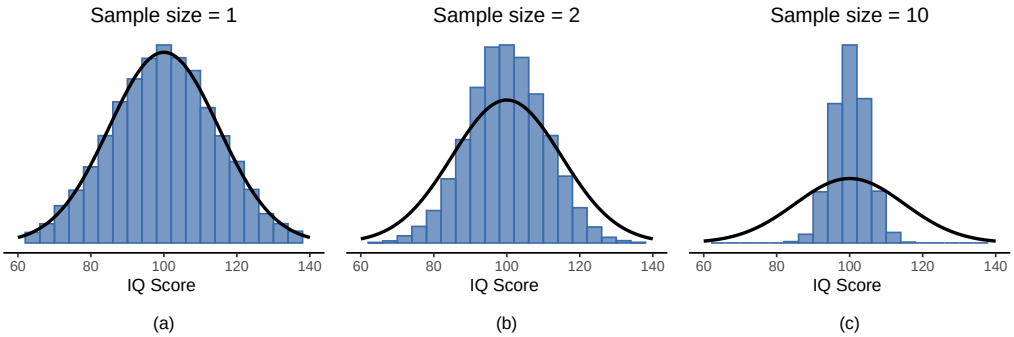


Figure 8.8: Ilustración de cómo la distribución muestral de la media depende del tamaño de la muestra. En cada panel, he generado 10 000 muestras de datos de coeficiente intelectual y he calculado la media de CI observada en cada uno de estos conjuntos de datos. Los histogramas de estos gráficos muestran la distribución de estas medias (es decir, la distribución muestral de la media). Cada puntuación individual de CI se extrajo de una distribución normal con una media de 100 y una desviación estándar de 15, que se muestra como la línea negra continua. En el panel (a), cada conjunto de datos contiene una única observación, por lo que la media de cada muestra es la puntuación de CI de una sola persona. En consecuencia, la distribución muestral de la media es, por supuesto, idéntica a la distribución poblacional de las puntuaciones de CI. Sin embargo, cuando aumentamos el tamaño de la muestra a 2, la media de cualquier muestra tiende a estar más cerca de la media de la población que la puntuación de CI de cualquier persona, por lo que el histograma (es decir, la distribución muestral) es un poco más estrecho que la distribución de la población. En el momento en que aumentamos el tamaño de la muestra a 10 (panel (c)), podemos ver que la distribución de las medias muestrales tiende a agruparse bastante estrechamente en torno a la media real de la población.

la distribución muestral de la media empieza a parecerse más a una distribución normal. Para que te hagas una idea, he realizado algunas simulaciones. Para ello, empecé con la distribución "en rampa" que se muestra en el histograma en Figure 8.9. Como se puede ver al comparar el histograma de forma triangular con la curva de campana trazada por la línea negra, la distribución de la población no se parece mucho a una distribución normal. A continuación, simulé los resultados de un gran número de experimentos. En cada experimento, tomé $N = 2$ muestras de esta distribución y calculé la media muestral. Figure 8.9 (b) representa el histograma de estas medias muestrales (es decir, la distribución muestral de la media para $N = 2$). Esta vez, el histograma produce una distribución en forma de χ^2. Sigue sin ser normal, pero está mucho más cerca de la línea negra que la distribución de la población en Figure 8.9 (a). Cuando aumento el tamaño de la muestra a $N = 4$, la distribución muestral de la media es muy cercana a la normal (Figure 8.9 (c)), y cuando llegamos a un tamaño de muestra de N = 8 es casi perfectamente normal. En otras palabras, mientras el tamaño de la muestra no sea pequeño, la distribución muestral de la media será aproximadamente normal, ¡independientemente de cómo sea la distribución de la población!

A partir de estas cifras, parece que tenemos pruebas de todas las afirmaciones siguientes sobre la distribución muestral de la media.

- La media de la distribución muestral es la misma que la media de la población
- La desviación estándar de la distribución muestral (es decir, el error estándar) disminuye a medida que aumenta el tamaño de la muestra
- La forma de la distribución muestral se vuelve normal a medida que aumenta el tamaño de la muestra.

Resulta que no solo todas estas afirmaciones son ciertas, sino que hay un teorema muy famoso en estadística que demuestra las tres cosas, conocido como el **teorema central del límite**. Entre otras cosas, el teorema central del límite nos dice que si la distribución de la población tiene media μ y desviación estándar σ, entonces la distribución muestral de la media también tiene media μ y el error estándar de la media es

$$SEM = \frac{\sigma}{\sqrt{N}}$$

Como dividimos la desviación estándar de la población σ por la raíz cuadrada del tamaño de la muestra N, el SEM se hace más pequeño a medida que aumenta el tamaño de la muestra. También nos dice que la forma de la distribución muestral se vuelve normal.[4]

Este resultado es útil para todo tipo de cosas. Nos dice por qué los experimentos grandes son más fiables que los pequeños, y como nos da una fórmula explícita para el error estándar, nos dice cuánto más fiable es un experimento grande. Nos dice por qué la distribución normal es, bueno, normal. En los experimentos reales, muchas de las cosas que queremos medir son en realidad promedios de muchas cantidades diferentes (por

[4]Como de costumbre, estoy siendo un poco descuidada aquí. El teorema central del límite es un poco más general de lo que parece en esta sección. Como en la mayoría de los textos de introducción a la estadística, he tratado una situación en la que se cumple el teorema central del límite: cuando se toma la media de muchos sucesos independientes extraídos de la misma distribución. Sin embargo, el teorema central del límite es mucho más amplio que esto. Por ejemplo, hay toda una clase de cosas llamadas "estadísticos U", todas las cuales cumplen el teorema central del límite y, por lo tanto, se distribuyen normalmente para muestras de gran tamaño. La media es uno de esos estadísticos, pero no es el único.

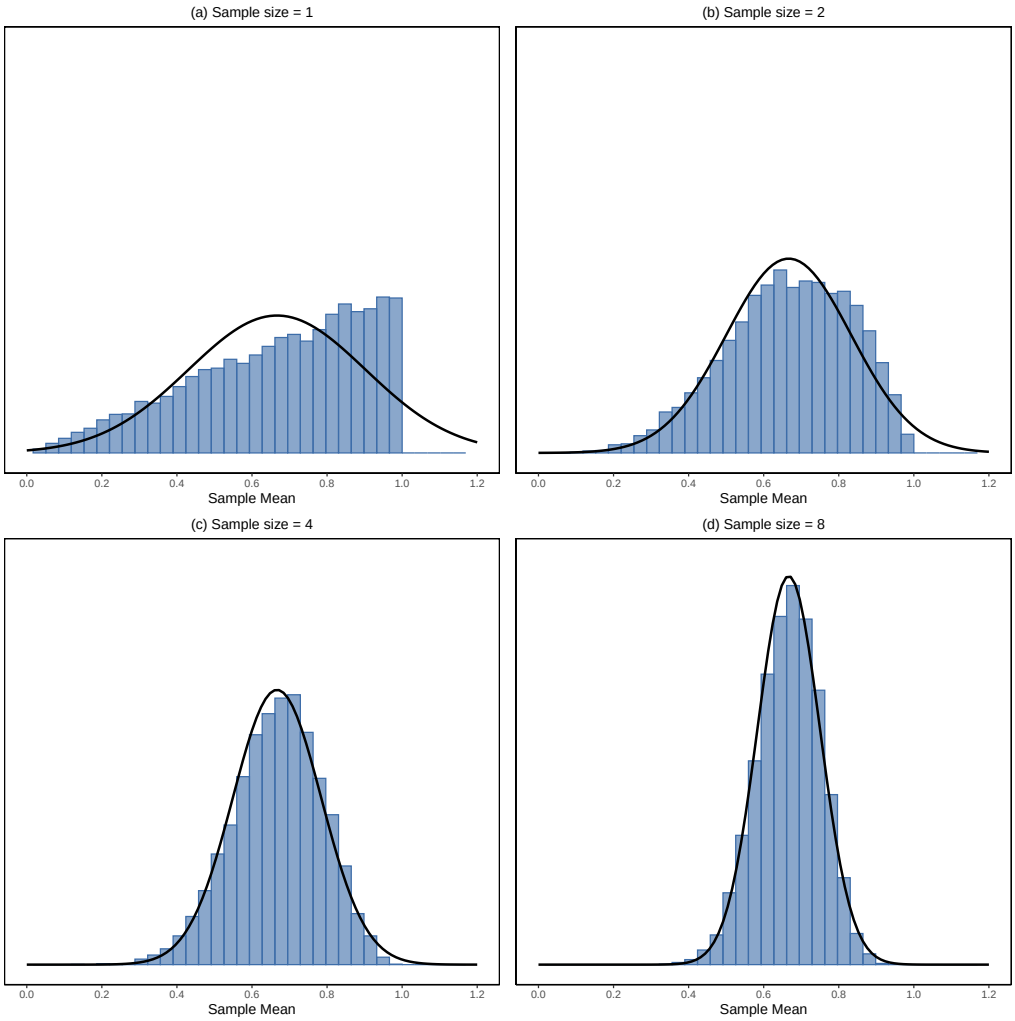


Figure 8.9: Demostración del teorema central del límite. En el panel (a), tenemos una distribución de población no normal, y los paneles (b)-(d) muestran la distribución muestral de la media para muestras de tamaño 2, 4 y 8 para datos extraídos de la distribución en el panel (a). Como se puede ver, aunque la distribución original de la población no es normal, la distribución muestral de la media se aproxima bastante a la normalidad cuando se tiene una muestra de 4 observaciones.

ejemplo, podría decirse que la inteligencia "general" medida por el CI es un promedio de una gran cantidad de habilidades y capacidades "específicas"), y cuando esto ocurre, la cantidad promediada debería seguir una distribución normal. Debido a esta ley matemática, la distribución normal aparece una y otra vez en los datos reales.

8.4 Estimación de los parámetros poblacionales

En todos los ejemplos de CI de las secciones anteriores conocíamos de antemano los parámetros de la población. Como se enseña a todos los estudiantes universitarios en su primera lección sobre la medición de la inteligencia, las puntuaciones del CI se definen con una media de 100 y una desviación estándar de 15. Sin embargo, esto es un poco mentira. ¿Cómo sabemos que las puntuaciones de CI tienen una media poblacional real de 100? Bueno, lo sabemos porque las personas que diseñaron los tests los han administrado a muestras muy grandes y luego han "amañado" las reglas de puntuación para que su muestra tenga una media de 100. Eso no es malo, por supuesto, es una parte importante del diseño de una medida psicológica. Sin embargo, es importante tener en cuenta que esta media teórica de 100 solo se aplica a la población que los diseñadores de los tests usaron para diseñarlos. De hecho, los buenos diseñadores de tests harán todo lo posible para proporcionar "normas de tests" que puedan aplicarse a muchas poblaciones diferentes (por ejemplo, diferentes grupos de edad, nacionalidades, etc.).

Esto es muy práctico, pero, por supuesto, casi todos los proyectos de investigación de interés implican analizar una población de personas diferente a la utilizada en las normas de la prueba. Por ejemplo, supongamos que queremos medir el efecto de la intoxicación por plomo de bajo nivel en el funcionamiento cognitivo en Port Pirie, una ciudad industrial del sur de Australia con una fundición de plomo. Tal vez decidas comparar las puntuaciones de CI entre las personas de Port Pirie con una muestra comparable de Whyalla, una ciudad industrial del sur de Australia con una refinería de acero.[5] Independientemente de la ciudad en la que estés pensando, no tiene mucho sentido suponer simplemente que la media real de CI de la población sea 100. Que yo sepa, nadie ha elaborado datos normativos razonables que puedan aplicarse automáticamente a las ciudades industriales del sur de Australia. Tendremos que **estimar** los parámetros de la población a partir de una muestra de datos. ¿Y cómo lo hacemos?

[5]Ten en cuenta que si realmente estuvieras interesada en esta cuestión, tendrías que ser mucho más cuidadosa que yo. No se puede comparar sin más las puntuaciones de CI de Whyalla con las de Port Pirie y suponer que cualquier diferencia se debe a la intoxicación por plomo. Aunque fuera cierto que las únicas diferencias entre las dos ciudades correspondieran a las diferentes refinerías (y no lo es, ni mucho menos), hay que tener en cuenta que la gente ya cree que la contaminación por plomo provoca déficits cognitivos. Si volvemos a Chapter 2, esto significa que hay diferentes efectos de demanda para la muestra de Port Pirie que para la muestra de Whyalla. En otras palabras, es posible que los datos muestren una diferencia de grupo ilusoria en tus datos, causada por el hecho de que la gente cree que existe una diferencia real. Me parece bastante inverosímil pensar que los lugareños no se darían cuenta de lo que se estaba intentando hacer si un grupo de investigadores apareciera en Port Pirie con batas de laboratorio y tests de CI, y aún menos verosímil pensar que mucha gente estaría bastante resentida contigo por hacerlo. Esas personas no cooperarán tanto en las pruebas. Otras personas de Port Pirie podrían estar más motivadas para hacerlo bien porque no quieren que su ciudad natal quede mal. Es probable que los efectos motivacionales que se aplicarían en Whyalla sean más débiles, porque la gente no tiene el mismo concepto de "intoxicación por mineral de hierro" que de "intoxicación por plomo". La psicología es difícil.

8.4.1 Estimación de la media poblacional

Supongamos que vamos a Port Pirie y 100 de los lugareños tienen la amabilidad de someterse a un test de CI. La puntuación media de CI entre estas personas resulta ser $\bar{X} = 98.5$. Entonces, ¿cuál es el verdadero CI medio de toda la población de Port Pirie? Obviamente, no sabemos la respuesta a esa pregunta. Podría ser 97,2, pero también podría ser 103,5. Nuestro muestreo no es exhaustivo, así que no podemos dar una respuesta definitiva. No obstante, si me obligaran a punta de pistola a dar una "mejor estimación", tendría que decir que es 98.5. Esa es la esencia de la estimación estadística: dar la mejor estimación posible.

En este ejemplo, la estimación del parámetro de población desconocido es sencilla. Calculo la media muestral y la utilizo como mi **estimación de la media poblacional**. Es bastante sencillo, y en la siguiente sección explicaré la justificación estadística de esta respuesta intuitiva. Sin embargo, por el momento lo que quiero hacer es asegurarme de que reconoces que el estadístico muestral y la estimación del parámetro poblacional son cosas conceptualmente diferentes. Un estadístico muestral es una descripción de tus datos, mientras que la estimación es una conjetura sobre la población. Teniendo esto en cuenta, los estadísticos suelen utilizar diferentes notaciones para referirse a ellos. Por ejemplo, si la media poblacional verdadera se denota μ, entonces usaríamos $\hat{\mu}$ para referirnos a nuestra estimación de la media poblacional. En cambio, la media muestral se denota $\bar{X}$ o, a veces, m. Sin embargo, en muestras aleatorias simples la estimación de la media poblacional es idéntica a la media muestral. Si observo una media muestral de $\bar{X} = 98,5$, entonces mi estimación de la media poblacional también es $\hat{\mu} = 98,5$. Para ayudar a mantener clara la notación, aquí hay una tabla práctica (Table 8.2):

Table 8.2: Notación para la media

Symbol	What is it?	Do we know what it is?
$\hat{X}$	Sample mean	Yes, calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes, identical to the sample mean in simple random samples

8.4.2 Estimación de la desviación estándar de la población

Hasta ahora, la estimación parece bastante sencilla, y puede que te preguntes por qué te he obligado a leer todo eso sobre la teoría del muestreo. En el caso de la media, nuestra estimación del parámetro poblacional (es decir, $\hat{\mu}$) resultó ser idéntica al estadístico muestral correspondiente (es decir, $\bar{X}$). Sin embargo, esto no siempre es cierto. Para verlo, pensemos en cómo construir una **estimación de la desviación estándar poblacional**, que denotaremos $\hat{\sigma}$. ¿Qué utilizaremos como estimación en este caso? Tu primer pensamiento podría ser que podríamos hacer lo mismo que hicimos al estimar la media, y sólo tienes que utilizar el estadístico muestral como nuestra estimación. Eso es casi lo correcto, pero no del todo.

He aquí por qué. Supongamos que tengo una muestra que contiene una única observación. Para este ejemplo, es útil considerar una muestra en la que no tengas ninguna intuición sobre cuáles podrían ser los valores reales de la población, así que usemos algo completamente ficticio. Supongamos que la observación en cuestión mide la cromulencia de mis zapatos. Resulta que mis zapatos tienen una cromulencia de 20. Así que aquí está mi muestra:

Se trata de una muestra perfectamente legítima, aunque tenga un tamaño muestral de $N = 1$. Tiene una media muestral de 20 y dado que cada observación de esta muestra es igual a la media muestral (¡obviamente!) tiene una desviación estándar muestral de 0. Como descripción de la *muestra* parece bastante correcta, la muestra contiene una única observación y, por tanto, no se observa ninguna variación dentro de la muestra. Una desviación estándar muestral de $s = 0$ es la respuesta correcta en este caso. Pero como estimación de la desviación estándar de la *población* parece una completa locura, ¿verdad? Es cierto que tú y yo no sabemos nada en absoluto sobre lo que es la "cromulencia", pero sabemos algo sobre datos. La única razón por la que no vemos ninguna variabilidad en la *muestra* es que la muestra es demasiado pequeña para mostrar ninguna variación. Por lo tanto, si tenemos un tamaño de muestra de $N = 1$, parece que la respuesta correcta es simplemente decir "ni idea".

Observa que *no* tienes la misma intuición cuando se trata de la media muestral y la media poblacional. Si nos vemos obligadas a hacer una suposición sobre la media de la población, no nos parecerá una locura adivinar que la media de la población es 20. Claro, probablemente no te sentirías muy segura de esa suposición porque solo tienes una observación con la que trabajar, pero sigue siendo la mejor suposición que puedes hacer.

Ampliemos un poco este ejemplo. Supongamos que ahora hago una segunda observación. Mi conjunto de datos ahora tiene $N = 2$ observaciones de la cromulencia de los zapatos, y la muestra completa tiene ahora este aspecto:

$$20, 22$$

Esta vez, nuestra muestra es lo suficientemente grande como para que podamos observar cierta variabilidad: ¡dos observaciones es el número mínimo necesario para observar cualquier variabilidad! Para nuestro nuevo conjunto de datos, la media muestral es $\bar{X} = 21$ y la desviación estándar muestral es $s = 1$. ¿Qué intuiciones tenemos sobre la población? Una vez más, en lo que respecta a la media poblacional, la mejor suposición que podemos hacer es la media muestral. Si tuviéramos que adivinar, probablemente diríamos que la cromulencia media de la población es de 21. ¿Qué pasa con la desviación estándar? Esto es un poco más complicado. La desviación estándar de la muestra solo se basa en dos observaciones, y si eres como yo, probablemente tengas la intuición de que, con solo dos observaciones, no le hemos dado a la población "suficientes oportunidades" para que nos revele su verdadera variabilidad. a nosotros. No es solo que sospechemos que la estimación es incorrecta, después de todo, con solo dos observaciones esperamos que lo sea en cierta medida. Lo que nos preocupa es que el error sea sistemático. En concreto, sospechamos que es probable que la desviación estándar de la muestra sea menor que la desviación estándar de la población.

Esta intuición parece correcta, pero estaría bien poder demostrarlo de alguna manera. De hecho, existen pruebas matemáticas que confirman esta intuición, pero a menos que

se tengan los conocimientos matemáticos adecuados, no sirven de mucho. En cambio, lo que haré será simular los resultados de algunos experimentos. Con eso en mente, volvamos a nuestros estudios sobre el CI. Supongamos que la media real de CI de la población es de 100 y la desviación estándar es de 15. Primero realizaré un experimento en el que mediré $N = 2$ puntuaciones de CI y calcularé la desviación estándar de la muestra. Si hago esto una y otra vez y trazo un histograma de estas desviaciones estándar muestrales, lo que tengo es la distribución muestral de la desviación estándar. He representado esta distribución en Figure 8.10. Aunque la verdadera desviación estándar de la población es 15, la media de las desviaciones estándar muestrales es solo 8.5. Observa que este es un resultado muy diferente al que encontramos en Figure 8.8 (b) cuando trazamos la distribución muestral de la media, donde la media poblacional es 100 y la media de las medias muestrales también es 100 .

```
[1] 8.498853
```

Ahora ampliemos la simulación. En lugar de limitarnos a la situación en la que $N = 2$, repitamos el ejercicio para tamaños de muestra de 1 a 10. Si representamos gráficamente la media muestral promedio y la desviación estándar muestral promedio en función del tamaño de la muestra, obtendremos los resultados que se muestran en Figure 8.11. En el lado izquierdo (panel (a)) he representado la media muestral promedio y en el lado derecho (panel (b)) he representado la desviación estándar promedio. Los dos gráficos son bastante diferentes: en promedio, la media muestral promedio es igual a la media poblacional. Es un **estimador insesgado**, que es esencialmente la razón por la que la mejor estimación de la media poblacional es la media muestral.[6] El gráfico de la derecha es bastante diferente: en promedio, la desviación estándar muestral s es menor que la desviación estándar poblacional σ. Se trata de un **estimador sesgado**. En otras palabras, si queremos hacer una "mejor estimación" $\hat{\sigma}$ sobre el valor de la desviación estándar poblacional $\hat{\sigma}$ debemos asegurarnos de que nuestra estimación es un poco mayor que la desviación estándar muestral s.

La solución a este sesgo sistemático es muy sencilla. Así es como funciona. Antes de abordar la desviación estándar, veamos la varianza. Si recuerdas la sección sobre [Estimación de los parámetros de la población], la varianza de la muestra se define como la media de las desviaciones al cuadrado de la media de la muestra. Es decir:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

La varianza muestral s^2 es un estimador sesgado de la varianza poblacional σ^2. Pero resulta que solo tenemos que hacer un pequeño ajuste para transformar esto en un estimador insesgado. Todo lo que tenemos que hacer es dividir por $N - 1$ en lugar de por N.

Se trata de un estimador insesgado de la varianza poblacional σ. Además, esto responde finalmente a la pregunta que planteamos en [Estimación de los parámetros de la población]. ¿Por qué jamovi nos da respuestas ligeramente diferentes para la varianza? Es porque jamovi calcula $\hat{\sigma}^2$ no s^2, por eso. Una historia similar se aplica a la desviación estándar. Si dividimos entre $N - 1$ en lugar de N, nuestra estimación

[6] Debo señalar que estoy ocultando algo aquí. La insesgadez es una característica deseable para un estimador, pero hay otras cosas que importan además del sesgo. Sin embargo, está fuera del alcance de este libro discutir esto en detalle. Solo quiero llamar tu atención sobre el hecho de que hay una cierta complejidad oculta aquí.

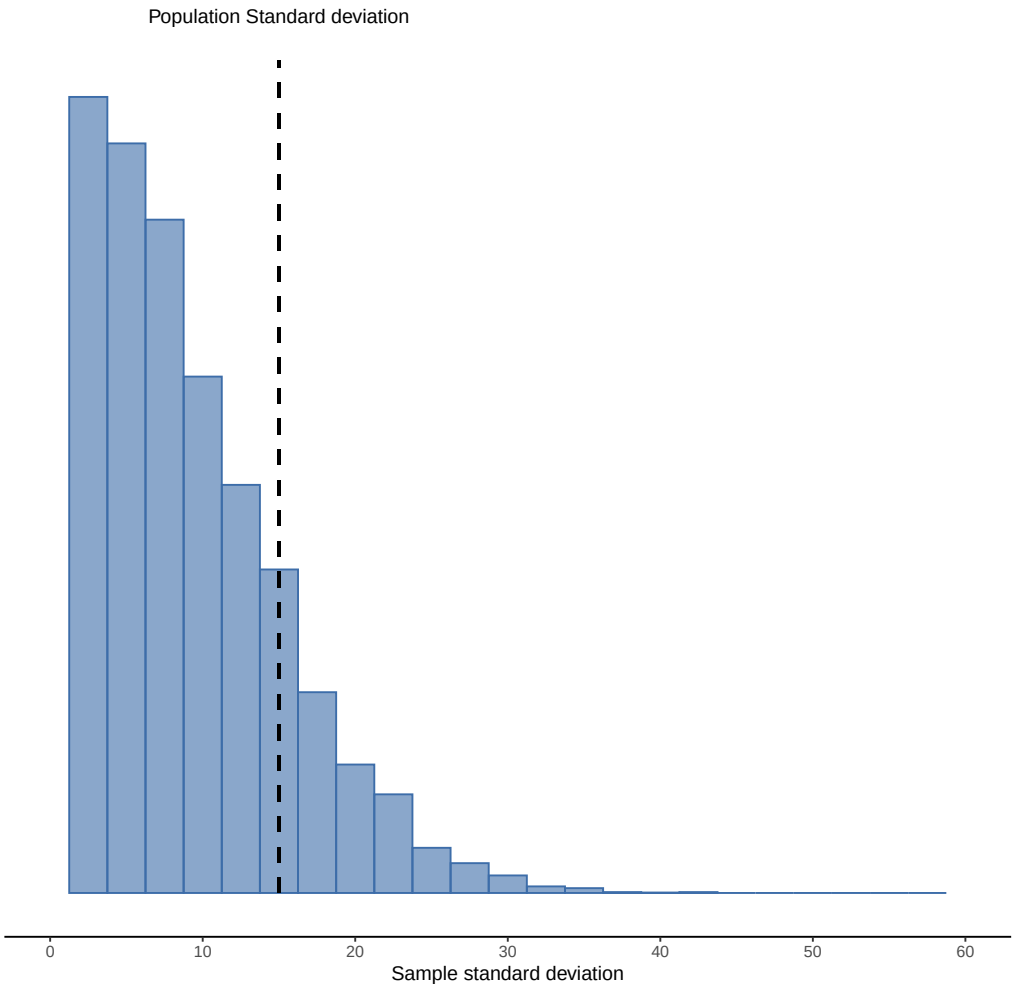


Figure 8.10: Distribución muestral de la desviación estándar muestral para un experimento de 'dos puntuaciones de CI'. La verdadera desviación estándar de la población es 15 (línea discontinua), pero como se puede ver en el histograma, la gran mayoría de los experimentos producirán una desviación estándar muestral mucho menor que esta. En promedio, este experimento produciría una desviación estándar estimada de solo 8,4, ¡muy por debajo del valor real! En otras palabras, la desviación estándar de la muestra es una estimación sesgada de la desviación estándar de la población.

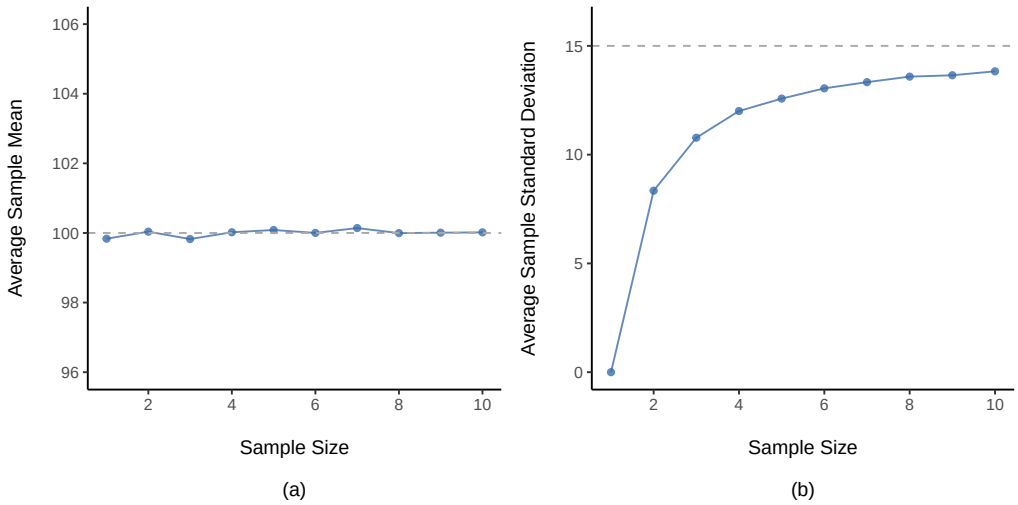


Figure 8.11: Ilustración del hecho de que la media muestral es un estimador insesgado de la media poblacional (panel a), pero la desviación estándar muestral es un estimador sesgado de la desviación estándar poblacional (panel b). Para la figura, generé conjuntos de datos simulados de $ 10,000 $ con 1 observación cada uno, $ 10,000 $ más con 2 observaciones, y así sucesivamente hasta un tamaño de muestra de 10. Cada conjunto de datos estaba formado por datos de CI falsos, es decir, los datos se distribuían normalmente con una media poblacional real de 100 y una desviación estándar de 15. En promedio, las medias muestrales resultan ser 100, independientemente del tamaño de la muestra (panel a). Sin embargo, las desviaciones estándar de la muestra resultan ser sistemáticamente demasiado pequeñas (panel b), sobre todo para tamaños de muestra pequeños.

de la desviación estándar de la población es insesgada, y cuando usamos la función de desviación estándar de jamovi, lo que está haciendo es calcular $\hat{\sigma}$ no $ s$.[7]

Un último punto. En la práctica, mucha gente tiende a referirse a $\hat{\sigma}$ (es decir, la fórmula en la que dividimos por $N - 1$) como la desviación estándar de la muestra. Técnicamente, esto es incorrecto. La desviación estándar de la muestra debería ser igual a s (es decir, la fórmula en la que dividimos por N). No son lo mismo, ni conceptual ni numéricamente. Una es una propiedad de la muestra, la otra es una característica estimada de la población. Sin embargo, en casi todas las aplicaciones de la vida real, lo que realmente nos importa es la estimación del parámetro de la población, por lo que la gente siempre informa $\hat{\sigma}$ en lugar de s. Este es el número correcto para informar, por supuesto. Es solo que la gente tiende a ser un poco imprecisa con la terminología cuando lo escriben, porque "desviación estándar de la muestra" es más corta que "desviación estándar estimada de la población". No es gran cosa, y en la práctica hago lo mismo que todo el mundo. Sin embargo, creo que es importante mantener los dos conceptos separados. Nunca es buena idea confundir "propiedades conocidas de la muestra" con "suposiciones sobre la población de la que procede". En el momento en que empiezas a pensar que s y $\hat{\sigma}$ son lo mismo, empiezas a hacer exactamente eso.

Para terminar esta sección, aquí hay otro par de tablas para ayudar a mantener las cosas claras (Table 8.3 y Table 8.4).

Table 8.3: Notación para la desviación estándar

Symbol	What is it?	Do we know what it is?
s	Sample standard deviation	Yes, calculated from the raw data
σ	Population standard deviation	Almost never known for sure
$\hat{\sigma}$	Estimate of the population standard deviation	Yes, but not the same as the sample standard deviation

[7] Dividiendo por $N - 1$ obtenemos una estimación insesgada de la varianza poblacional:

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

, y lo mismo para la desviación estándar:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

. Bien, estoy ocultando algo más aquí. En un giro extraño y contraintuitivo, dado que $\hat{\sigma}^2$ es un estimador insesgado de σ^2, se supondría que sacar la raíz cuadrada estaría bien y $\hat{sigma}$ sería un estimador insesgado de . ¿Cierto? Extrañamente, no lo es. En realidad, hay un sesgo sutil, pequeño en $\hat{\sigma}$. Esto es simplemente extraño: $\hat{\sigma}^2$ es una estimación insesgada de la varianza poblacional σ^2 , pero cuando sacas la raíz cuadrada, resulta que $\hat{\sigma}$ es un estimador sesgado de la desviación estándar de la población . Raro, raro, raro, ¿verdad? Entonces, ¿por qué es $\hat{\sigma}$ sesgada? La respuesta técnica es "porque las transformaciones no lineales (por ejemplo, la raíz cuadrada) no conmutan con la expectativa", pero eso suena como un galimatías para todos los que no han hecho un curso de estadística matemática. Afortunadamente, a efectos prácticos no importa. El sesgo es pequeño, y en la vida real todo el mundo utiliza $\hat{\sigma}$ y funciona bien. A veces las matemáticas son simplemente molestas.

Table 8.4: Notación para la varianza

Symbol	What is it?	Do we know what it is?
s^2	Sample variance	Yes, calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

8.5 Estimación de un intervalo de confianza

Estadística significa nunca tener que decir que estás seguro
– Origen desconocido [8]

Hasta este punto del capítulo, he descrito los fundamentos de la teoría del muestreo en la que se basan los estadísticos para hacer conjeturas sobre los parámetros de la población a partir de una muestra de datos. Como ilustra esta discusión, una de las razones por las que necesitamos toda esta teoría del muestreo es que cada conjunto de datos nos deja algo de incertidumbre, por lo que nuestras estimaciones nunca van a ser perfectamente precisas. Lo que ha faltado en este debate es un intento de cuantificar la cantidad de incertidumbre que acompaña a nuestra estimación. No basta con adivinar que, por ejemplo, el CI medio de los estudiantes universitarios de psicología es de 115 dólares (sí, me acabo de inventar esa cifra). También queremos poder decir algo que exprese el grado de certeza que tenemos en nuestra conjetura. Por ejemplo, estaría bien poder decir que hay un 95% de probabilidades de que la verdadera media se encuentre entre 109 y 121. El nombre para esto es un **intervalo de confianza** para la media.

Si se conocen las distribuciones muestrales, construir un intervalo de confianza para la media es bastante fácil. Así es como funciona. Supongamos que la media real de la población es μ y la desviación estándar es σ. Acabo de terminar mi estudio que tiene N participantes, y la media del CI entre esos participantes es $\bar{X}$. Sabemos por nuestra discusión de El teorema central del límite que la distribución muestral de la media es aproximadamente normal. También sabemos por nuestra discusión sobre la distribución normal en Section 7.5 que existe una probabilidad de 95% de que una cantidad distribuida normalmente caiga dentro de aproximadamente dos desviaciones estándar de la media real.

Para ser más precisos, la respuesta más correcta es que existe una probabilidad del 95% de que una cantidad distribuida normalmente caiga dentro de $1,96$ desviaciones estándar de la media real. A continuación, recordemos que la desviación estándar de la distribución muestral se denomina error estándar y el error estándar de la media se escribe como SEM. Cuando juntamos todas estas piezas, aprendemos que existe una probabilidad del 95% de que la media muestral $\bar{X}$ que hemos observado realmente se encuentre dentro de $1,96$ errores estándar de la media poblacional.

[8] Esta cita aparece en muchas camisetas y sitios web, e incluso se menciona en algunos artículos académicos (p. ej., http://www.amstat.org/publications/jse/v10n3/friedman.html , pero nunca encontré la fuente original.

Por supuesto, el número 1.96 no tiene nada de especial. Resulta que es el multiplicador que hay que utilizar si se desea un intervalo de confianza de 95%. Si hubieras querido un intervalo de confianza de 70%, habrías usado $1,04$ como número mágico en lugar de $1,96$.

[Detalle técnico adicional [9]]

Por supuesto, no hay nada especial en el número 1,96. Resulta que es el multiplicador que necesita usar si desea un intervalo de confianza del 95%. Si hubiera querido un intervalo de confianza del 70 %, habría utilizado 1,04 como número mágico en lugar de 1,96.

8.5.1 Interpretar un intervalo de confianza

Lo más difícil de los intervalos de confianza es entender lo que significan. Cuando la gente se encuentra por primera vez con intervalos de confianza, el primer instinto casi siempre es decir que "hay un 95% de probabilidad de que la media real se encuentre dentro del intervalo de confianza". Es sencillo y parece captar la idea de sentido común de lo que significa decir que tengo un "95% de confianza". Por desgracia, no es del todo correcto. La definición intuitiva depende en gran medida de las creencias personales sobre el valor de la media de la población. Digo que tengo un 95% de confianza porque esas son mis creencias. En la vida cotidiana eso está perfectamente bien, pero si recuerdas la sección ¿Qué significa probabilidad?, te darás cuenta de que hablar de creencias personales y confianza es una idea bayesiana. Sin embargo, los intervalos de confianza no son herramientas bayesianas. Como todo lo demás en este capítulo, los intervalos de confianza son herramientas frecuentistas, y si vas a utilizar métodos frecuentistas, entonces no es apropiado darles una interpretación bayesiana. Si utilizas métodos frecuentistas, debes adoptar interpretaciones frecuentistas. Bien, si esa no es la respuesta correcta, ¿cuál es? Recuerda lo que dijimos sobre la probabilidad frecuentista. La única forma que tenemos de hacer "afirmaciones de probabilidad" es hablar de una secuencia de sucesos y contar las frecuencias de los diferentes sucesos. Desde esa perspectiva, la interpretación de un intervalo de confianza del 95% debe tener algo que ver con la replicación. En concreto, si replicamos el experimento una y otra vez y calculamos un intervalo de confianza del 95% para cada repetición, entonces el 95% de esos intervalos

[9] Matemáticamente, escribimos esto como:

$$\mu - (1.96 \times SEM) \leq \bar{X} \leq \mu + (1.96 \times SEM)$$

donde el SEM es igual a $\frac{\sigma}{\sqrt{N}}$ N y podemos estar seguras al 95% de que esto es cierto. Sin embargo, eso no responde a la pregunta que realmente nos interesa. La ecuación anterior nos dice lo que debemos esperar sobre la media muestral dado que sabemos cuáles son los parámetros de la población. Lo que queremos es que funcione al revés. Queremos saber qué debemos creer sobre los parámetros poblacionales, dado que hemos observado una muestra concreta. Sin embargo, no es demasiado difícil hacer esto. Usando un poco de álgebra de secundaria, una forma astuta de reescribir nuestra ecuación es la siguiente:

$$\bar{X} - (1.96 \times SEM) \leq \mu \leq \bar{X} + (1.96 \times SEM)$$

Lo que esto nos dice es que el rango de valores tiene una probabilidad del 95% de contener la media poblacional μ. Nos referimos a este rango como un **intervalo de confianza del 95 %**, denominado CI_{95}. En resumen, siempre que N sea lo suficientemente grande (lo suficientemente grande par que creamos que la distribución muestral de la media es normal), entonces podemos escribir esto como nuestra fórmula para el intervalo de confianza del 95%:

$$CI_{95} = \bar{X} \pm (1,96 \times \frac{\sigma}{\sqrt{N}})$$

contendrían la media verdadera. De manera más general, el 95% de todos los intervalos de confianza construidos mediante este procedimiento deberían contener la media poblacional verdadera. Esta idea se ilustra en Figure 8.12, que muestra 50 intervalos de confianza construidos para un experimento de "medida de 10 puntuaciones de CI" (panel superior) y otros 50 intervalos de confianza para un experimento de "medida de 25 puntuaciones de CI" (panel inferior). Esperaríamos que alrededor de 95 de nuestros intervalos de confianza contuvieran la verdadera media poblacional, y eso es lo que encontramos en Figure 8.12. La diferencia fundamental aquí es que la afirmación bayesiana hace una declaración de probabilidad sobre la media de la población (es decir, se refiere a nuestra incertidumbre sobre la media de la población), lo que no está permitido según la interpretación frecuentista de la probabilidad porque no se puede "replicar" una población. En la afirmación frecuentista, la media poblacional es fija y no se pueden hacer afirmaciones probabilísticas sobre ella. Sin embargo, los intervalos de confianza son repetibles, por lo que podemos replicar experimentos. Por lo tanto, un *frecuentista* puede hablar de la probabilidad de que el *intervalo de confianza* (una variable aleatoria) contenga la media real, pero no puede hablar de la probabilidad de que la *media poblacional real* (no es un suceso repetible) se encuentre dentro del intervalo de confianza. Sé que esto parece un poco pedante, pero es importante. Importa porque la diferencia en la interpretación conduce a una diferencia en las matemáticas. Existe una alternativa bayesiana a los intervalos de confianza, conocida como *intervalos creíbles*. En la mayoría de las situaciones, los intervalos creíbles son bastante similares a los intervalos de confianza, pero en otros casos son drásticamente diferentes. Sin embargo, como prometí, hablaré más sobre la perspectiva bayesiana en Chapter 16.

8.5.2 Cálculo de intervalos de confianza en jamovi

jamovi incluye una forma sencilla de calcular los intervalos de confianza para la media como parte de la funcionalidad 'Descriptivos'. En 'Descriptivos'-'Estadísticas' hay una casilla de verificación tanto para 'Error estándar de la media' como para el 'Intervalo de confianza para la media', por lo que puedes usar esto para averiguar el intervalo de confianza del 95% (que es el valor predeterminado). Así, por ejemplo, si cargo el archivo IQsim.omv, marco la casilla 'Intervalo de confianza para la media', puedo ver el intervalo de confianza asociado con el CI medio simulado: IC del 95 % inferior = 99,39 y IC del 95 % superior = 99,97. Así, en nuestros datos de muestra grande con N = 10 000, la puntuación media del CI es 99,68 con un IC del 95 % de 99,39 a 99,97.

Cuando se trata de trazar intervalos de confianza en jamovi, puede especificar que la media se incluya como opción en un diagrama de caja. Además, cuando aprendamos sobre pruebas estadísticas específicas, por ejemplo, en Chapter 13, veremos que también podemos trazar intervalos de confianza como parte del análisis de datos. Eso está muy bien, así que te mostraremos cómo hacerlo más adelante.

8.6 Resumen

En este capítulo hemos tratado dos temas principales. La primera mitad del capítulo trata sobre la teoría del muestreo, y la segunda mitad trata sobre cómo podemos usar la teoría del muestreo para construir estimaciones de los parámetros de la población. El desglose de las secciones es el siguiente:

- Ideas básicas sobre Muestras, poblaciones y muestreo

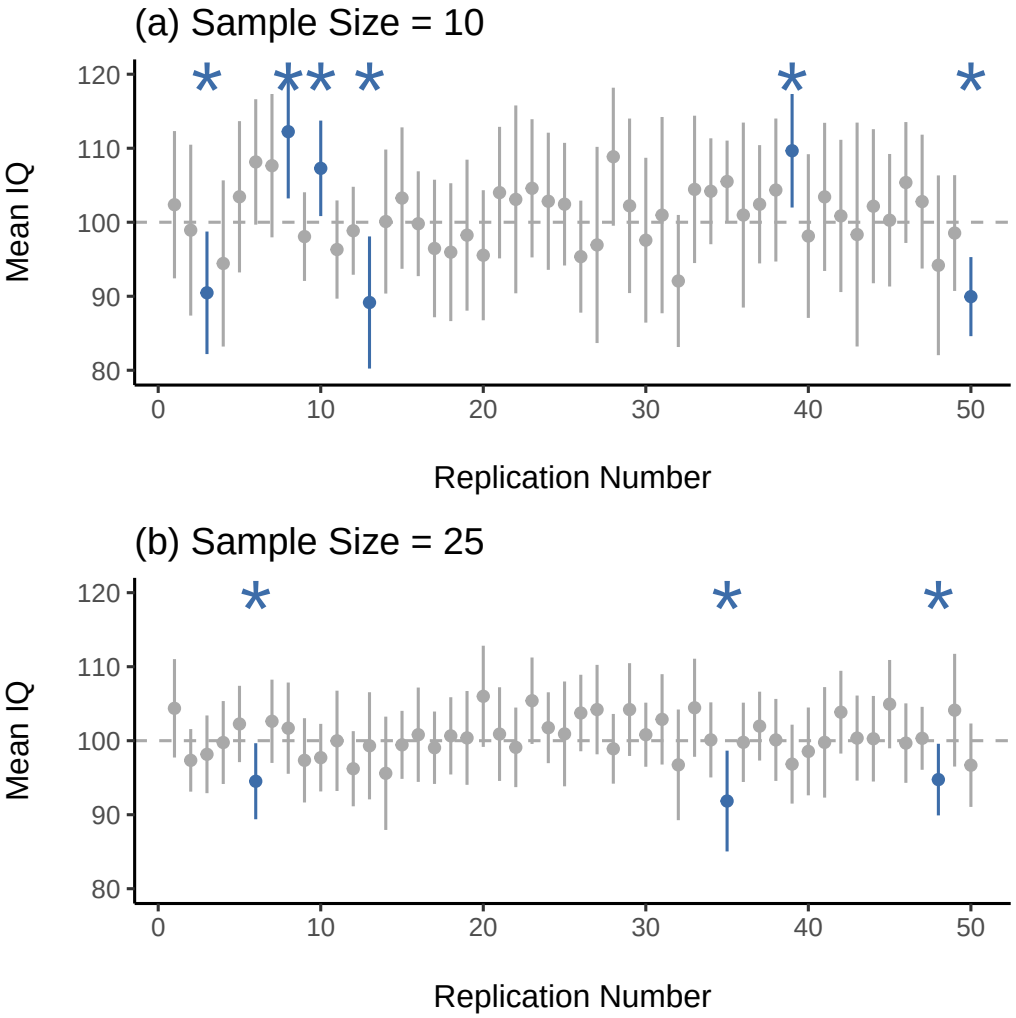


Figure 8.12: Intervalos de confianza del 95%. El panel superior (a) muestra 50 réplicas simuladas de un experimento en el que medimos el CI de 10 personas. El punto marca la posición de la media muestral y la línea muestra el intervalo de confianza del 95%. La mayoría de los 50 intervalos de confianza contienen media real (es decir, 100), pero algunos, en azul y marcados con asteriscos, no la contienen. El gráfico inferior (panel b) muestra una simulación similar, pero esta vez simulamos réplicas de un experimento que mide el CI de 25 personas.

- Teoría estadística del muestreo: La ley de los grandes números y Distribuciones muestrales y el teorema central del límite
- [Estimación de parámetros poblacionales]. Medias y desviaciones estándar
- Estimación de un intervalo de confianza

Como siempre, hay muchos temas relacionados con el muestreo y la estimación que no se tratan en este capítulo, pero creo que para una clase de introducción a la psicología es bastante completo. Para la mayoría de los investigadores aplicados, no necesitará mucha más teoría que esta. Una cuestión importante que no he tocado en este capítulo es qué hacer cuando no se dispone de una muestra aleatoria simple. Hay mucha teoría estadística a la que se puede recurrir para manejar esta situación, pero va mucho más allá del alcance de este libro.

Chapter 9

Prueba de hipótesis

> El proceso de inducción consiste en asumir la ley más simple que se pueda armonizar con nuestra experiencia. Este proceso, sin embargo, no tiene fundamento lógico sino sólo psicológico. Está claro que no hay motivos para creer que el curso más simple de los acontecimientos vaya a suceder realmente. Es una hipótesis que el sol saldrá mañana, lo que significa que no sabemos si saldrá. – Ludwig Wittgenstein [1]

En el último capítulo discutí las ideas en las que se basa la estimación, que es una de las dos "grandes ideas" de la estadística inferencial. Ahora es el momento de centrar nuestra atención en la otra gran idea, que es la *prueba de hipótesis*. En su forma más abstracta, la prueba de hipótesis es realmente una idea muy simple. El investigador tiene una teoría sobre el mundo y quiere determinar si los datos apoyan o no esa teoría. Sin embargo, los detalles son complicados y la mayoría de la gente considera que la teoría de la prueba de hipótesis es la parte más frustrante de la estadística. La estructura del capítulo es la siguiente. En primer lugar, describiré cómo funcionan las prueba de hipótesis con bastante detalle, utilizando un ejemplo sencillo para mostrar cómo se "construye" una prueba de hipótesis. Intentaré no ser demasiado dogmática centrarme en la lógica subyacente del procedimiento de prueba.[2] Luego, dedicaré un poco de tiempo a hablar sobre los diversos dogmas, reglas y herejías que rodean la teoría de la prueba de hipótesis.

9.1 Una colección de hipótesis

Al final todos sucumbimos a la locura. Para mí, ese día llegará cuando por fin me asciendan a catedrática. Instalada en mi torre de marfil, felizmente protegida por la cátedra, podré por fin despedirme de mis sentidos (por así decirlo) y dedicarme a esa línea de

[1] La cita proviene del texto de Wittgenstein (1922), *Tractatus Logico-Philosphicus.*

[2] Nota técnica. La descripción que sigue difiere sutilmente de la descripción estándar que se da en muchos textos introductorios. La teoría ortodoxa de la prueba de hipótesis nula surgió del trabajo de Sir Ronald Fisher y Jerzy Neyman a principios del siglo XX; pero Fisher y Neymar en realidad tenían puntos de vista muy diferentes sobre cómo debería funcionar. El tratamiento estándar de las pruebas de hipótesis que utilizan la mayoría de los textos es un híbrido de los dos enfoques. El tratamiento aquí es un poco más al estilo de Neyman que la visión ortodoxa, especialmente en lo que respecta al significado del valor p.

investigación psicológica más improductiva, la búsqueda de la percepción extrasensorial (PES).[3]

Supongamos que este glorioso día ha llegado. Mi primer estudio es sencillo y consiste en probar si existe la clarividencia. Cada participante se sienta en una mesa y un experimentador le muestra una tarjeta. La tarjeta es negra por un lado y blanca por el otro. El experimentador retira la tarjeta y la coloca sobre una mesa en una habitación contigua. La tarjeta se coloca con el lado negro hacia arriba o el lado blanco hacia arriba de forma totalmente aleatoria, y la aleatorización se produce después de que el experimentador haya salido de la habitación con el participante. Entra un segundo experimentador y le pregunta al participante qué lado de la tarjeta está ahora hacia arriba. Es un experimento de una sola vez. Cada persona ve solo una tarjeta y da solo una respuesta, y en ningún momento el participante está en contacto con alguien que sepa la respuesta correcta. Mi conjunto de datos, por lo tanto, es muy sencillo. He hecho la pregunta a N personas y un número X de ellas ha dado la respuesta correcta. Para concretar, supongamos que he hecho la prueba a $N = 100$ personas y $X = 62$ de ellas han dado la respuesta correcta. Un número sorprendentemente grande, sin duda, pero ¿es lo suficientemente grande como para que pueda afirmar que he encontrado pruebas de la PES? Esta es la situación en la que las pruebas de hipótesis resultan útiles. Sin embargo, antes de hablar de cómo probar hipótesis, debemos tener claro qué entendemos por hipótesis.

9.1.1 Hipótesis de investigación versus hipótesis estadísticas

La primera distinción que debes tener clara es entre hipótesis de investigación e hipótesis estadísticas. En mi estudio sobre la PES, mi objetivo científico general es demostrar que existe la clarividencia. En esta situación, tengo un objetivo de investigación claro: espero descubrir pruebas de la PES. En otras situaciones, podría ser mucho más neutral que eso, por lo que podría decir que mi objetivo de investigación es determinar si existe o no la clarividencia. Independientemente de cómo me presente, lo que quiero decir es que una hipótesis de investigación implica hacer una afirmación científica sustantiva y comprobable. Si eres psicóloga, tus hipótesis de investigación se refieren fundamentalmente a constructos psicológicos. Cualquiera de las siguientes contaría como **hipótesis de investigación**:

- *Escuchar música reduce la capacidad de prestar atención a otras cosas.* Se trata de una afirmación sobre la relación causal entre dos conceptos psicológicamente significativos (escuchar música y prestar atención a las cosas), por lo que es una hipótesis de investigación perfectamente razonable.
- *La inteligencia está relacionada con la personalidad.* Al igual que la anterior, se trata de una afirmación relacional sobre dos constructos psicológicos (inteligencia y personalidad), pero la afirmación es más débil: correlacional, no causal.
- *La inteligencia es la velocidad de procesamiento de la información.* Esta hipótesis tiene un carácter bastante diferente. En realidad, no es una afirmación relacional en absoluto. Es una afirmación ontológica sobre el carácter fundamental de la

[3] Mis disculpas a cualquiera que realmente crea en estas cosas, pero según mi lectura de la literatura sobre PES no es razonable pensar que esto sea real. Sin embargo, para ser justos, algunos de los estudios están rigurosamente diseñados, por lo que en realidad es un área interesante para pensar sobre el diseño de la investigación psicológica. Y, por supuesto, es un país libre, así que puedes dedicar tu tiempo y esfuerzo a demostrar que me equivoco si quieres, pero no creo que sea un uso muy práctico de tu intelecto.

inteligencia (y estoy bastante segura de que en realidad es ésta). Normalmente es más fácil pensar en cómo construir experimentos para probar hipótesis de investigación del tipo "¿afecta X a Y?" que abordar afirmaciones como "¿qué es X?" Y en la práctica, lo que suele ocurrir es que se encuentran formas de probar las afirmaciones relacionales que se derivan de las ontológicas. Por ejemplo, si creo que la inteligencia es la velocidad de procesamiento de la información en el cerebro, mis experimentos consistirán a menudo en buscar relaciones entre medidas de inteligencia y medidas de velocidad. En consecuencia, la mayoría de las preguntas de investigación cotidianas tienden a ser de naturaleza relacional, pero casi siempre están motivadas por preguntas ontológicas más profundas sobre el estado de naturaleza.

Ten en cuenta que en la práctica, mis hipótesis de investigación podrían solaparse mucho. Mi objetivo final en el experimento de PES podría ser probar una afirmación ontológica como "la PES existe", pero podría restringirme operativamente a una hipótesis más limitada como "algunas personas pueden 'ver' objetos de manera clarividente". Dicho esto, hay algunas cosas que realmente no cuentan como hipótesis de investigación adecuadas en ningún sentido significativo:

- *El amor es un campo de batalla.* Esto es demasiado vago para ser comprobable. Aunque está bien que una hipótesis de investigación tenga cierto grado de vaguedad, tiene que ser posible operacionalizar tus ideas teóricas. Tal vez no soy lo bastante creativa para verlo, pero no veo cómo se puede convertir esto en un diseño de investigación concreto. Si eso es cierto, entonces esta no es una hipótesis de investigación científica, es una canción pop. Eso no significa que no sea interesante. Muchas preguntas profundas que se hacen los humanos entran en esta categoría. Quizá algún día la ciencia sea capaz de construir teorías comprobables sobre el amor, o comprobar si Dios existe, etcétera. Pero ahora mismo no podemos, y yo no apostaría por ver nunca una aproximación científica satisfactoria a ninguna de las dos cosas.
- *La primera regla del club de la tautología es la primera regla del club de la tautología.* No es una afirmación sustantiva de ningún tipo. Es cierta por definición. Ningún estado de la naturaleza concebible podría ser incompatible con esta afirmación. Decimos que se trata de una hipótesis infalsable, y como tal está fuera del dominio de la ciencia. Independientemente de lo que se haga en ciencia, tus afirmaciones deben tener la posibilidad de ser erróneas.
- *En mi experimento más gente dirá "sí" que "no".* Esto falla como hipótesis de investigación porque es una afirmación sobre el conjunto de datos, no sobre la psicología (a menos, por supuesto, que tu pregunta de investigación real sea si las personas tienen algún tipo de sesgo hacia el "sí"). En realidad, esta hipótesis empieza a parecer más una hipótesis estadística que una hipótesis de investigación.

Como puedes ver, las hipótesis de investigación pueden ser algo complicadas a veces y, en última instancia, son afirmaciones científicas. Las **hipótesis estadísticas** no son ninguna de estas dos cosas. Las hipótesis estadísticas deben ser precisas y deben corresponder a afirmaciones concretas sobre las características del mecanismo de generación de datos (es decir, la "población"). Aun así, la intención es que las hipótesis estadísticas guarden una relación clara con las hipótesis de investigación sustantivas que te interesan. Por ejemplo, en mi estudio sobre PES, mi hipótesis de investigación es que algunas personas son capaces de ver a través de las paredes o lo que sea. Lo que quiero hacer es "mapear" esto en una afirmación sobre cómo se generaron los datos. Así que

vamos a pensar en lo que sería esa afirmación. La cantidad que me interesa dentro del experimento es $P(correcta)$, la probabilidad verdadera pero desconocida con la que los participantes en mi experimento responden la pregunta correctamente. Usemos la letra griega θ (theta) para referirnos a esta probabilidad. Aquí hay cuatro hipótesis estadísticas diferentes:

- Si la PES no existe y si mi experimento está bien diseñado, entonces mis participantes solo están adivinando. Así que debería esperar que acierten la mitad de las veces, y entonces mi hipótesis estadística es que la verdadera probabilidad de elegir correctamente es $\theta = 0.5$.
- Alternativamente, supongamos que la PES existe y que los participantes pueden ver la tarjeta. Si eso es cierto, la gente obtendrá mejores resultados que el azar y la hipótesis estadística es que $\theta > 0,5$.
- Una tercera posibilidad es que la PES exista, pero los colores estén todos invertidos y la gente no se dé cuenta (vale, es una locura, pero nunca se sabe). Si es así, es de esperar que los resultados sean inferiores al azar. Esto correspondería a una hipótesis estadística de que $\theta < 0.5$.
- Por último, supongamos que la PES existe pero no tengo idea si la gente ve el color correcto o el incorrecto. En ese caso, la única afirmación que podría hacer sobre los datos sería que la probabilidad de acertar la respuesta correcta no es igual a 0,5. Esto corresponde a la hipótesis estadística de que $\theta \neq 0.5$.

Todos estos son ejemplos legítimos de una hipótesis estadística porque son afirmaciones sobre un parámetro de la población y están relacionados de forma significativa con mi experimento.

Lo que esta discusión deja claro, espero, es que cuando se intenta construir una prueba de hipótesis estadística, el investigador tiene que tener en cuenta dos hipótesis muy distintas. En primer lugar, tiene una hipótesis de investigación (una afirmación sobre la psicología), que corresponde a una hipótesis estadística (una afirmación sobre la población que genera los datos). En mi ejemplo de PES, podrían ser las que se muestran en Table 9.1.

Table 9.1: Investigación e hipótesis estadísticas

Dani's research hypothesis:	**"ESP exists"**
Dani's statistical hypothesis:	$\theta \neq 0.5$

Y una cosa clave que hay que reconocer es lo siguiente. Una prueba de hipótesis estadística es una prueba de la hipótesis estadística, no de la hipótesis de investigación. Si el estudio está mal diseñado, se rompe el vínculo entre la hipótesis de investigación y la hipótesis estadística. Para poner un ejemplo tonto, supongamos que mi estudio de PES se realizara en una situación en la que el participante pudiera ver realmente la tarjeta reflejada en una ventana. Si eso sucede, podrías encontrar pruebas muy sólidas de que $\theta \neq 0.5$, pero esto no nos diría nada sobre si "la PES existe".

9.1.2 Hipótesis nulas e hipótesis alternativas

Hasta aquí, todo bien. Tengo una hipótesis de investigación que corresponde a lo que quiero creer sobre el mundo, y puedo mapearla en una hipótesis estadística que corresponde a lo que quiero creer sobre cómo se generaron los datos. Es en este punto

donde las cosas se vuelven contraintuitivas para mucha gente. Porque lo que estoy a punto de hacer es inventar una nueva hipótesis estadística (la hipótesis "nula", H_0) que corresponde exactamente a lo contrario de lo que quiero creer, y luego centrarme exclusivamente en ella casi en detrimento de lo que realmente me interesa (que ahora se llama la hipótesis "alternativa", H1). En nuestro ejemplo de PES, la hipótesis nula es que $\theta = 0.5$, ya que eso es lo que esperaríamos si la PES no existiera. Mi esperanza, por supuesto, es que la PES es totalmente real, y por lo tanto la alternativa a esta hipótesis nula es $\theta \neq 0.5$. En esencia, lo que estamos haciendo aquí es dividir los posibles valores de θ en dos grupos: aquellos valores que realmente espero que no sean ciertos (la nula) y aquellos valores con los que estaría contenta si resultaran ser correctos (la alternativa). Una vez hecho esto, lo importante es reconocer que el objetivo de una prueba de hipótesis no es demostrar que la hipótesis alternativa es (probablemente) cierta. El objetivo es demostrar que la hipótesis nula es (probablemente) falsa. A la mayoría de la gente esto le parece bastante extraño.

según mi experiencia, la mejor manera de pensar en ello es imaginar que una prueba de hipótesis es un juicio penal[4], **el juicio de la hipótesis nula**. La hipótesis nula es el acusado, el investigador es el fiscal y la prueba estadística es el juez. Al igual que en un juicio penal, existe la presunción de inocencia. La hipótesis nula se considera cierta a menos que tú, la investigadora, puedas probar más allá de toda duda razonable que es falsa. Eres libre de diseñar tu experimento como quieras (dentro de lo razonable, obviamente) y tu objetivo al hacerlo es maximizar la probabilidad de que los datos generen una condena por el delito de ser falsos. El truco está en que la prueba estadística establece las reglas del juicio y esas reglas están diseñadas para proteger la hipótesis nula, concretamente para garantizar que, si la hipótesis nula es realmente cierta, las posibilidades de una condena falsa están garantizadas para ser bajas. Esto es muy importante. Después de todo, la hipótesis nula no tiene abogado, y dado que el investigador está intentando desesperadamente demostrar que es falsa, alguien tiene que protegerla.

9.2 Dos tipos de errores

Antes de entrar en detalles sobre cómo se construye una prueba estadística, es útil entender la filosofía que hay detrás. Lo he insinuado al señalar la similitud entre una prueba de hipótesis nula y un juicio penal, pero ahora debo ser explícita. Idealmente, nos gustaría construir nuestra prueba de forma que nunca cometiéramos errores. Por desgracia, dado que el mundo está desordenado, esto nunca es posible. A veces simplemente tienes mala suerte. Por ejemplo, supongamos que lanzamos una moneda 10 veces seguidas y sale cara las 10 veces. Eso parece una prueba muy sólida para llegar a la conclusión de que la moneda está sesgada, pero, por supuesto, hay una probabilidad de 1 entre 1024 de que esto ocurriera incluso si la moneda fuera totalmente justa. En otras palabras, en la vida real siempre tenemos que aceptar que existe la posibilidad de que nos hayamos equivocado. En consecuencia, el objetivo de las pruebas de hipótesis estadística no es eliminar los errores, sino minimizarlos.

Llegados a este punto, debemos ser un poco más precisas sobre lo que entendemos por "errores". En primer lugar, digamos lo obvio. O bien la hipótesis nula es verdadera, o bien es falsa, y nuestra prueba mantendrá la hipótesis nula o la rechazará.[5] Así que,

[4] esta analogía solo funciona si procedes de un sistema jurídico acusatorio como Reino Unido/Estados Unidos/Australia. Según tengo entendido, el sistema inquisitorial francés es bastante diferente.

[5] un inciso sobre el lenguaje que utilizas para hablar sobre la prueba de hipótesis. En primer lugar, hay

como ilustra Table 9.2, después de ejecutar la prueba y hacer nuestra elección, podría haber ocurrido una de cuatro cosas:

Table 9.2: prueba estadística de hipótesis nula (NHST)

	retain H_0	**reject** H_0
H_0 is true	correct decision	error (type I)
H_0 is false	error (type II)	correct decision

Por consiguiente, en realidad hay dos tipos de error. Si rechazamos una hipótesis nula que en realidad es cierta, cometemos un **error de tipo I**. Por otro lado, si mantenemos la hipótesis nula cuando en realidad es falsa, cometemos un **error de tipo II**.

¿Recuerdas que dije que las pruebas estadísticas eran como un juicio penal? Pues lo decía en serio. Un juicio penal requiere que se demuestre "más allá de toda duda razonable" que el acusado lo hizo. Todas las normas probatorias están (al menos en teoría) diseñadas para garantizar que no haya (casi) ninguna probabilidad de condenar injustamente a un acusado inocente. El juicio está diseñado para proteger los derechos de un acusado, como dijo el famoso jurista inglés William Blackstone, es "mejor que escapen diez culpables a que sufra un inocente". En otras palabras, un juicio penal no trata de la misma manera los dos tipos de error. Castigar al inocente se considera mucho peor que dejar libre al culpable. Una prueba estadística es más o menos lo mismo. El principio de diseño más importante de la prueba es controlar la probabilidad de un error de tipo I, para mantenerla por debajo de una probabilidad fija. Esta probabilidad, que se denota α, se llama **nivel de significación** de la prueba. Y lo diré de nuevo, porque es fundamental para todo el montaje: se dice que una prueba de hipótesis tiene un nivel de significación α si la tasa de error tipo I no es mayor que α.

¿Y qué pasa con la tasa de error tipo II? Bueno, también nos gustaría tenerla bajo control, y denotamos esta probabilidad por β. Sin embargo, es mucho más común referirse a la **potencia** de la prueba, que es la probabilidad con la que rechazamos una hipótesis nula cuando realmente es falsa, que es $1 - \beta$. Para que no nos equivoquemos, aquí tenemos de nuevo la misma tabla pero con los números correspondientes añadidos (Table 9.3):

Una prueba de hipótesis "potente" es aquella que tiene un valor pequeño de β, mientras mantiene α fijo en algún nivel (pequeño) deseado. Por convención, los científicos utilizan tres niveles α diferentes: .05, .01 y .001. Fíjate en la asimetría aquí; las pruebas están diseñadas para garantizar que el nivel de α se mantiene bajo, pero no hay ninguna

que evitar la palabra "demostrar". Una prueba estadística realmente no demuestra que una hipótesis sea verdadera o falsa. La prueba implica certeza y, como dice el refrán, la estadística significa nunca tener que decir que estás seguro. En eso casi todo el mundo está de acuerdo. Sin embargo, más allá de eso, hay bastante confusión. Algunas personas sostienen que solo se pueden hacer afirmaciones como "rechazó la nula", "no rechazó la nula" o posiblemente "retuvo la nula". Según esta línea de pensamiento, no se pueden decir cosas como "acepta la alternativa" o "acepta la nula". Personalmente creo que esto es demasiado fuerte. En mi opinión, confunde la prueba de hipótesis nulas con la visión falsacionista del proceso científico de Karl Popper. Aunque hay similitudes entre el falsacionismo y la prueba de hipótesis nula, no son equivalentes. Sin embargo, aunque personalmente creo que está bien hablar de aceptar una hipótesis (con la condición de que "aceptar" no significa que sea necesariamente cierta, especialmente en el caso de la hipótesis nula), mucha gente no estará de acuerdo. Y lo que es más, deberías ser consciente de que esta rareza particular existe para que no te pille desprevenida cuando escribas tus propios resultados.

Table 9.3: prueba estadística de hipótesis nula (NHST) - detalle adicional

	retain H_0	**reject** H_0
H_0 is true	1-α (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

garantía correspondiente con respecto a β. Sin duda, nos gustaría que la tasa de error de tipo II fuera pequeña y tratamos de diseñar pruebas que la mantengan pequeña, pero esto suele ser secundario frente a la abrumadora necesidad de controlar la tasa de error de tipo I. Como habría dicho Blackstone si fuera estadístico, es "mejor retener 10 hipótesis nulas falsas que rechazar una única verdadera". Para ser sincera, no sé si estoy de acuerdo con esta filosofía. Hay situaciones en las que creo que tiene sentido y situaciones en las que creo que no, pero eso no viene al caso. Es como se construyen las pruebas.

9.3 Pruebas estadísticas y distribuciones muestrales

Llegados a este punto, tenemos que empezar a hablar en concreto de cómo se construye una prueba de hipótesis. Para ello, volvamos al ejemplo de la PES. Ignoremos los datos reales que obtuvimos, por el momento, y pensemos en la estructura del experimento. Independientemente de cuáles sean los números reales, la forma de los datos es que X de N personas identificaron correctamente el color de la carta oculta. Además, supongamos por el momento que la hipótesis nula es realmente cierta, que la PES no existe y que la probabilidad real de que alguien elija el color correcto es exactamente $\theta = 0,5$. ¿Cómo esperaríamos que fueran los datos? Bueno, obviamente esperaríamos que la proporción de personas que dan la respuesta correcta fuera bastante cercana al 50%. O, para expresarlo en términos más matemáticos, diríamos que $\frac{X}{N}$ es aproximadamente $0,5$. Por supuesto, no esperaríamos que esta fracción fuera exactamente 0.5. Si, por ejemplo, probamos $N = 100$ personas y $X = 53$ de ellas acertaron la pregunta, probablemente nos veríamos obligadas a admitir que los datos son bastante coherentes con la hipótesis nula. Por otro lado, si $X = 99$ de nuestros participantes acertaron la pregunta, estaríamos bastante seguras de que la hipótesis nula es errónea. Del mismo modo, si solo $X = 3$ personas acertaron la respuesta, estaríamos igualmente seguras de que la hipótesis nula era incorrecta. Seamos un poco más técnicas. Tenemos una cantidad X que podemos calcular observando nuestros datos. Tras observar el valor de X, decidimos si creemos que la hipótesis nula es correcta o si rechazamos la hipótesis nula a favor de la alternativa. El nombre de esta cosa que calculamos para guiar nuestras decisiones es la **prueba estadística**.

Una vez elegida la prueba estadística, el siguiente paso es establecer con precisión qué valores de la prueba estadística harían que se rechazara la hipótesis nula y qué valores harían que la mantuviéramos. Para ello, debemos determinar cuál sería la **distribución muestral de la prueba estadística** si la hipótesis nula fuera realmente cierta (ya hemos hablado de las distribuciones muestrales en Section 8.3.1 ¿Por qué necesitamos esto? Porque esta distribución nos dice exactamente qué valores de X nos llevaría a esperar nuestra hipótesis nula. Y, por tanto, podemos usar esta distribución como

una herramienta para evaluar hasta qué punto la hipótesis nula concuerda con nuestros datos.

¿Cómo determinamos realmente la distribución muestral de la prueba estadística? Para muchas pruebas de hipótesis, este paso es bastante complicado, y más adelante en el libro verás que soy un poco evasiva al respecto para algunas de las pruebas (algunas ni yo misma las entiendo). Sin embargo, a veces es muy fácil. Y, afortunadamente para nosotras, nuestro ejemplo PES nos proporciona uno de los casos más fáciles. Nuestro parámetro poblacional θ es simplemente la probabilidad global de que la gente responda correctamente a la pregunta, y nuestra prueba estadística X es el recuento del número de personas que lo hicieron de una muestra de tamaño N. Ya hemos visto una distribución como esta antes, en Section 7.4, ¡y eso es exactamente lo que describe la distribución binomial! Así que, para usar la notación y la terminología que introduje en esa sección, diríamos que la hipótesis nula predice que X se distribuye binomialmente, lo cual se escribe

$$X \sum Binomial(\theta, N)$$

Dado que la hipótesis nula establece que $\theta = 0.5$ y nuestro experimento tiene $N = 100$ personas, tenemos la distribución muestral que necesitamos. Esta distribución muestral se representa en Figure 9.1. En realidad, no hay sorpresas, la hipótesis nula dice que $X = 50$ es el resultado más probable, y dice que es casi seguro que veamos entre 40 y 60 respuestas correctas.

9.4 Tomando decisiones

Bien, estamos muy cerca de terminar. Hemos construido una prueba estadística (X) y la elegimos de tal manera que estamos bastante seguras de que si X está cerca de $\frac{N}{2}$ entonces deberíamos mantener la nula, y si no, debemos rechazarla. La pregunta que queda es la siguiente. ¿Exactamente qué valores de la prueba estadística deberíamos asociar con la hipótesis nula y exactamente qué valores van con la hipótesis alternativa? En mi estudio PES, por ejemplo, he observado un valor de $X = 62$. ¿Qué decisión debo tomar? ¿Debo optar por creer en la hipótesis nula o en la hipótesis alternativa?

9.4.1 Regiones críticas y valores críticos

Para responder a esta pregunta necesitamos introducir el concepto de **región crítica** para la prueba estadística X. La región crítica de la prueba corresponde a aquellos valores de X que nos llevarían a rechazar la hipótesis nula (razón por la cual la región crítica también se denomina a veces región de rechazo). ¿Cómo encontramos esta región crítica? Consideremos lo que sabemos:

- X debe ser muy grande o muy pequeña para rechazar la hipótesis nula
- Si la hipótesis nula es verdadera, la distribución muestral de X es $Binomial(0.5, N)$
- Si $\alpha = .05$, la región crítica debe cubrir el 5% de esta distribución muestral.

Es importante que comprendas este último punto. La región crítica corresponde a aquellos valores de X para los que rechazaríamos la hipótesis nula, y la distribución muestral en cuestión describe la probabilidad de que obtuviéramos un valor particular de X si la hipótesis nula fuera realmente cierta. Ahora, supongamos que elegimos una

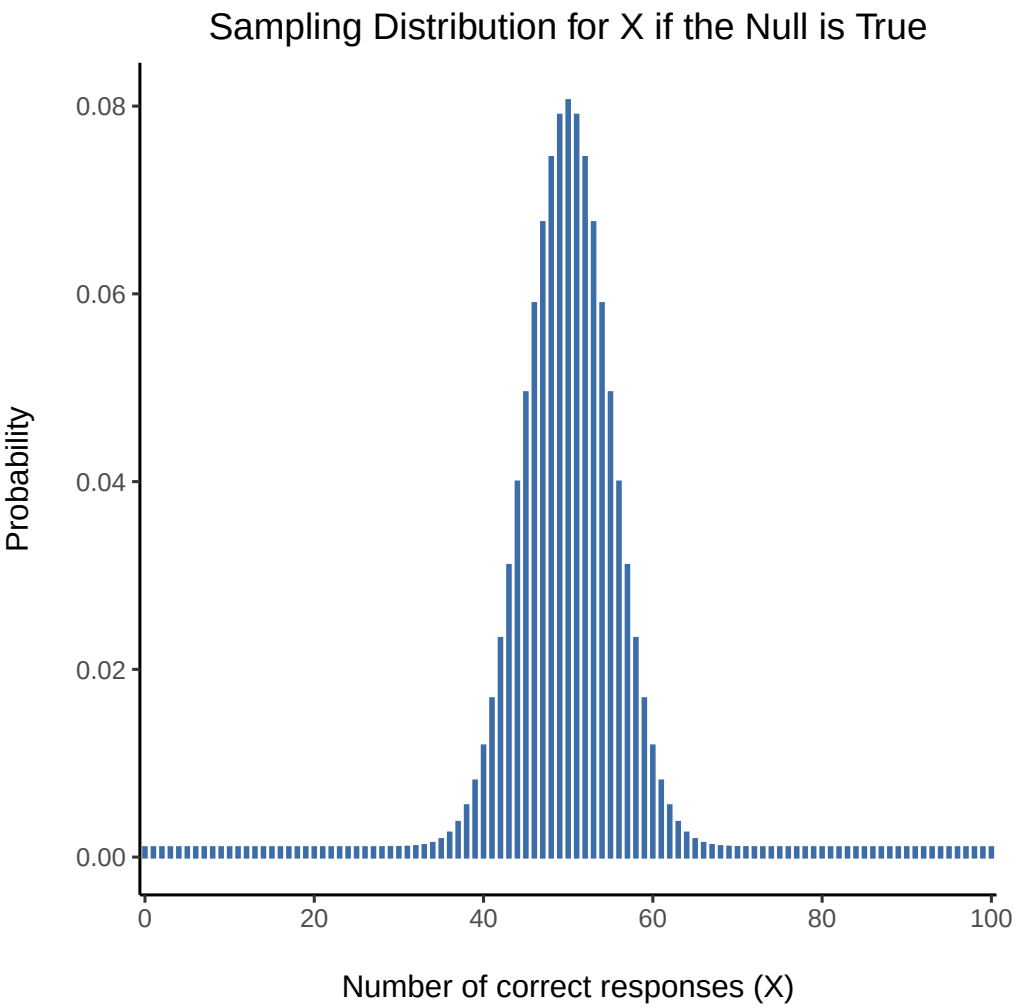


Figure 9.1: La distribución muestral para nuestra prueba estadística X cuando la hipótesis nula es verdadera. Para nuestro escenario PES se trata de una distribución binomial. No es de extrañar, dado que la hipótesis nula dice que la probabilidad de una respuesta correcta es $\theta = .5$, la distribución muestral dice que el valor más probable es 50 (de 100) respuestas correctas. La mayor parte de la masa de probabilidad se encuentra entre 40 y 60

región crítica que cubre 20% de la distribución muestral, y supongamos que la hipótesis nula es realmente cierta. ¿Cuál sería la probabilidad de rechazar incorrectamente la nula? La respuesta es, por supuesto, 20%. Y, por tanto, habríamos construido una prueba que tuviera un nivel de 0.2. Si queremos $\alpha = .05$, la región crítica solo puede cubrir el 5% de la distribución muestral de nuestra prueba estadística.

Resulta que esas tres cosas resuelven el problema de forma única. Nuestra región crítica consiste en los valores más extremos, conocidos como las **colas** de la distribución. Esto se ilustra en Figure 9.2. Si queremos $\alpha = .05$ entonces nuestras regiones críticas corresponden a $X \leq 40$ y $X \geq 60$.[6] Es decir, si el número de personas que dicen " verdadero" está entre 41 y 59, entonces deberíamos mantener la hipótesis nula. Si el número está entre 0 y 40, o entre 60 y 100, debemos rechazar la hipótesis nula. Los números 40 y 60 suelen denominarse **valores críticos**, ya que definen los bordes de la región crítica.

En este punto, nuestra prueba de hipótesis está prácticamente completa:

1. Elegimos un nivel (por ejemplo, $\alpha = .05$);
2. Obtenemos alguna prueba estadística (por ejemplo, X) que haga un buen trabajo (en algún sentido significativo) al comparar H_0 con H_1;
3. Calculamos la distribución muestral de la prueba estadística suponiendo que la hipótesis nula es verdadera (en este caso, binomial); y entonces
4. Calculamos la región crítica que produce un nivel apropiado (0-40 y 60-100).

Todo lo que tenemos que hacer ahora es calcular el valor de la prueba estadística para los datos reales (por ejemplo, X = 62) y luego compararlo con los valores críticos para tomar nuestra decisión. Dado que 62 es mayor que el valor crítico de 60, rechazaríamos la hipótesis nula. O, dicho de otro modo, decimos que la prueba produjo un resultado estadísticamente **significativo**.

9.4.2 Una nota sobre la "significación" estadística

> *Al igual que otras técnicas ocultas de adivinación, el método estadístico tiene una jerga privada deliberadamente concebida para ocultar sus métodos a los no practicantes.*
> – Atribuido a GO Ashley [7]

Llegados a este punto, conviene hacer una breve digresión sobre la palabra "significativo". El concepto de significación estadística es en realidad muy sencillo, pero tiene un nombre muy desafortunado. Si los datos nos permiten rechazar la hipótesis nula, decimos que "el resultado es estadísticamente significativo", que a menudo se abrevia como "el resultado es significativo". Esta terminología es bastante antigua y se remonta a una época en la que "significativo" solo significaba algo así como "indicado", en lugar de su significado moderno, que es mucho más cercano a "importante". Como resultado, muchos lectores modernos se confunden mucho cuando comienzan a aprender estadística porque piensan que un "resultado significativo" debe ser importante. No significa eso en absoluto. Lo único que significa "estadísticamente significativo" es que los datos nos han permitido

[6] Estrictamente hablando, la prueba que acabo de construir tiene $\alpha = .057$, que es un poco demasiado generosa. Sin embargo, si hubiera elegido 39 y 61 como límites de la región crítica, ésta solo cubriría 3.5% de la distribución. Pensé que tiene más sentido usar 40 y 60 como mis valores críticos, y estar dispuesta a tolerar una tasa de error tipo I de 5.7%, ya que eso es lo más cerca que puedo llegar a un valor de $\alpha = .05$.

[7] Internet parece bastante convencido de que Ashley dijo esto, aunque no puedo encontrar a nadie dispuesto a dar una fuente para la afirmación.

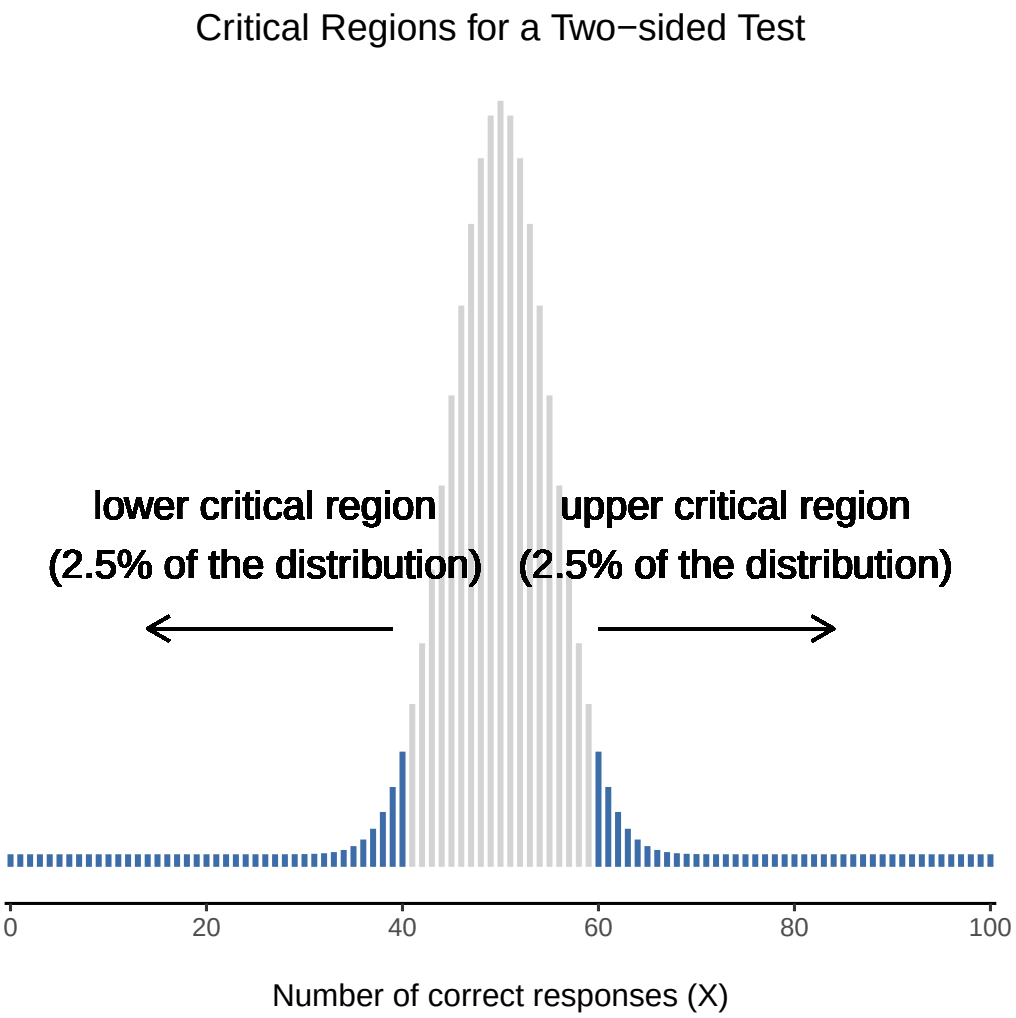


Figure 9.2: La región crítica asociada con la prueba de hipótesis para el estudio PES, para una prueba de hipótesis con un nivel de significación de $\alpha = .05$. El gráfico muestra la distribución muestral de X bajo la hipótesis nula (es decir, igual que Figure 9.1). Las barras grises corresponden a aquellos valores de X para los que mantendríamos la hipótesis nula. Las barras azules (sombreadas más oscuras) muestran la región crítica, aquellos valores de X para los que rechazaríamos la hipótesis nula. Debido a que la hipótesis alternativa es bilateral (es decir, permite tanto $\theta < .5$ como $\theta > .5$, la región crítica cubre ambas colas de la distribución. Para asegurar un nivel de α de \$.05 \$, debemos asegurarnos de que cada una de las dos regiones abarca 2.5% de la distribución muestral

rechazar una hipótesis nula. Si el resultado es realmente importante o no en el mundo real es una cuestión muy diferente, y depende de muchas otras cosas.

9.4.3 La diferencia entre pruebas unilaterales y bilaterales

Hay una cosa más que quiero señalar sobre la prueba de hipótesis que acabo de construir. Si nos tomamos un momento para pensar en las hipótesis estadísticas que he estado usando,

$$H_0 : \theta = 0.5$$

$$H_1 : \theta \neq 0.5$$

nos damos cuenta de que la hipótesis alternativa cubre tanto la posibilidad de que $ $< .5$ como la posibilidad de que θ .5. Esto tiene sentido si realmente creo que PES podría producir un rendimiento mejor que el azar o un rendimiento peor que el azar (y hay algunas personas que piensas así). En lenguaje estadístico, este es un ejemplo de una **prueba bilateral**. Se llama así porque la hipótesis alternativa cubre el área a ambos "lados" de la hipótesis nula y, en consecuencia, la región crítica de la prueba cubre ambas colas de la distribución muestral (2.5 % a cada lado si = .05), como se ilustró anteriormente en Figure 9.2. Sin embargo, esa no es la única posibilidad. Es posible que solo estés dispuesta a creer en PES si produce un rendimiento mejor que el azar. Si es así, entonces mi hipótesis alternativa solo cubriría la posibilidad de que $\theta > .5$, y como consecuencia la hipótesis nula ahora se convierte en

$$H_0 : \theta \leq 0.5$$

$$H_1 : \theta > 0.5$$

Cuando esto ocurre, tenemos lo que se llama una **prueba unilateral** y la región crítica solo cubre una cola de la distribución muestral. Esto se ilustra en Figure 9.3.

9.5 El valor p de una prueba

En cierto sentido, nuestra prueba de hipótesis está completa. Hemos construido una prueba estadística, calculado su distribución muestral si la hipótesis nula es verdadera y a continuación construido la región crítica para la prueba. Sin embargo, en realidad he omitido el número más importante de todos, **el valor p**. A este tema nos referimos ahora. Hay dos formas algo diferentes de interpretar el valor p, una propuesta por Sir Ronald Fisher y la otra por Jerzy Neyman. Ambas versiones son legítimas, aunque reflejan formas muy diferentes de pensar sobre las pruebas de hipótesis. La mayoría de los libros de texto introductorios tienden a dar solo la versión de Fisher, pero creo que es una lástima. En mi opinión, la versión de Neyman es más limpia y en realidad refleja mejor la lógica de la prueba de hipótesis nula. Sin embargo, puede que no estés de acuerdo, así que he incluido ambas. Empezaré con la versión de Neyman.

9.5.1 Una visión más suave de la toma de decisiones

Un problema con el procedimiento de prueba de hipótesis que he descrito es que no distingue entre un resultado que es "apenas significativo" y los que son "altamente significativos". Por ejemplo, en mi estudio de PES, los datos que obtuve apenas cayeron dentro de la región crítica, por lo que obtuve un efecto significativo, pero por muy poco.

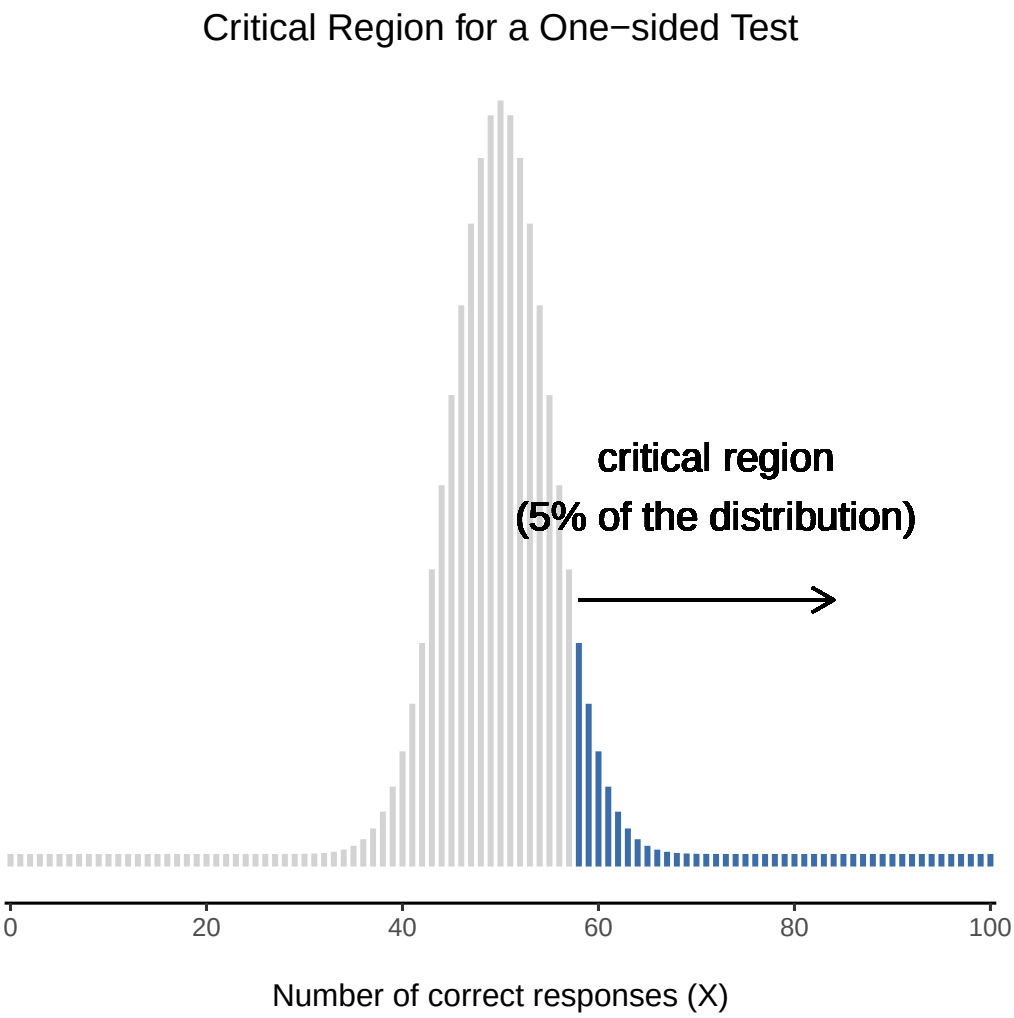


Figure 9.3: La región crítica para una prueba unilateral. En este caso, la hipótesis alternativa es que $\theta \geq .5$ por lo que solo rechazaríamos la hipótesis nula para valores grandes de X. Como consecuencia, la región crítica solo cubre la cola superior de la distribución muestral, concretamente el 5% superior de la distribución. Contrasta esto con la versión bilateral en Figure 9.2

Por el contrario, supongamos que hubiera realizado un estudio en el que $X = 97$ de mis $N = 100$ participantes hubieran acertado la respuesta. Obviamente, esto también sería significativo, pero con un margen mucho mayor, por lo que no habría ambigüedad al respecto. El procedimiento que ya he descrito no distingue entre ambos. Si adopto la convención estándar de permitir $\alpha = .05$ como tasa de error Tipo I aceptable, entonces ambos resultados son significativos.

Aquí es donde el valor p resulta útil. Para entender cómo funciona, supongamos que realizamos muchas pruebas de hipótesis en el mismo conjunto de datos, pero con un valor diferente de en cada caso. Cuando hacemos eso para mis datos de PES originales, lo que obtendríamos es algo como Table 9.4.

Table 9.4: Rechazo de la hipótesis nula a diferentes niveles de alfa

Value of α	**0.05**	**0.04**	**0.03**	**0.02**	**0.01**
Reject the null?	Yes	Yes	Yes	No	No

Cuando probamos los datos PES ($X = 62$ éxitos de $N = 100$ observaciones), usando α niveles de .03 y superiores, siempre nos encontramos rechazando la hipótesis nula. Para niveles de α de .02 e inferiores, siempre terminamos manteniendo la hipótesis nula. Por lo tanto, en algún lugar entre .02 y .03 debe haber un valor más pequeño que α que nos permita rechazar la hipótesis nula para estos datos. Este es el valor de p. Resulta que los datos PES tienen $p = .021$. En resumen, p se define como la tasa de error Tipo I más pequeña (α) que debemos estar dispuestas a tolerar si queremos rechazar la hipótesis nula.

Si resulta que p describe una tasa de error que te parece intolerable, entonces debes mantener la hipótesis nula. Si te sientes cómoda con una tasa de error igual a p, entonces está bien rechazar la hipótesis nula a favor de tu alternativa preferida.

En efecto, p es un resumen de todas las posibles pruebas de hipótesis que podrías haber realizado, tomadas a través de todos los valores posibles de . Y como consecuencia tiene el efecto de "suavizar" nuestro proceso de decisión. Para aquellas pruebas en las que p ď habría rechazado la hipótesis nula, mientras que para aquellas pruebas en las que p ą habría mantenido la nula. En mi estudio PES obtuve $X = 62$ y como consecuencia terminé con $p = .021$. Así que la tasa de error que debo tolerar es de 2.1%. Por el contrario, supongamos que mi experimento arrojó $X = 97$. ¿Qué sucede con mi valor p ahora? Esta vez se ha reducido a $p = 1.36$ x 10^{-25}, que es una tasa de error de Tipo I minúscula, minúscula[8]. En este segundo caso, podrías rechazar la hipótesis nula con mucha más confianza, porque solo tengo que estar "dispuesta" a tolerar una tasa de error tipo I de aproximadamente $ 1 $ en $ 10 $ billones de billones para justificar mi decisión de rechazar.

9.5.2 La probabilidad de datos extremos

La segunda definición del valor p proviene de Sir Ronald Fisher, y en realidad es esta la que suele aparecer en la mayoría de los libros de texto de introducción a la estadística.

[8] ¡Eso es $p = .0000000000000000000000000136$ para las personas a las que no les gusta la notación científica!

¿Te das cuenta de que, cuando construí la región crítica, correspondía a las colas (es decir, valores extremos) de la distribución muestral? Eso no es una coincidencia, casi todas las pruebas "buenas" tienen esta característica (buenas en el sentido de minimizar nuestra tasa de error tipo II, β). La razón es que una buena región crítica casi siempre corresponde a aquellos valores de la prueba estadística que es menos probable que se observen si la hipótesis nula es cierta. Si esta regla es cierta, podemos definir el valor p como la probabilidad de que hubiéramos observado una prueba estadística que sea al menos tan extrema como la que realmente obtuvimos. En otras palabras, si los datos son extremadamente inverosímiles según la hipótesis nula, entonces es probable que la hipótesis nula sea errónea.

9.5.3 Un error común

De acuerdo, puedes ver que hay dos formas bastante diferentes pero legítimas de interpretar el valor p, una basada en el enfoque de Neyman para la prueba de hipótesis y la otra basada en el de Fisher. Desgraciadamente, hay una tercera explicación que la gente da a veces, especialmente cuando están aprendiendo estadística por primera vez, y es *absoluta y completamente incorrecta*. Este enfoque erróneo consiste en referirse al valor de p como "la probabilidad de que la hipótesis nula sea verdadera". Es una forma de pensar intuitivamente atractiva, pero errónea en dos aspectos clave. En primer lugar, la prueba de hipótesis nula es una herramienta frecuentista y el enfoque frecuentista de la probabilidad no permite asignar probabilidades a la hipótesis nula. Según esta visión de la probabilidad, la hipótesis nula o es cierta o no lo es, no puede tener un "5% de probabilidad" de ser cierta. En segundo lugar, incluso dentro del enfoque bayesiano, que sí permite asignar probabilidades a las hipótesis, el valor de p no correspondería a la probabilidad de que la nula sea cierta. Esta interpretación es totalmente incoherente con las matemáticas de cómo se calcula el valor p. Dicho sin rodeos, a pesar del atractivo intuitivo de pensar así, no hay justificación para interpretar un valor p de esta manera. No lo hagas nunca.

9.6 Informar los resultados de una prueba de hipótesis

Cuando se escriben los resultados de una prueba de hipótesis, suele haber varios elementos que hay que informar, pero varían bastante de una prueba a otra. A lo largo del resto del libro, dedicaré algo de tiempo a hablar sobre cómo informar de los resultados de diferentes pruebas (consulta Section 10.1.9 para ver un ejemplo especialmente detallado, para que puedas hacerte una idea de cómo se hace normalmente). Sin embargo, independientemente de la prueba que estés haciendo, lo único que siempre tienes que hacer es decir algo sobre el valor de p y si el resultado fue significativo o no.

El hecho de tener que hacer esto no es sorprendente, es el objetivo de la prueba. Lo que puede sorprender es que haya cierta controversia sobre cómo hacerlo exactamente. Dejando a un lado a las personas que están completamente en desacuerdo con todo el marco en el que se basa la prueba de hipótesis nula, existe cierta tensión sobre si se debe informar o no el valor exacto de p que se ha obtenido, o si sólo se debe indicar que $p < \alpha$ para un nivel de significación que se ha elegido de antemano (por ejemplo, $p < .05$).

9.6.1 La cuestión

Para ver por qué esto es un problema, la clave es reconocer que los valores p son terriblemente convenientes. En la práctica, el hecho de que podamos calcular el valor p significa que en realidad no tenemos que especificar ningún nivel α para realizar la prueba. En su lugar, lo que puedes hacer es calcular su valor p e interpretarlo directamente. Si obtienes $p = 0,062$, significa que tendrías que estar dispuesta a tolerar una tasa de error de tipo I de $6,2\%$ para justificar el rechazo de la hipótesis nula. Si tú personalmente encuentras 6.2% intolerable entonces retienes la hipótesis nula. Por lo tanto, se argumenta, ¿por qué no nos limitamos a comunicar el valor real de p y dejamos que el lector decida por sí mismo cuál es la tasa de error de Tipo I aceptable? Este enfoque tiene la gran ventaja de "suavizar" el proceso de toma de decisiones. De hecho, si aceptas la definición de Neyman del valor p, ese es el punto central del valor p. Ya no tenemos un nivel de significación fijo de $\alpha = .05$ como una línea brillante que separa las decisiones de "aceptar" de las de "rechazar", y esto elimina el problema bastante patológico de verse obligado a tratar $p = .051$ de una manera fundamentalmente diferente a $p = .049$.

Esta flexibilidad es a la vez una ventaja y un inconveniente del valor p. La razón por la que a mucha gente no le gusta la idea de comunicar un valor p exacto es que le da demasiada libertad al investigador. En particular, le permite cambiar de opinión sobre la tolerancia de error que está dispuesto a tolerar después de ver los datos. Por ejemplo, consideremos mi experimento PES. Supongamos que realicé mi prueba y terminé con un valor de p de .09. ¿Debo aceptar o rechazar? Para ser sincera, todavía no me he molestado en pensar qué nivel de error Tipo I estoy "realmente" dispuesta a aceptar. No tengo una opinión sobre ese tema. Pero *sí* tengo una opinión sobre si la PES existe o no, y *definitivamente* tengo una opinión sobre si mi investigación debería publicarse en una revista científica de prestigio. Y sorprendentemente, ahora que he mirado los datos, estoy empezando a pensar que una tasa de error de 9% no es tan mala, especialmente cuando se compara con lo molesto que sería tener que admitirle al mundo que mi experimento ha fracasado. Así que, para evitar que parezca que lo inventé a posteriori, ahora digo que mi α es .1, con el argumento de que una tasa de error tipo I de 10% no es tan mala y en ese nivel mi prueba es significativa! Yo gano.

En otras palabras, lo que me preocupa es que aunque tenga las mejores intenciones y sea la persona más honesta, la tentación de "matizar" las cosas aquí y allá es muy, muy fuerte. Como puede atestiguar cualquiera que haya realizado un experimento alguna vez, es un proceso largo y difícil y, a menudo, te apegas mucho a tus hipótesis. Es difícil dejarlo ir y admitir que el experimento no encontró lo que querías que encontrara. Y ese es el peligro. Si usamos el valor p "en bruto", la gente empezará a interpretar los datos en términos de lo que quieren creer, no de lo que los datos dicen en realidad y, si permitimos eso, ¿por qué nos molestamos en hacer ciencia? ? ¿Por qué no dejar que todo el mundo crea lo que quiera sobre cualquier cosa, independientemente de los hechos? Vale, eso es un poco extremo, pero de ahí viene la preocupación. Según este punto de vista, realmente hay que especificar el valor α de antemano y luego solo informar de si la prueba fue significativa o no. Es la única manera de mantenernos honestos.

9.6.2 Dos soluciones propuestas

En la práctica, es bastante raro que un investigador especifique un único nivel de antemano. En su lugar, la convención es que los científicos se basen en tres niveles de significación estándar: .05, .01 y .001. Al comunicar los resultados, se indica cuál de

Table 9.5: Traducciones típicas de los niveles del valor p

Usual notation	**Signif. stars**	**English translation**	**The null is...**
$p > .05$		The test wasn't significant	Retained
$p < .05$	*	The test was significant at $\alpha = .05$ but not at $\alpha = .01$ or $\alpha = .001$.	Rejected
$p < .01$	**	The test was significant at $\alpha = .05$ and $\alpha = .01$ but not at $\alpha = .001$.	Rejected
$p < .001$	***	The test was significant at all levels	Rejected

estos niveles de significación (si es que hay alguno) permite rechazar la hipótesis nula. Esto se resume en Table 9.5. Esto nos permite suavizar un poco la regla de decisión, ya que $p < .01$ implica que los datos cumplen con un estándar probatorio más estricto que $p < .05$. Sin embargo, dado que estos niveles se fijan de antemano por convención, se evita que la gente elija su nivel después de ver los datos.

Sin embargo, mucha gente todavía prefiere comunicar valores de p exactos. Para muchas personas, la ventaja de permitir que el lector tome sus propias decisiones sobre cómo interpretar p = 0,06 supera cualquier desventaja. Sin embargo, en la práctica, incluso entre aquellos investigadores que prefieren valores de p exactos, es bastante común escribir $p < .001$ en lugar de informar un valor exacto para p pequeño. Esto se debe en parte a que una gran cantidad de software en realidad no imprime el valor p cuando es tan pequeño (p. ej., SPSS solo escribe $p = .000$ siempre que $p < .001$), y en parte porque un valor muy pequeño de p puede ser engañoso. La mente humana ve un número como .0000000001 y es difícil suprimir la sensación visceral de que las pruebas a favor de la hipótesis alternativa son casi seguras. En la práctica, sin embargo, esto suele ser erróneo. La vida es algo grande, desordenado y complicado, y todas las pruebas estadísticas que se han inventado se basan en simplificaciones, aproximaciones y suposiciones. Como consecuencia, probablemente no sea razonable salir de ningún análisis estadístico con una sensación de confianza mayor de la que implica $p < .001$. En otras palabras, $p < .001$ es en realidad un código para "en lo que respecta a esta prueba, las pruebas son abrumadoras".

A la luz de todo esto, es posible que te preguntes qué debes hacer exactamente. Hay bastantes consejos contradictorios sobre el tema, con algunas personas que sostienen que debes informar el valor p exacto y otras que debes usar el enfoque escalonado ilustrado en Table 9.1. Como resultado, el mejor consejo que puedo dar es sugerir que mires

los artículos/informes escritos en tu campo y veas cuál parece ser la convención. Si no parece haber ningún patrón coherente, utiliza el método que prefieras.

9.7 Ejecutando la prueba de hipótesis en la práctica

Llegados a este punto, algunas os estaréis preguntando si se trata de una prueba de hipótesis "real" o solo de un ejemplo de juguete que me he inventado. Es real. En la discusión anterior construí la prueba a partir de los primeros principios, pensando que era el problema más simple que podrías encontrarte en la vida real. Sin embargo, esta prueba ya existe. Se llama *prueba binomial*, y jamovi la implementa como uno de los análisis estadísticos disponibles cuando pulsas el botón 'Frecuencias'. Para probar la hipótesis nula de que la probabilidad de respuesta es la mitad de $p = .5$,[9] y usando datos en los que $x = 62$ de $n = 100$ personas dieron la respuesta correcta respuesta, disponible en el archivo de datos binomialtest.omv, obtenemos los resultados que se muestran en Figure 9.4.

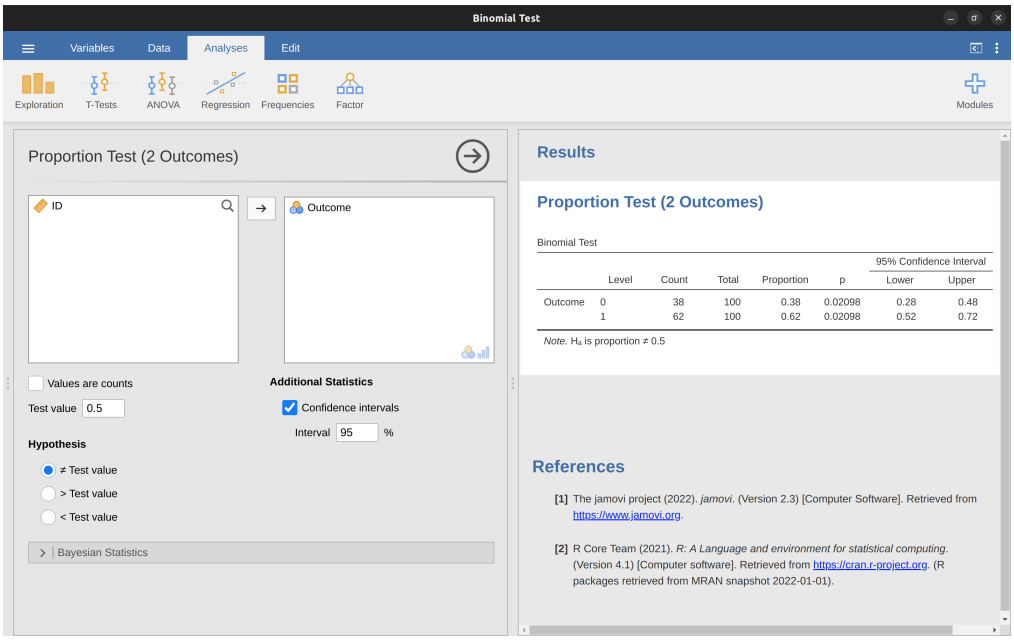


Figure 9.4: análisis de prueba binomial y resultados en jamovi

En este momento, esta salida te parece bastante desconocida, pero puedes ver que te está diciendo más o menos las cosas correctas. En concreto, el valor p de $0,02$ es menor que la elección habitual de $\alpha = 0,05$, por lo que puedes rechazar la hipótesis nula. Hablaremos mucho más sobre cómo leer este tipo de salida a medida que avancemos, y después de un tiempo, con suerte, lo encontrarás bastante fácil de leer y comprender.

[9] Ten en cuenta que la p aquí no tiene nada que ver con un valor de p. El argumento p en la prueba binomial de jamovi corresponde a la probabilidad de dar una respuesta correcta, según la hipótesis nula. En otras palabras, es el valor θ.

9.8 Tamaño del efecto, tamaño de la muestra y potencia

En secciones anteriores, he hecho hincapié en el hecho de que el principal principio de diseño que subyace a las pruebas de hipótesis estadísticas es que intentamos controlar nuestra tasa de error Tipo I. Cuando fijamos $\alpha = .05$ estamos intentando asegurarnos que solo 5% de las hipótesis nulas verdaderas se rechacen incorrectamente. Sin embargo, esto no significa que no nos importen los errores de tipo II. De hecho, desde la perspectiva del investigador, el error de no rechazar la nula cuando en realidad es falsa es extremadamente molesto. Teniendo eso en cuenta, un objetivo secundario de las pruebas de hipótesis es tratar de minimizar β, la tasa de error de Tipo II, aunque normalmente no hablamos en términos de minimizar los errores de Tipo II. En su lugar, hablamos de maximizar la potencia de la prueba. Dado que la potencia se define como $1 - \beta$, es lo mismo.

9.8.1 La función de potencia

Pensemos un momento en qué es realmente un error de tipo II. Un error de tipo II se produce cuando la hipótesis alternativa es verdadera, pero sin embargo no somos capaces de rechazar la hipótesis nula. Lo ideal sería poder calcular un único número β que nos indicara la tasa de error de Tipo II, de la misma manera que podemos establecer $\alpha = .05$ para la tasa de error de Tipo I. Desafortunadamente, esto es mucho más complicado. Para ver esto, observa que en mi estudio PES la hipótesis alternativa en realidad corresponde a un montón de posibles valores de θ. De hecho, la hipótesis alternativa corresponde a cada valor de θ excepto 0,5. Supongamos que la probabilidad real de que alguien elija la respuesta correcta es del 55% (es decir, $\theta = .55$). Si es así, entonces la verdadera distribución muestral para X no es la misma que predice la hipótesis nula, ya que el valor más probable para X ahora es 55 de 100. No solo eso, toda la distribución muestral se ha desplazado, como se muestra en Figure 9.5. Las regiones críticas, por supuesto, no cambian. Por definición, las regiones críticas se basan en lo que predice la hipótesis nula. Lo que vemos en esta figura es el hecho de que cuando la hipótesis nula es errónea, una proporción mucho mayor de la distribución muestral cae en la región crítica. Y, por supuesto, eso es lo que debería suceder. ¡La probabilidad de rechazar la hipótesis nula es mayor cuando la hipótesis nula es realmente falsa! Sin embargo $\theta = .55$ no es la única posibilidad consistente con la hipótesis alternativa. Supongamos que el verdadero valor de θ es en realidad $0,7$. ¿Qué sucede con la distribución muestral cuando esto ocurre? La respuesta, que se muestra en Figure 9.6, es que casi la totalidad de la distribución muestral ahora se ha movido a la región crítica. Por tanto, si $\theta = 0,7$, la probabilidad de que rechacemos correctamente la hipótesis nula (es decir, la potencia de la prueba) es mucho mayor que si $\theta = 0,55$. En resumen, aunque $\theta = .55$ y $\theta = .70$ forman parte de la hipótesis alternativa, la tasa de error de Tipo II es diferente.

Lo que todo esto significa es que la potencia de una prueba (es decir, $1 - \beta$) depende del verdadero valor de θ. Para ilustrar esto, he calculado la probabilidad esperada de rechazar la hipótesis nula para todos los valores de θ y la he representado en Figure 9.7. Este gráfico describe lo que normalmente se denomina función de potencia de la prueba. Es un buen resumen de lo buena que es la prueba, porque en realidad nos dice la potencia $(1 - \beta)$ para todos los valores posibles de θ. Como se puede ver, cuando el valor verdadero de θ está muy cerca de $0,5$, la potencia de la prueba cae bruscamente,

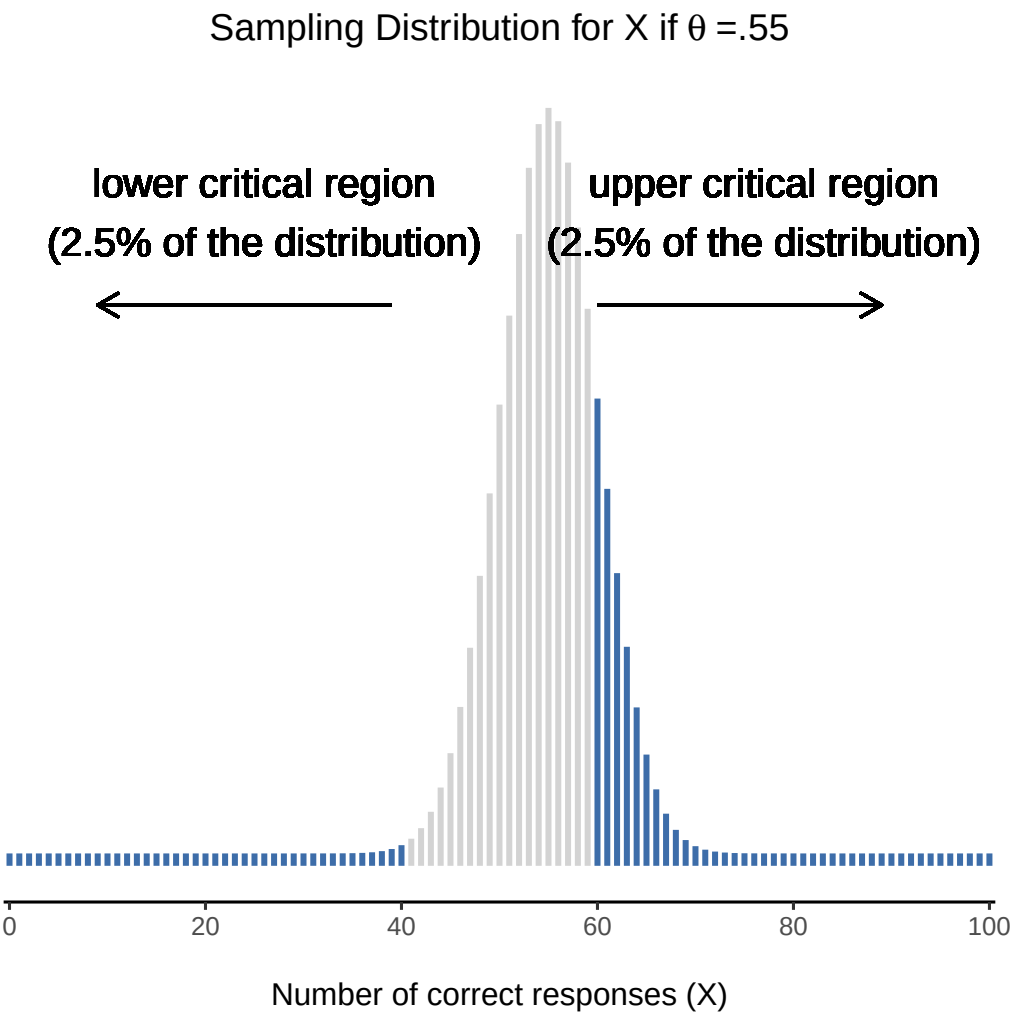


Figure 9.5: Distribución muestral bajo la hipótesis alternativa para un valor de parámetro poblacional de $\theta = 0.55$. Una proporción razonable de la distribución se encuentra en la región de rechazo.

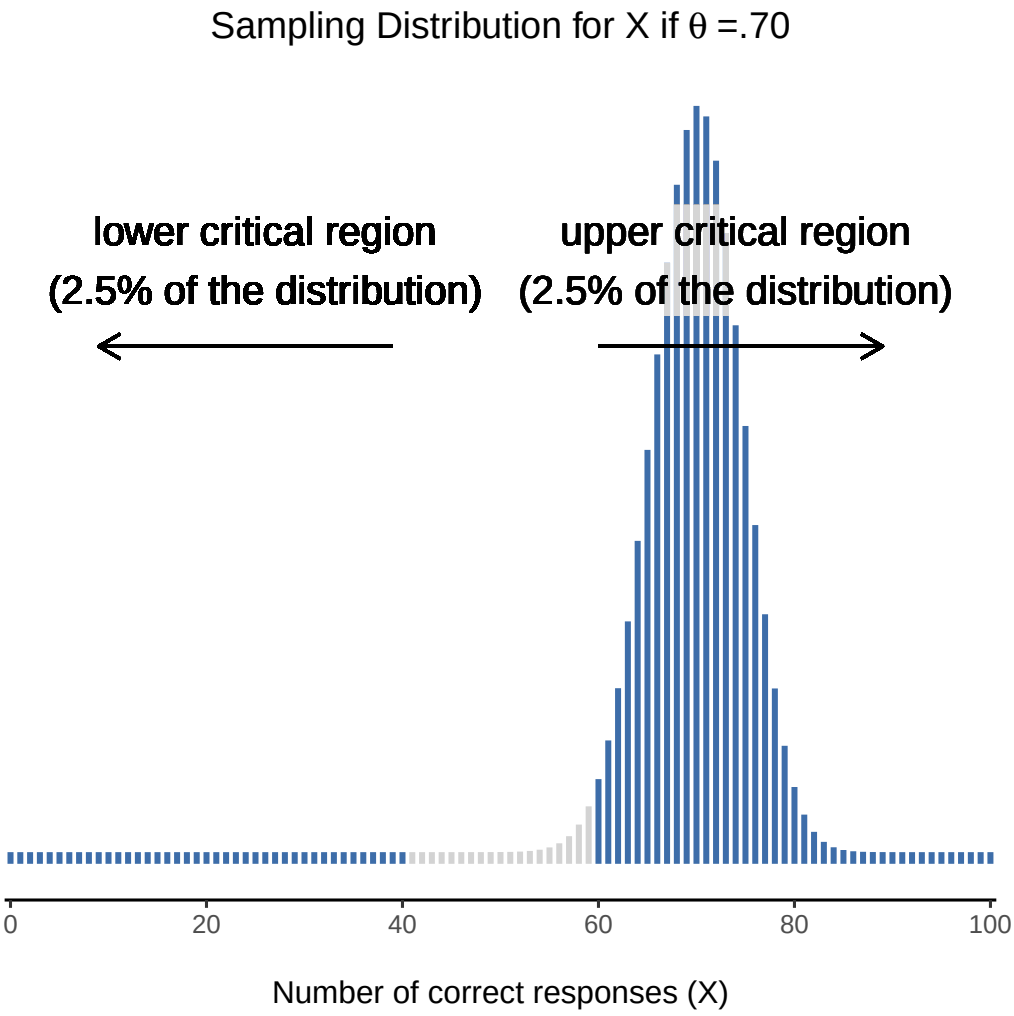


Figure 9.6: Distribución muestral bajo la hipótesis *alternativa* para un valor del parámetro poblacional de $\theta = 0.70$. Casi toda la distribución se encuentra en la región de rechazo.

pero cuando está más lejos, la potencia es grande.

9.8.2 La función de potencia

> *Dado que todos los modelos son erróneos, el científico debe estar alerta a lo que es erróneo de manera importante. No es apropiado preocuparse por los ratones cuando hay tigres en el exterior*
> - Caja de George (Box 1976, p. 792)

El gráfico que se muestra en Figure 9.7 refleja un aspecto básico de las pruebas de hipótesis. Si el estado real del mundo es muy diferente de lo que predice la hipótesis nula, la potencia será muy alta, pero si el estado real del mundo es similar a la hipótesis nula (pero no idéntico), la potencia de la prueba será muy baja. Por lo tanto, es útil poder tener alguna forma de cuantificar lo "similar" que es el verdadero estado del mundo a la hipótesis nula. Un estadístico que hace esto se llama medida del **tamaño del efecto** (p. ej., Cohen (1988); Ellis (2010)). El tamaño del efecto se define de forma ligeramente diferente en diferentes contextos (por lo que esta sección solo habla en términos generales), pero la idea cualitativa que intenta captar es siempre la misma (ver, por ejemplo, Table 9.6). ¿Cuán grande es la diferencia entre los parámetros poblacionales *verdaderos* y los valores de los parámetros asumidos por la hipótesis nula? En nuestro ejemplo PES, si dejamos que $\theta_0 = 0.5$ denote el valor asumido por la hipótesis nula y dejamos que θ denote el valor verdadero, entonces una medida simple del tamaño del efecto podría ser algo así como la diferencia entre el valor verdadero y el nulo (es decir, $\theta - \theta_0$), o posiblemente solo la magnitud de esta diferencia, $abs(\theta - \theta_0)$.

Table 9.6: Una guía básica para entender la relación entre la significación estadística y los tamaños del efecto. Básicamente, si no se obtiene un resultado significativo, el tamaño del efecto carece de sentido porque no hay pruebas de que sea real. Por otro lado, si se obtiene un efecto significativo pero el tamaño del efecto es pequeño, es muy probable que el resultado (aunque sea real) no sea tan interesante. Sin embargo, esta guía es muy rudimentaria. Depende mucho de lo que se esté estudiando exactamente. Los pequeños efectos pueden tener una enorme importancia práctica en algunas situaciones. Así que no te tomes esta tabla demasiado en serio. Como mucho, es una guía aproximada.

	big effect size	**small effect size**
significant result	difference is real, and of practical importance	difference is real, but might not be interesting
non-significant result	no effect observed	no effect observed

¿Por qué calcular el tamaño del efecto? Supongamos que has realizado el experimento, has recogido los datos y has obtenido un efecto significativo al realizar la prueba de hipótesis. ¿No basta con decir que se ha obtenido un efecto significativo? ¿Seguro que ese es el objetivo de las pruebas de hipótesis? Bueno, más o menos. Sí, el objetivo de hacer una prueba de hipótesis es intentar demostrar que la hipótesis nula es errónea, pero eso no es lo único que nos interesa. Si la hipótesis nula afirmaba que $\theta = .5$ y demostramos que es errónea, en realidad sólo hemos contado la mitad de la historia. Rechazar la hipótesis nula implica que creemos que $\theta \neq .5$, pero hay una gran diferencia entre $\theta = .51$ y $\theta = .8$. Si encontramos que $\theta = .8$, entonces no solo descubrimos que la hipótesis nula es errónea, sino que parece ser muy errónea. Por otro lado, supongamos

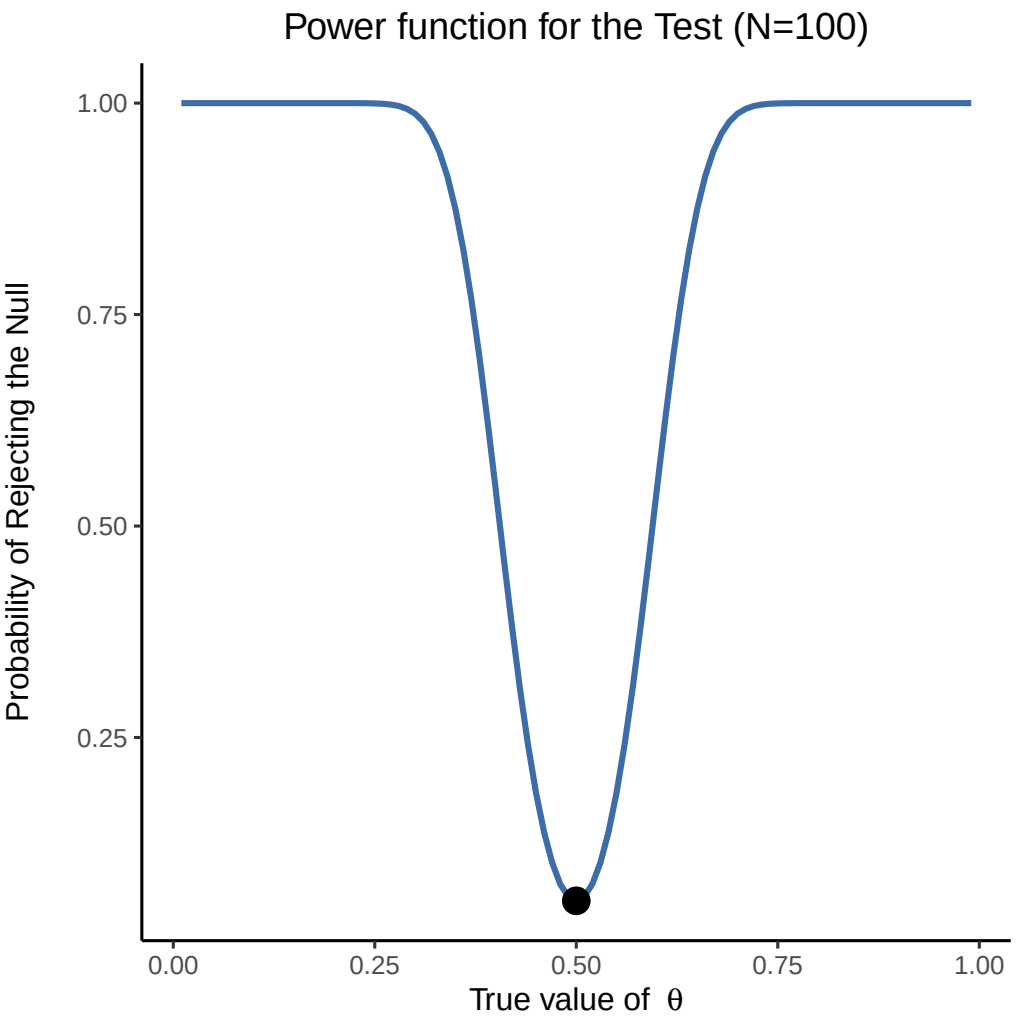


Figure 9.7: La probabilidad de que rechacemos la hipótesis nula, representada como una función del valor verdadero de θ. Obviamente, la prueba es más poderosa (mayor probabilidad de rechazo correcto) si el valor verdadero de θ es muy diferente del valor que especifica la hipótesis nula (es decir, $\theta = .5$). Observa que cuando θ en realidad es igual a .5 (representado como un punto negro), la hipótesis nula es de hecho verdadera y rechazar la hipótesis nula en este caso sería un error de Tipo I

que hemos rechazado con éxito la hipótesis nula, pero parece que el verdadero valor de θ es solo 0,51 (esto solo sería posible con un estudio muy grande). Claro que la hipótesis nula es errónea, pero no está nada claro que nos importe porque el tamaño del efecto es muy pequeño. En el contexto de mi estudio sobre la percepción extrasensorial, es posible que sí nos importe, ya que cualquier demostración de poderes psíquicos reales sería muy interesante[10], pero en otros contextos, una diferencia de 1% generalmente no es muy interesante, aunque sea una diferencia real. Por ejemplo, supongamos que estamos estudiando las diferencias en los resultados de los exámenes de la escuela secundaria entre hombres y mujeres y resulta que los resultados de las mujeres son 1% más altas en promedio que los de los hombres. Si tengo datos de miles de estudiantes, es casi seguro que esta diferencia será estadísticamente significativa, pero independientemente de lo pequeño que sea el valor p, simplemente no es muy interesante. Difícilmente querrías ir por ahí proclamando una crisis en la educación de los chicos basándote en una diferencia tan pequeña, ¿verdad? Por este motivo cada vez es más habitual (lenta, pero inexorablemente) comunicar algún tipo de medida estándar del tamaño del efecto junto con los resultados de la prueba de hipótesis. La prueba de hipótesis en sí te dice si debes creer que el efecto que has observado es real (es decir, que no se debe al azar), mientras que el tamaño del efecto te dice si debes preocuparte o no.

9.8.3 Aumentando la potencia de tu estudio

No es de extrañar que los científicos estén obsesionados con maximizar la potencia de sus experimentos. Queremos que nuestros experimentos funcionen y, por tanto, maximizar la probabilidad de rechazar la hipótesis nula si es falsa (y por supuesto, por lo general, queremos creer que es falsa). Como hemos visto, un factor que influye en la potencia es el tamaño del efecto. Así que lo primero que puedes hacer para aumentar tu potencia es aumentar el tamaño del efecto. En la práctica, esto significa que hay que diseñar el estudio de forma que aumente el tamaño del efecto. Por ejemplo, en mi estudio sobre la percepción extrasensorial podría creer que los poderes psíquicos funcionan mejor en una habitación tranquila y oscura con menos distracciones que nublen la mente. Por lo tanto, trataría de realizar mis experimentos en un entorno así. Si puedo reforzar de algún modo las capacidades de PES de las personas, entonces el valor real de θ aumentará [11] y, por tanto, el tamaño de mi efecto será mayor. En resumen, un diseño experimental inteligente es una forma de aumentar la potencia, ya que puede alterar el tamaño del efecto.

Por desgracia, a menudo ocurre que incluso con el mejor de los diseños experimentales sólo se obtiene un efecto pequeño. Tal vez, por ejemplo, la PES exista realmente , pero incluso en las mejores condiciones es muy, muy débil. En esas circunstancias, lo mejor para aumentar la potencia es aumentar el tamaño de la muestra. En general, cuantas más observaciones tengas disponibles, más probable es que puedas discriminar entre dos hipótesis. Si realizara mi experimento de PES con 10 participantes y 7 de ellos adivinaron correctamente el color de la carta oculta, no estarías muy impresionada. Pero si lo realizara con 10.000 participantes, y 7.000 de ellos acertaran la respuesta,

[10] Ten en cuenta que la p aquí no tiene nada que ver con un valor de p. El argumento p en la prueba binomial de jamovi corresponde a la probabilidad de dar una respuesta correcta, según la hipótesis nula. En otras palabras, es el valor θ.

[11] Observa que el verdadero parámetro poblacional θ no corresponde necesariamente a un hecho inmutable de la naturaleza. En este contexto, θ no es más que la probabilidad real de que la gente adivine correctamente el color de la carta de la otra habitación. Como tal, el parámetro poblacional puede verse influido por todo tipo de cosas. Por supuesto, todo esto suponiendo que la PES exista.

sería mucho más probable que pensaras que había descubierto algo. En otras palabras, la potencia aumenta con el tamaño de la muestra. Esto se ilustra en Figure 9.8, que muestra la potencia de la prueba para un parámetro verdadero de $\theta = 0.7$ para todos los tamaños de muestra N desde 1 hasta 100, donde asumo que la hipótesis nula predice que $\theta_0 = 0.5$.

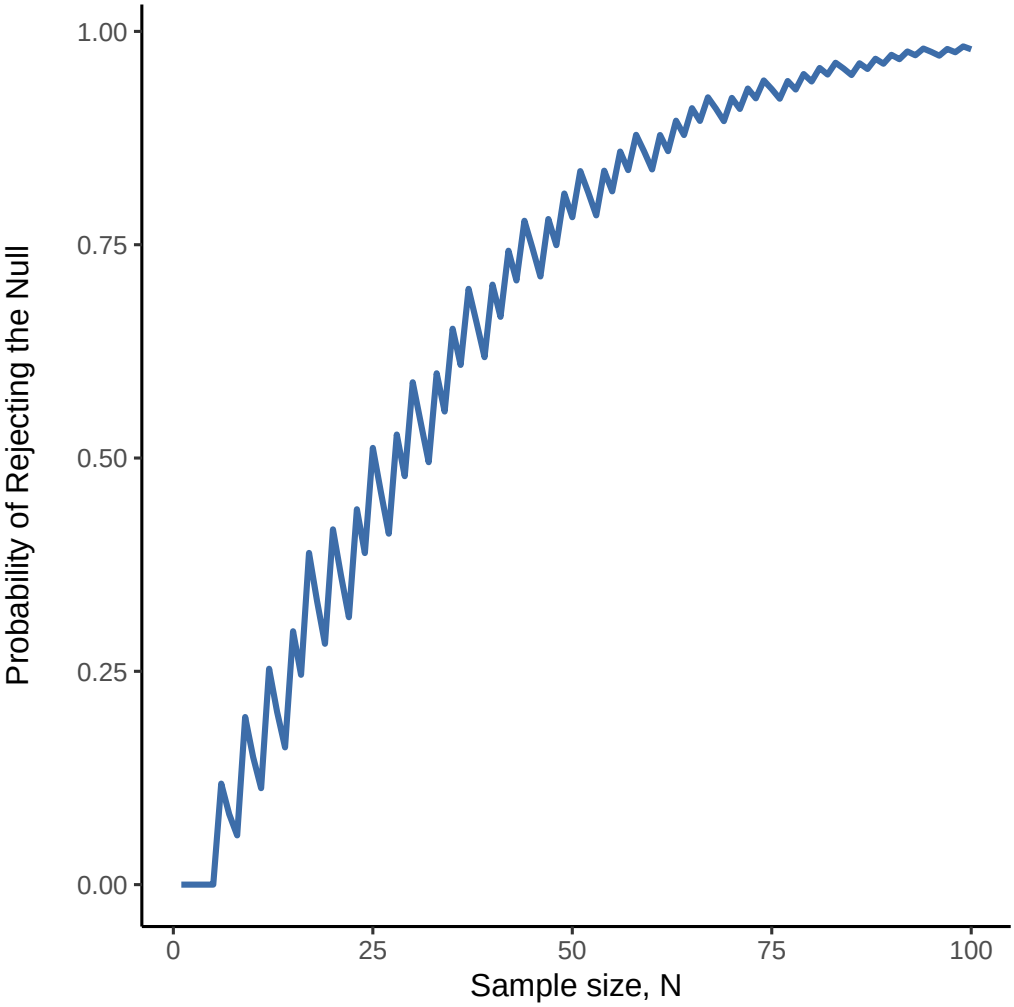


Figure 9.8: La potencia de nuestra prueba en función del tamaño de la muestra N. En este caso, el valor verdadero de θ es 0,7 pero la hipótesis nula es que $\theta = 0,5$. En general, N más grande significa mayor potencia. (Los pequeños zig-zags en esta función se producen debido a algunas interacciones extrañas entre θ, α y el hecho de que la distribución binomial es discreta, no importa para ningún propósito serio)

Dado que la potencia es importante, siempre que te plantees realizar un experimento, sería bastante útil saber cuánta potencia es probable que tengas. Nunca se puede saber con seguridad, ya que es imposible conocer el tamaño del efecto real. Sin embargo, a menudo (bueno, a veces) es posible adivinar cuál debería ser. Si es así, puedes adivinar

qué tamaño de muestra necesitas. Esta idea se llama **análisis de potencia**, y si es posible hacerlo, resulta muy útil. Puede decirte algo sobre si tienes suficiente tiempo o dinero para poder llevar a cabo el experimento con éxito. Cada vez es más frecuente ver a gente que defiende que el análisis de potencia debería ser una parte obligatoria del diseño experimental, por lo que merece la pena conocerlo. Sin embargo, no hablo del análisis de potencia en este libro. Esto es en parte por una razón aburrida y en parte por una razón sustantiva. La razón aburrida es que todavía no he tenido tiempo de escribir sobre el análisis de potencia. La sustantiva es que todavía desconfío un poco del análisis de potencia. Hablando como investigadora, muy rara vez me he encontrado en situación de poder hacer uno. O bien (a) mi experimento es un poco atípico y no sé cómo definir el tamaño del efecto correctamente, o (b) literalmente tengo tan poca idea sobre cuál será el tamaño del efecto que no sabría cómo interpretar las respuestas. No solo eso, después de extensas conversaciones con alguien que se gana la vida haciendo consultoría estadística (mi esposa, por cierto), no puedo evitar darme cuenta de que en la práctica, la única vez que alguien le pide un análisis de potencia es cuando está ayudando a alguien a escribir una solicitud de subvención. En otras palabras, la única vez que un científico parece querer un análisis de potencia en la vida real es cuando se ve obligados a hacerlo por un proceso burocrático. No forma parte del trabajo diario de nadie. En resumen, siempre he sido de la opinión de que, aunque la potencia es un concepto importante, el análisis de potencia no es tan útil como la gente lo hace parecer, excepto en los raros casos en los que (a) alguien ha descubierto cómo calcular la potencia para tu diseño experimental real y (b) tienes una idea bastante buena de cuál es probable que sea el tamaño del efecto.[12] Tal vez otras personas hayan tenido mejores experiencias que yo, pero personalmente nunca he estado en una situación en la que tanto (a) como (b) fueran ciertas. Puede que en el futuro me convenzan de lo contrario, y probablemente una versión futura de este libro incluya una discusión más detallada del análisis de potencia, pero por ahora esto es todo lo que puedo decir sobre el tema.

9.9 Algunas cuestiones a tener en cuenta

Lo que te he descrito en este capítulo es el marco ortodoxo de las pruebas de significación de hipótesis nula (PSHN). Comprender cómo funciona PSHN es una necesidad absoluta porque ha sido el enfoque dominante de la estadística inferencial desde que cobró importancia a principios del siglo XX. Es en lo que la gran mayoría de los científicos en activo confían para sus análisis de datos, por lo que incluso si lo odias, debes conocerlo. Sin embargo, el enfoque no está exento de problemas. Hay una serie de peculiaridades en el marco, rarezas históricas sobre cómo llegó a ser, disputas teóricas sobre si el marco es correcto o no, y muchas trampas prácticas para los incautos. No voy a entrar en muchos detalles sobre este tema, pero creo que vale la pena discutir brevemente algunas de estas cuestiones.

[12] Una posible excepción es cuando se estudia la efectividad de un nuevo tratamiento médico y se especifica de antemano cuál sería un tamaño de efecto importante de detectar, por ejemplo, por encima de cualquier tratamiento existente. De esta forma se puede obtener cierta información sobre el valor potencial de un nuevo tratamiento.

9.9.1 Neyman contra Fisher

Lo primero que debe tener en cuenta es que la PSHN ortodoxa es en realidad una combinación de dos enfoques bastante diferentes para las pruebas de hipótesis, uno propuesto por Sir Ronald Fisher y el otro por Jerzy Neyman (ver Lehmann (2011) para un resumen histórico). La historia es confusa porque Fisher y Neyman eran personas reales cuyas opiniones cambiaron con el tiempo, y en ningún momento ninguno de ellos ofreció "la declaración definitiva" de cómo debemos interpretar su trabajo muchas décadas después. Dicho esto, he aquí un rápido resumen de lo que considero que son estos dos enfoques.

Primero, hablemos del enfoque de Fisher. Hasta donde yo sé, Fisher suponía que solo se tenía una hipótesis (la nula) y que lo que se quería hacer era averiguar si la hipótesis nula era inconsistente con los datos. Desde su perspectiva, lo que deberías hacer es comprobar si los datos son "suficientemente improbables" según la nula. De hecho, si recuerdas nuestra discusión anterior, así es como Fisher define el valor p. Según Fisher, si la hipótesis nula proporcionara una explicación muy pobre de los datos, entonces podrías rechazarla con seguridad. Pero, como no tenemos ninguna otra hipótesis con la que compararla, no hay forma de "aceptar la alternativa" porque no tenemos necesariamente una alternativa explícita. Eso es más o menos todo.

Por el contrario, Neyman pensaba que el objetivo de las pruebas de hipótesis era servir de guía para la acción y su enfoque era algo más formal que el de Fisher. Su punto de vista era que hay varias cosas que se pueden hacer (aceptar la nula o aceptar la alternativa) y el objetivo de la prueba era decir cuál es compatible con los datos. Desde esta perspectiva, es fundamental especificar correctamente la hipótesis alternativa. Si no se sabe cuál es la hipótesis alternativa, entonces no sabe lo potente que es la prueba, ni siquiera qué acción tiene sentido. Su marco requiere realmente una competición entre diferentes hipótesis. Para Neyman, el valor p no medía directamente la probabilidad de los datos (o datos más extremos) bajo la nula, era más una descripción abstracta sobre qué "posibles pruebas" te decían que aceptaras la nula, y qué "posibles pruebas" te decían que aceptaras la alternativa.

Como puedes ver, lo que tenemos hoy es una mezcla extraña de los dos. Hablamos de tener tanto una hipótesis nula como una alternativa (Neyman), pero generalmente [13] definimos el valor de p en términos de datos extremos (Fisher), pero seguimos teniendo α valores (Neyman). Algunas de las pruebas estadísticas han especificado explícitamente alternativas (Neyman), pero otras son bastante vagas al respecto (Fisher). Y, según algunas personas al menos, no se nos permite hablar de aceptar la alternativa (Fisher). Es un lío, pero espero que esto al menos explique por qué es un lío.

9.9.2 Bayesianos versus frecuentistas

Anteriormente en este capítulo, fui bastante enfática sobre el hecho de que *no* puedes interpretar el valor p como la probabilidad de que la hipótesis nula sea verdadera. PSHN es fundamentalmente una herramienta frecuentista (consulta Chapter 7) y, como tal, no permite asignar probabilidades a las hipótesis. La hipótesis nula es cierta o no lo es. El enfoque bayesiano de la estadística interpreta la probabilidad como un grado de creencia, por lo que es totalmente correcto decir que existe una probabilidad del 10% de que la

[13] Aunque este libro describe la definición del valor de p tanto de Neyman como de Fisher, la mayoría no lo hace. La mayoría de los libros de texto introductorios solo le darán la versión de Fisher.

hipótesis nula sea cierta. Eso es solo un reflejo del grado de confianza que tienes en esta hipótesis. Esto no está permitido dentro del enfoque frecuentista. Recuerda, si eres frecuentista, una probabilidad solo se puede definir en términos de lo que sucede después de un gran número de repeticiones independientes (es decir, una frecuencia de largo plazo). Si esta es tu interpretación de la probabilidad, hablar de la "probabilidad" de que la hipótesis nula sea cierta es un completo galimatías: una hipótesis nula o es verdadera o es falsa. Es imposible hablar de una frecuencia de largo plazo para esta afirmación. Hablar de "la probabilidad de la hipótesis nula" es tan absurdo como "el color de la libertad". No tiene uno.

Lo más importante es que no se trata de una cuestión puramente ideológica. Si decides que eres bayesiana y que te parece bien hacer afirmaciones probabilísticas sobre hipótesis, tienes que seguir las reglas bayesianas para calcular esas probabilidades. Hablaré más sobre esto en Chapter 16, pero por ahora lo que quiero señalarte es que el valor p es una terrible aproximación a la probabilidad de que H_0 sea cierta. Si lo que quieres saber es la probabilidad de la nula, ¡entonces el valor p no es lo que estás buscando!

9.9.3 Trampas

Como puedes ver, la teoría que subyace a las pruebas de hipótesis es un lío, e incluso ahora hay discusiones en estadística sobre cómo "debería" funcionar. Sin embargo, los desacuerdos entre los estadísticos no son nuestra verdadera preocupación aquí. Nuestra verdadera preocupación es el análisis práctico de datos. Y aunque el enfoque "ortodoxo" de la prueba de significancia de la hipótesis nula tiene muchos inconvenientes, incluso una bayesiana impenitente como yo estaría de acuerdo en que pueden ser útiles si se usan de manera responsable. La mayoría de las veces dan respuestas sensatas y se pueden utilizar para aprender cosas interesantes. Dejando a un lado las diversas ideologías y confusiones históricas que hemos discutido, el hecho es que el mayor peligro en toda la estadística es la *irreflexión*. No me refiero a la estupidez, sino literalmente a la irreflexión. La prisa por interpretar un resultado sin dedicar tiempo a pensar qué dice realmente cada prueba sobre los datos y comprobar si es coherente con la interpretación que se ha hecho. Ahí es donde está la mayor trampa.

Para dar un ejemplo de esto, considera el siguiente ejemplo (ver Gelman & Stern (2006)). Supongamos que estoy realizando mi estudio sobre PES y he decidido analizar los datos por separado para los participantes masculinos y femeninos. De los participantes masculinos, 33 de 50 adivinaron correctamente el color de la carta. Se trata de un efecto significativo ($p = .03$). Las mujeres acertaron 29 de cada 50. No es un efecto significativo ($p = .32$). Al observar esto, es muy tentador que la gente empiece a preguntarse por qué existe una diferencia entre hombres y mujeres en cuanto a sus habilidades psíquicas. Sin embargo, esto es erróneo. Si lo piensas bien, en realidad no hemos realizado una prueba que compare explícitamente a los hombres con las mujeres. Todo lo que hemos hecho es comparar a los hombres con el azar (la prueba binomial fue significativa) y comparar a las mujeres con el azar (la prueba binomial no fue significativa). Si queremos argumentar que hay una diferencia real entre los hombres y las mujeres, probablemente deberíamos realizar una prueba de la hipótesis nula de que no hay diferencia. Podemos hacerlo usando una prueba de hipótesis diferente,[14] pero cuando lo hacemos resulta que no tenemos pruebas de que los hombres y las mujeres sean significativamente diferentes ($p = .54$). ¿Crees que hay alguna diferencia fundamental entre los dos grupos? Por

[14] En este caso, la prueba de independencia ji-cuadrado de Pearson (ver Chapter 10)

supuesto que no. Lo que sucedió aquí es que los datos de ambos grupos (hombres y mujeres) están bastante en el límite. Por pura casualidad, uno de ellos acabó en el lado mágico de la línea $p = .05$, y el otro no. Eso no implica que los hombres y las mujeres sean diferentes. Este error es tan común que siempre hay que tener cuidado con él. La diferencia entre significativo y no significativo no es prueba de una diferencia real. Si quieres decir que hay una diferencia entre dos grupos, tienes que probar esa diferencia.

El ejemplo anterior es solo eso, un ejemplo. Lo he seleccionado porque es muy común, pero lo más importante es que el análisis de datos puede ser difícil de hacer bien. Piensa qué es lo que quieres probar, por qué quieres probarlo y si las respuestas que da tu prueba podrían tener algún sentido en el mundo real.

9.10 Resumen

Las pruebas de hipótesis nulas son uno de los elementos más ubicuos de la teoría estadística. La inmensa mayoría de artículos científicos presentan los resultados de una u otra prueba de hipótesis. Como consecuencia, es casi imposible desenvolverse en el mundo de la ciencia sin tener al menos una comprensión superficial de lo que significa un valor p, lo que hace que este sea uno de los capítulos más importantes del libro. Como de costumbre, terminaré el capítulo con un resumen rápido de las ideas clave de las que hemos hablado:

- Una colección de hipótesis. Hipótesis de investigación e hipótesis estadísticas. Hipótesis nula y alternativa.
- Dos tipos de errores. Tipo I y Tipo II.
- [Estadísticas de prueba y distribuciones muestrales].
- Contraste de hipótesis para [Tomar decisiones]
- El valor p de una prueba. valores p como decisiones “suaves”
- [Comunicar los resultados de una prueba de hipótesis]
- [Ejecución de la prueba de hipótesis en la práctica]
- Tamaño del efecto, tamaño de la muestra y potencia
- [Algunos temas a considerar] con respecto a la prueba de hipótesis

Más adelante en el libro, en Chapter 16, revisaré la teoría de las pruebas de hipótesis nulas desde una perspectiva bayesiana y presentaré una serie de herramientas nuevas que puedes usar si no te gusta mucho el enfoque ortodoxo. Pero, por ahora, hemos terminado con la teoría estadística abstracta y podemos empezar a hablar de herramientas específicas de análisis de datos.

Part V

Instrumentos estadística

Chapter 10

Análisis de datos categóricos

Ahora que hemos cubierto la teoría básica de las pruebas de hipótesis, es hora de comenzar a buscar pruebas específicas que se usan habitualmente en psicología. ¿Por dónde empezar? No todos los libros de texto se ponen de acuerdo sobre por dónde empezar, pero yo voy a empezar con "χ^2 tests" (este capítulo, pronunciado "chi-square"[1] y " pruebas t" en Chapter 11). Ambas herramientas se usan con mucha frecuencia en la práctica científica, y aunque no son tan potentes como la "regresión" y el "análisis de varianza" que trataremos en capítulos posteriores, son mucho más fáciles de entender.

El término "datos categóricos" no es más que otro nombre para "datos de escala nominal". No es nada que no hayamos discutido ya, sólo que en el contexto del análisis de datos, la gente tiende a usar el término "datos categóricos" en lugar de "datos de escala nominal". No sé por qué. En cualquier caso, **análisis de datos categóricos** se refiere a una colección de herramientas que puedes usar cuando tus datos son de escala nominal. Sin embargo, hay muchas herramientas diferentes que se pueden usar para el análisis de datos categóricos, y este capítulo cubre solo algunas de las más comunes.

10.1 La prueba de bondad de ajuste χ^2 (ji-cuadrado)

La prueba de bondad de ajuste χ^2 es una de las pruebas de hipótesis más antiguas que existen. Fue inventada por Karl Pearson alrededor del cambio de siglo (Pearson, 1900), con algunas correcciones hechas más tarde por Sir Ronald Fisher (Fisher, 1922a). Comprueba si una distribución de frecuencias observadas de una variable nominal coincide con una distribución de frecuencias esperadas. Por ejemplo, supongamos que un grupo de pacientes se sometió a un tratamiento experimental y se les evaluó la salud para ver si su condición mejoró, permaneció igual o empeoró. Se podría usar una prueba de bondad de ajuste para determinar si los números en cada categoría (mejorado, sin cambios, empeorado) coinciden con los números que se esperarían dada la opción de tratamiento estándar. Pensemos en esto un poco más, con algo de psicología.

[1] también conocido como "ji-cuadrado".

10.1.1 Los datos de las cartas

A lo largo de los años, se han realizado muchos estudios que muestran que a los humanos les resulta difícil simular la aleatoriedad. Por mucho que intentemos "actuar" al azar, pensamos en términos de patrones y estructura y, por lo tanto, cuando se nos pide que "hagamos algo al azar", lo que la gente realmente hace es cualquier cosa menos aleatorio. Como consecuencia, el estudio de la aleatoriedad humana (o la no aleatoriedad, según sea el caso) abre muchas preguntas psicológicas profundas sobre cómo pensamos sobre el mundo. Con esto en mente, consideremos un estudio muy simple. Supongamos que le pido a la gente que imagine un mazo de cartas barajado y que elija mentalmente una carta de este mazo imaginario "al azar". Después de que hayan elegido una carta, les pido que seleccionen mentalmente una segunda. Para ambas opciones, lo que vamos a ver es el palo (corazones, tréboles, picas o diamantes) que eligió la gente. Después de pedir, digamos, $N = 200$ personas que haga esto, me gustaría ver los datos y averiguar si las cartas que las personas pretendían seleccionar eran realmente aleatorias o no. Los datos están contenidos en el archivo randomness.csv en el que, cuando lo abres en jamovi y echas un vistazo a la vista de hoja de cálculo, verás tres variables. Estas son: una variable id que asigna un identificador único a cada participante, y las dos variables elección_1 y elección_2 que indican los palos de cartas que eligieron las personas.

Por el momento, concentrémonos en la primera elección que hizo la gente. Usaremos la opción Tablas de frecuencia en 'Exploración' - 'Descriptivos' para contar la cantidad de veces que observamos a las personas elegir cada palo. Esto es lo que obtenemos (Table 10.1):

Table 10.1: Número de veces que se eligió cada palo

clubs	diamonds	hearts	spades
35	51	64	50

Esa pequeña tabla de frecuencias es bastante útil. Al mirarla, hay un indicio de que es más probable que la gente elija corazones que tréboles, pero no es del todo obvio si eso es cierto o si se debe al azar. Así que probablemente tendremos que hacer algún tipo de análisis estadístico para averiguarlo, que es de lo que voy a hablar en la siguiente sección.

Excelente. A partir de este momento, trataremos esta tabla como los datos que buscamos analizar. Sin embargo, dado que voy a tener que hablar sobre estos datos en términos matemáticos (¡lo siento!), podría ser una buena idea aclarar cuál es la notación. En notación matemática, acortamos la palabra legible por humanos "observado" a la letra O, y usamos subíndices para indicar la posición de la observación. Entonces, la segunda observación en nuestra tabla se escribe como O_2 en matemáticas. La relación entre las descripciones en español y los símbolos matemáticos se ilustra en Table 10.2.

Espero que haya quedado claro. También vale la pena señalar que los matemáticos prefieren hablar de cosas generales en lugar de específicas, por lo que también verás la notación O_i, que se refiere al número de observaciones que se encuentran dentro de la i-ésima categoría (donde i podría ser 1, 2, 3 o 4). Finalmente, si queremos referirnos al conjunto de todas las frecuencias observadas, los estadísticos agrupan todos los valores observados en un vector [2], al que me referiré como O.

[2] un vector es una secuencia de elementos de datos del mismo tipo básico.

Table 10.2: Relación entre las descripciones en español y los símbolos matemáticos

label	index, i	math. symbol	the value
clubs, ♣	1	O_1	35
diamonds, ♢	2	O_2	51
hearts, ♡	3	O_3	64
spades, ♠	4	O_4	50

$$O = (O_1, O_2, O_3, O_4)$$

Una vez más, esto no es nada nuevo o interesante. Es solo notación. Si digo que $O = (35, 51, 64, 50)$ todo lo que estoy haciendo es describir la tabla de frecuencias observadas (es decir, observadas), pero me estoy refiriendo a ella usando notación matemática.

10.1.2 La hipótesis nula y la hipótesis alternativa

Como se indicó en la última sección, nuestra hipótesis de investigación es que "la gente no elige cartas al azar". Lo que vamos a querer hacer ahora es traducir esto en algunas hipótesis estadísticas y luego construir una prueba estadística de esas hipótesis. La prueba que te voy a describir es la prueba de bondad de ajuste de **Pearson** χ^2 (ji-cuadrado) y, como ocurre a menudo, tenemos que comenzar construyendo cuidadosamente nuestra hipótesis nula. En este caso, es bastante fácil. Primero, expresemos la hipótesis nula en palabras:

$$H_0 : \text{Los cuatro palos se eligen con la misma probabilidad}$$

Ahora, debido a que esto es estadística, tenemos que poder decir lo mismo de manera matemática. Para hacer esto, usemos la notación P_j para referirnos a la verdadera probabilidad de que se elija el j-ésimo palo. Si la hipótesis nula es verdadera, entonces cada uno de los cuatro palos tiene un 25% de posibilidades de ser seleccionado. En otras palabras, nuestra hipótesis nula afirma que $P_1 = .25$, $P_2 = .25$, $P3 = .25$ y finalmente que $P_4 = .25$. Sin embargo, de la misma manera que podemos agrupar nuestras frecuencias observadas en un vector O que resume todo el conjunto de datos, podemos usar P para referirnos a las probabilidades que corresponden a nuestra hipótesis nula. Entonces, si permito que el vector $P = (P_1, P_2, P_3, P_4)$ se refiera a la colección de probabilidades que describen nuestra hipótesis nula, entonces tenemos:

$$H_0 : P = (.25, .25, .25, .25)$$

En este caso particular, nuestra hipótesis nula corresponde a un vector de probabilidades P en el que todas las probabilidades son iguales entre sí. Pero esto no tiene por qué ser así. Por ejemplo, si la tarea experimental fuera que las personas imaginaran que estaban sacando de una baraja que tenía el doble de tréboles que cualquier otro palo, entonces la hipótesis nula correspondería a algo así como $P = (.4, .2, .2, .2)$. Mientras las probabilidades sean todas positivas y sumen 1, entonces es una opción perfectamente

legítima para la hipótesis nula. Sin embargo, el uso más común de la prueba de bondad de ajuste es probar la hipótesis nula de que todas las categorías tienen la misma probabilidad, por lo que nos ceñiremos a eso para nuestro ejemplo.

¿Qué pasa con nuestra hipótesis alternativa, H_1? Todo lo que realmente nos interesa es demostrar que las probabilidades involucradas no son todas idénticas (es decir, las elecciones de las personas no fueron completamente aleatorias). En consecuencia, las versiones "humanas" de nuestras hipótesis son las siguientes:

H_0 : Los cuatro palos se eligen con la misma probabilidad H_1 : Al menos una de las probabilidades de elec

… y la versión "apta para matemáticos" es:

$H_0 : P = (.25, .25, .25, .25)$ $H_1 : P \neq (.25, .25, .25, .25)$

10.1.3 La prueba estadística de "bondad de ajuste"

En este punto, tenemos nuestras frecuencias observadas O y una colección de probabilidades P correspondientes a la hipótesis nula que queremos probar. Lo que ahora queremos hacer es construir una prueba de la hipótesis nula. Como siempre, si queremos probar H_0 contra H_1, vamos a necesitar una prueba estadística. El truco básico que utiliza una prueba de bondad de ajuste es construir una prueba estadística que mida cuán "cercanos" están los datos a la hipótesis nula. Si los datos no se parecen a lo que "esperaría" ver si la hipótesis nula fuera cierta, entonces probablemente no sea cierta. Bien, si la hipótesis nula fuera cierta, ¿qué esperaríamos ver? O, para usar la terminología correcta, ¿cuáles son las **frecuencias esperadas**? Hay $N = 200$ observaciones y (si el valor nulo es verdadero) la probabilidad de que cualquiera de ellas elija un corazón es $P_3 = 0,25$, así que supongo que esperamos $200x0,25 = 50$ corazones, ¿verdad? O, más específicamente, si dejamos que Ei se refiera a "el número de respuestas de la categoría i que esperamos si el valor nulo es verdadero", entonces

$$E_i = N \times P_i$$

Esto es bastante fácil de calcular. Si hay 200 observaciones que pueden clasificarse en cuatro categorías, y pensamos que las cuatro categorías son igualmente probables, entonces, en promedio, esperaríamos ver 50 observaciones en cada categoría, ¿verdad?

Ahora, ¿cómo traducimos esto en una prueba estadística? Claramente, lo que queremos hacer es comparar el número esperado de observaciones en cada categoría (E_i) con el número observado de observaciones en esa categoría (O_i). Y sobre la base de esta comparación deberíamos poder llegar a una buena prueba estadística. Para empezar, calculemos la diferencia entre lo que la hipótesis nula esperaba que encontráramos y lo que realmente encontramos. Es decir, calculamos la puntuación de diferencia "observada menos esperada", $O_i - E_i$. Esto se ilustra en Table 10.3.

Así, según nuestros cálculos, está claro que la gente eligió más corazones y menos tréboles de lo que predijo la hipótesis nula. Sin embargo, un momento de reflexión sugiere que estas diferencias en bruto no son exactamente lo que estamos buscando. Intuitivamente, parece que es tan malo cuando la hipótesis nula predice muy pocas observaciones (que es lo que sucedió con los corazones) como cuando predice demasiadas (que es lo que sucedió con los tréboles). Entonces es un poco extraño que tengamos un número negativo para los tréboles y un número positivo para los corazones. Una manera fácil de arreglar esto

Table 10.3: frecuencias esperadas y observadas

	♣	♢	♡	♠
expected frequency E_i	50	50	50	50
observed frequency O_i	35	51	64	50
difference score $O_i - E_i$	-15	1	14	0

es elevar todo al cuadrado, de modo que ahora calculemos las diferencias al cuadrado, $(E_i - O_i)^2$. Como antes, podemos hacer esto a mano (Table 10.4).

Table 10.4: elevar al cuadrado las diferencias en las puntuaciones

♣	♢	♡	♠
225	1	196	0

Ahora estamos progresando. Lo que tenemos ahora es una colección de números que son grandes cuando la hipótesis nula hace una mala predicción (tréboles y corazones), pero son pequeños cuando hace una buena (diamantes y picas). A continuación, por algunas razones técnicas que explicaré en un momento, también dividamos todos estos números por la frecuencia esperada Ei, de modo que en realidad estemos calculando $\frac{(E_i - O_i)^2}{E_i}$. Dado que $E_i = 50$ para todas las categorías en nuestro ejemplo, no es un cálculo muy interesante, pero hagámoslo de todos modos (Table 10.5).

Table 10.5: dividir las diferencias de puntuaciones al cuadrado por la frecuencia esperada para proporcionar una puntuación de 'error'

♣	♢	♡	♠
4.50	0.02	3.92	0.00

En efecto, lo que tenemos aquí son cuatro puntuaciones de "error" diferentes, cada una de las cuales nos indica la magnitud del "error" que cometió la hipótesis nula cuando intentamos usarla para predecir nuestras frecuencias observadas. Entonces, para convertir esto en una prueba estadística útil, una cosa que podríamos hacer es simplemente sumar estos números. El resultado se denomina estadístico de **bondad de ajuste**, conocido convencionalmente como χ^2 (ji-cuadrado) o GOF. Podemos calcularlo como en Table 10.6.

$$\sum((observado - esperado)^2/esperado)$$

Esto nos da un valor de 8,44.

[Detalle técnico adicional [3]]

[3] si hacemos que k se refiera al número total de categorías (es decir, k = 4 para los datos de nuestras

Como hemos visto en nuestros cálculos, en nuestro conjunto de datos de cartas tenemos un valor de $\chi^2 = 8{,}44$. Entonces, ahora la pregunta es si este es un valor lo suficientemente grande como para rechazar el la hipótesis nula.

10.1.4 La distribución muestral del estadístico GOF

Para determinar si un valor particular de χ^2 es o no lo suficientemente grande como para justificar el rechazo de la hipótesis nula, necesitaremos averiguar cuál sería la distribución muestral para χ^2 si la hipótesis nula fuera cierta. Así que eso es lo que voy a hacer en esta sección. Te mostraré con bastante detalle cómo se construye esta distribución muestral y luego, en la siguiente sección, la usaré para construir una prueba de hipótesis. Si quieres ir al grano y estás dispuesta a confiar en que la distribución muestral es una distribución χ^2 (ji-cuadrado) con $k-1$ grados de libertad, puedes omitir el resto de esta sección. Sin embargo, si deseas comprender *por qué* la prueba de bondad de ajuste funciona de la forma en que lo hace, sigue leyendo.

Bien, supongamos que la hipótesis nula es realmente cierta. Si es así, entonces la verdadera probabilidad de que una observación caiga en la i-ésima categoría es P_i. Después de todo, esa es más o menos la definición de nuestra hipótesis nula. Pensemos en lo que esto realmente significa. Esto es como decir que la "naturaleza" toma la decisión sobre si la observación termina o no en la categoría i al lanzar una moneda ponderada (es decir, una donde la probabilidad de obtener cara es P_j). Y, por lo tanto, podemos pensar en nuestra frecuencia observada O_i imaginando que la naturaleza lanzó N de estas monedas (una para cada observación en el conjunto de datos), y exactamente O_i de ellas salieron cara. Obviamente, esta es una forma bastante extraña de pensar en el experimento. Pero lo que hace (espero) es recordarte que en realidad hemos visto este escenario antes. Es exactamente la misma configuración que dio lugar a Section 7.4 en Chapter 7. En otras palabras, si la hipótesis nula es verdadera, se deduce que nuestras frecuencias observadas se generaron muestreando a partir de una distribución binomial:

$$O_i \sum Binomial(P_i, N)$$

Ahora bien, si recuerdas nuestra discusión sobre Section 8.3.3, la distribución binomial empieza a parecerse bastante a la distribución normal, especialmente cuando N es grande y cuando P_i no está demasiado cerca a 0 o 1. En otras palabras, siempre que N_i^P sea lo suficientemente grande. O, dicho de otro modo, cuando la frecuencia esperada Ei es lo suficientemente grande, entonces la distribución teórica de O_i es aproximadamente normal. Mejor aún, si O_i se distribuye normalmente, entonces también lo es $(O_i - E_i)/\sqrt{(E_i)}$. Dado que E_i es un valor fijo, restando Ei y dividiendo por ? Ei cambia la media y la desviación estándar de la distribución normal, pero eso es todo lo que hace. Bien, ahora echemos un vistazo a cuál es realmente nuestro estadístico de bondad de ajuste. Lo que estamos haciendo es tomar un montón de cosas que están normalmente distribuidas, elevarlas al cuadrado y sumarlas. Espera. ¡También lo

cartas), entonces el estadístico χ^2 está dado por:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Intuitivamente, está claro que si chi^2 es pequeño, entonces los datos observados Oi están muy cerca de lo que predijo la hipótesis nula E_i, por lo que vamos a necesitar un gran estadístico χ^2 para rechazar la hipótesis nula.

hemos visto antes! Como discutimos en la sección sobre Section 7.6, cuando tomas un montón de cosas que tienen una distribución normal estándar (es decir, media 0 y desviación estándar 1), las elevas al cuadrado y luego las sumas, la cantidad resultante tiene una distribución ji-cuadrado. Así que ahora sabemos que la hipótesis nula predice que la distribución muestral del estadístico de bondad de ajuste es una distribución de ji-cuadrado. Genial.

Hay un último detalle del que hablar, a saber, los grados de libertad. Si recuerdas Section 7.6, dije que si el número de cosas que está sumando es k, entonces los grados de libertad para la distribución de ji-cuadrado resultante es k. Sin embargo, lo que dije al comienzo de esta sección es que los grados de libertad reales para la prueba de bondad de ajuste de ji-cuadrado son $k - 1$. ¿Por qué? La respuesta aquí es que lo que se supone que estamos mirando es el número de cosas realmente independientes que se suman. Y, como continuaré hablando en la siguiente sección, aunque hay k cosas que estamos agregando solo $k - 1$ de ellas son realmente independientes, por lo que los grados de libertad en realidad son solo $k - 1$. Ese es el tema de la siguiente sección[4].

10.1.5 Grados de libertad

Cuando introduje la distribución de ji-cuadrado en Section 7.6, fui un poco imprecisa sobre lo que **"grados de libertad"** significa realmente. Obviamente, es importante. Si observamos Figure 10.1, podemos ver que si cambiamos los grados de libertad, la distribución de ji-cuadrado cambia de forma bastante sustancial. ¿Pero qué es exactamente? Una vez más, cuando presenté la distribución y expliqué su relación con la distribución normal, ofrecí una respuesta: es el número de "variables normalmente distribuidas" que estoy elevando al cuadrado y sumando. Pero, para la mayoría de las personas, eso es algo abstracto y no del todo útil. Lo que realmente necesitamos hacer es tratar de comprender los grados de libertad en términos de nuestros datos. Así que aquí va.

La idea básica detrás de los grados de libertad es bastante sencilla. Se calculan contando el número de "cantidades" distintas que se utilizan para describir los datos y restando todas las "restricciones" que esos datos deben satisfacer.[5] Esto es un poco vago, así que usemos los datos de nuestras cartas como un ejemplo concreto. Describimos nuestros datos utilizando cuatro números, $O1, O2, O3$ y O4 correspondientes a las frecuencias observadas de las cuatro categorías diferentes (corazones, tréboles, diamantes, picas). Estos cuatro números son los resultados aleatorios de nuestro experimento. Pero mi experimento en realidad tiene una restricción fija incorporada: el tamaño de la muestra

[4] si reescribes la ecuación para el estadístico de bondad de ajuste como una suma de k - 1 cosas independientes, obtienes la distribución muestral "adecuada", que es ji-cuadrado con k - 1 grados de libertad. Está fuera del alcance de un libro introductorio mostrar las matemáticas con tanto detalle. Todo lo que quería hacer es darte una idea de por qué el estadístico de bondad de ajuste está asociado con la distribución de ji-cuadrado.

[5] Me siento obligada a señalar que esto es una simplificación excesiva. Funciona bien en bastantes situaciones, pero de vez en cuando nos encontraremos con valores de grados de libertad que no son números enteros. No dejes que esto te preocupe demasiado; cuando te encuentres con esto, recuerda que los "grados de libertad" son en realidad un concepto un poco confuso, y que la bonita y simple historia que te estoy contando aquí no es toda la historia. Para una clase introductoria, por lo general es mejor ceñirse a la historia simple, pero creo que es mejor advertirte que esperes que esta historia simple se desmorone. Si no te hiciera esta advertencia, podrías comenzar a confundirte cuando veas $df = 3.4$ o algo así, pensando (incorrectamente) que has entendido mal algo de lo que te he enseñado en lugar de darte cuenta (correctamente) de que hay algo que no te he contado.

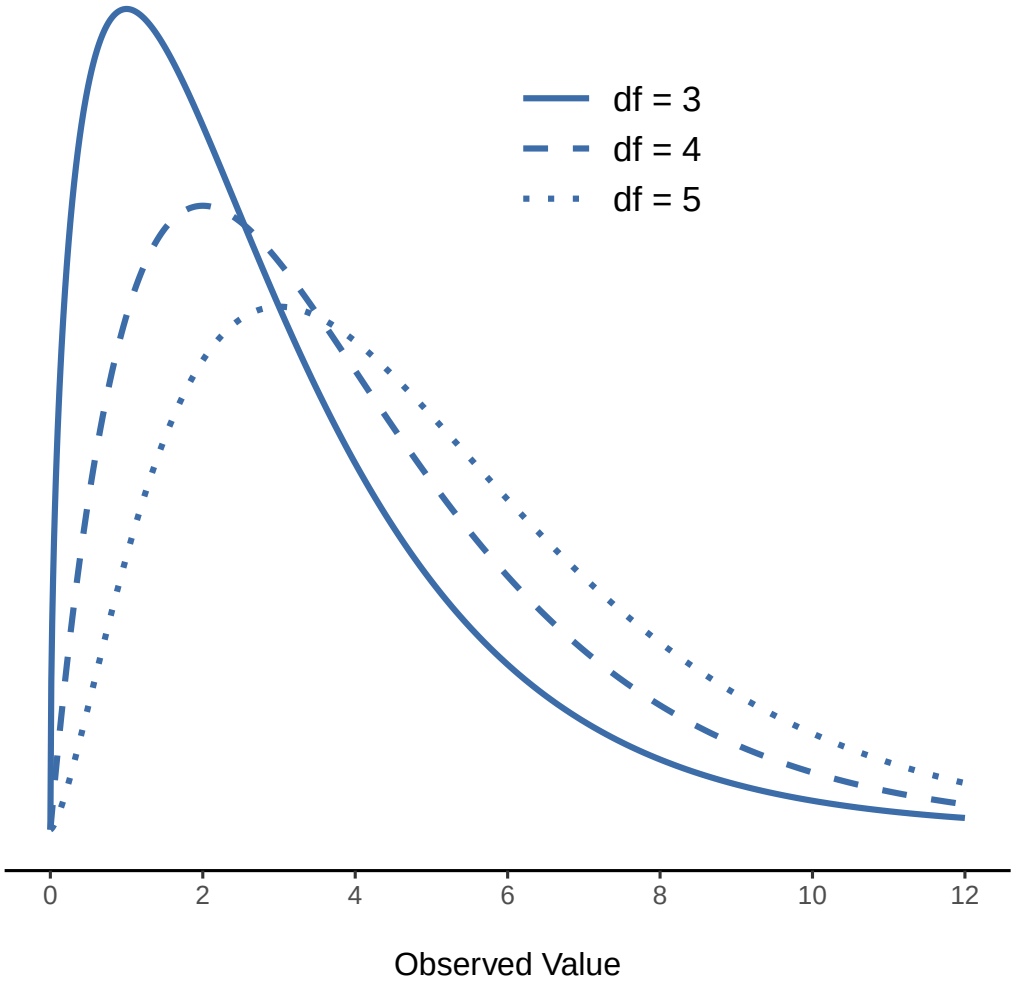


Figure 10.1: distribuciones χ^2 (ji-cuadrado) con diferentes valores para los 'grados de libertad'

N. [6] Es decir, si sabemos

cuántas personas eligieron corazones, cuántas eligieron diamantes y cuántas eligieron tréboles, entonces podríamos averiguar exactamente cuántas eligieron espadas. En otras palabras, aunque nuestros datos se describen usando cuatro números, en realidad solo corresponden a $4 - 1 = 3$ grados de libertad. Una forma ligeramente diferente de pensar al respecto es notar que hay cuatro probabilidades que nos interesan (nuevamente, correspondientes a las cuatro categorías diferentes), pero estas probabilidades deben sumar uno, lo que impone una restricción. Por lo tanto los grados de libertad son $4-1 = 3$. Independientemente de si deseas pensar en términos de frecuencias observadas o en términos de probabilidades, la respuesta es la misma. En general, cuando se ejecuta la prueba de bondad de ajuste χ^2(ji-cuadrado) para un experimento con k grupos, los grados de libertad serán $k - 1$.

10.1.6 Probando la hipótesis nula

El paso final en el proceso de construcción de nuestra prueba de hipótesis es averiguar cuál es la región de rechazo. Es decir, qué valores de χ^2 nos llevarían a rechazar la hipótesis nula. Como vimos anteriormente, los valores grandes de χ^2 implican que la hipótesis nula no ha hecho un buen trabajo al predecir los datos de nuestro experimento, mientras que los valores pequeños de χ^2 implican que en realidad se ha hecho bastante bien. Por lo tanto, una estrategia bastante sensata sería decir que hay algún valor crítico tal que si χ^2 es mayor que el valor crítico, rechazamos el valor nulo, pero si χ^2 es menor que este valor, mantenemos la hipótesis nula. En otras palabras, para usar el lenguaje que introdujimos en Chapter 9, la prueba de bondad de ajuste ji-cuadrado es siempre una **prueba unilateral**. Correcto, entonces todo lo que tenemos que hacer es averiguar cuál es este valor crítico. Y es bastante sencillo. Si queremos que nuestra prueba tenga un nivel de significación de $\alpha = .05$ (es decir, estamos dispuestas a tolerar una tasa de error Tipo I de 5), entonces tenemos que elegir nuestro valor crítico de modo que solo haya una probabilidad del 5% de que χ^2 pueda llegar a ser tan grande si la hipótesis nula es cierta. Esto se ilustra en Figure 10.2.

Ah, pero te escucho preguntar, ¿cómo encuentro el valor crítico de una distribución ji-cuadrado con $k-1$ grados de libertad? Hace muchos años, cuando tomé por primera vez una clase de estadística de psicología, solíamos buscar estos valores críticos en un libro de tablas de valores críticos, como el de Figure 10.3. Mirando esta figura, podemos ver que el valor crítico para una distribución χ^2 con 3 grados de libertad y p=0.05 es 7.815.

Así, si nuestro estadístico χ^2 calculado es mayor que el valor crítico de 7.815, entonces podemos rechazar la hipótesis nula (recuerda que la hipótesis nula, H_0, es que los cuatro palos se eligen con la misma probabilidad). Como en realidad ya lo calculamos antes (es decir, $\chi^2 = 8{,}44$), podemos rechazar la hipótesis nula. Y eso es todo, básicamente. Ahora conoces la "prueba de χ^2 de Pearson para la bondad de ajuste". Qué suerte tienes.

[6] en la práctica, el tamaño de la muestra no siempre es fijo. Por ejemplo, podemos ejecutar el experimento durante un período fijo de tiempo y la cantidad de personas que participan depende de cuántas personas se presenten. Eso no importa para los propósitos actuales.

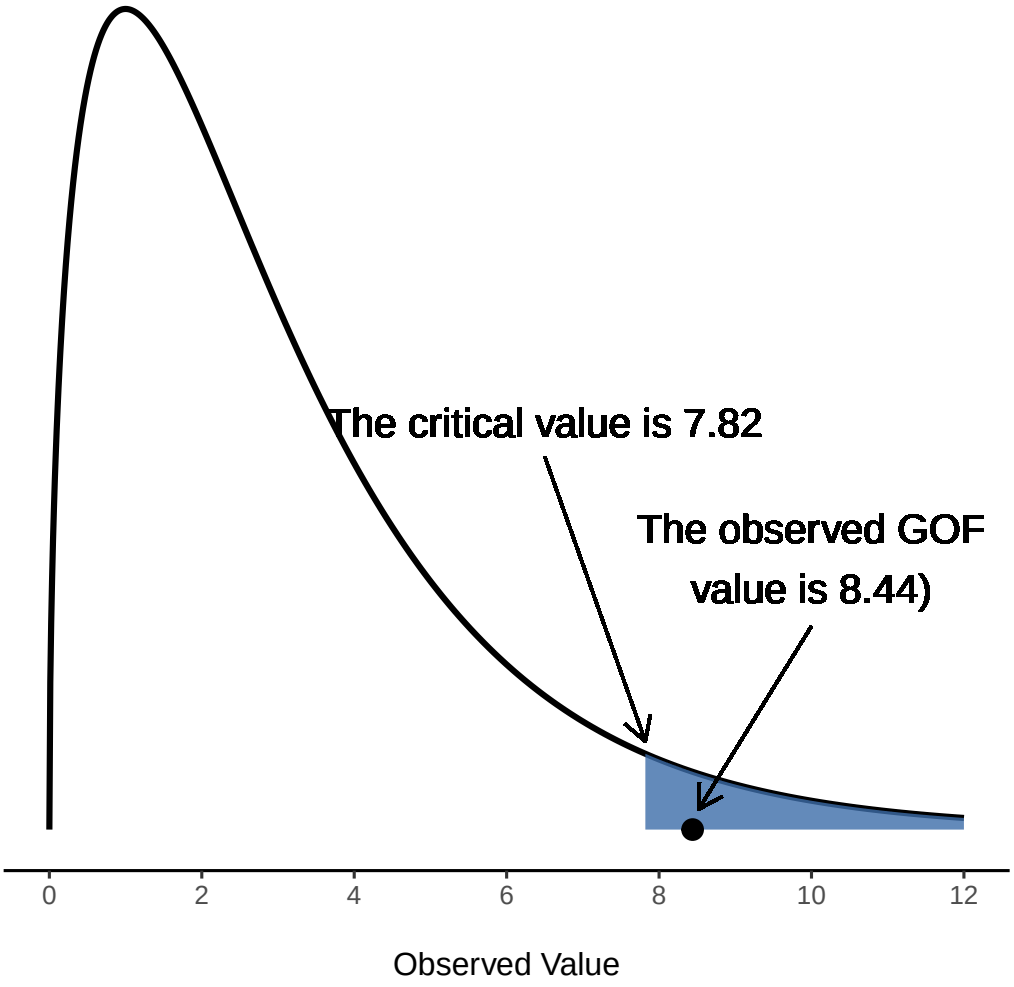


Figure 10.2: Ilustración de cómo funciona la prueba de hipótesis para la prueba de bondad de ajuste χ^2 (ji-cuadrado)

Degrees of Freedom	Probability								
	0.95	0.90	0.70	0.50	0.30	0.10	0.05	0.01	0.001
1	0.004	0.016	0.148	0.455	1.074	2.706	3.841	6.635	10.828
2	0.103	0.211	0.713	1.386	2.408	4.605	5.991	9.210	13.816
3	0.352	0.584	1.424	2.366	3.665	6.251	7.815	11.345	16.266
4	0.711	1.064	2.195	3.357	4.878	7.779	9.488	13.277	18.467
5	1.145	1.610	3.000	4.351	6.064	9.236	11.070	15.086	20.515
6	1.635	2.204	3.828	5.348	7.231	10.645	12.592	16.812	22.458
7	2.167	2.833	4.671	6.346	8.383	12.017	14.067	18.475	24.322
8	2.733	3.490	5.527	7.344	9.524	13.362	15.507	20.090	26.124
9	3.325	4.168	6.393	8.343	10.656	14.684	16.919	21.666	27.877
10	3.940	4.865	7.267	9.342	11.781	15.987	18.307	23.209	29.588
	Non-significant						Significant		

Figure 10.3: Tabla de valores críticos para la distribución ji-cuadrado

10.1.7 Haciendo la prueba en jamovi

No es sorprendente que jamovi proporcione un análisis que hará estos cálculos por ti. Usemos el archivo Randomness.omv. En la barra de herramientas principal de 'Análisis', selecciona 'Frecuencias' - 'Pruebas de proporción de una muestra' - 'N Resultados'. Luego, en la ventana de análisis que aparece, mueve la variable que deseas analizar (opción 1) al cuadro 'Variable'. Además, haz clic en la casilla de verificación 'Recuentos esperados' para que se muestren en la tabla de resultados. Cuando hayas terminado todo esto, deberías ver los resultados del análisis en jamovi como en Figure 10.4. No sorprende entonces que jamovi proporcione los mismos recuentos y estadísticos esperados que calculamos a mano anteriormente, con un valor de χ^2 de \$ (8.44\$ con 3 gl y $p = 0.038$. Ten en cuenta que ya no necesitamos buscar un valor umbral de valor p crítico, ya que jamovi nos da el valor p real del $\chi^c alculado2$ por 3

10.1.8 Especificando una hipótesis nula diferente

En este punto, es posible que te preguntes qué hacer si deseas realizar una prueba de bondad de ajuste, pero tu hipótesis nula no es que todas las categorías sean igualmente probables. Por ejemplo, supongamos que alguien hubiera hecho la predicción teórica de que las personas deberían elegir cartas rojas 60% del tiempo y cartas negras 40% del tiempo (no tengo ni idea de por qué predecirías eso), pero no tenía otras preferencias. Si ese fuera el caso, la hipótesis nula sería esperar que 30% de las opciones fueran corazones, 30% diamantes, 20% picas y 20% tréboles. En otras palabras, esperaríamos que los corazones y los diamantes aparecieran 1,5 veces más que las picas y los tréboles (la proporción 30% : 20% es lo mismo que 1,5 : 1). Esto me parece una teoría tonta, y es bastante fácil probar esta hipótesis nula explícitamente especificada con los datos de nuestro análisis jamovi. En la ventana de análisis (etiquetada como 'Prueba de propor-

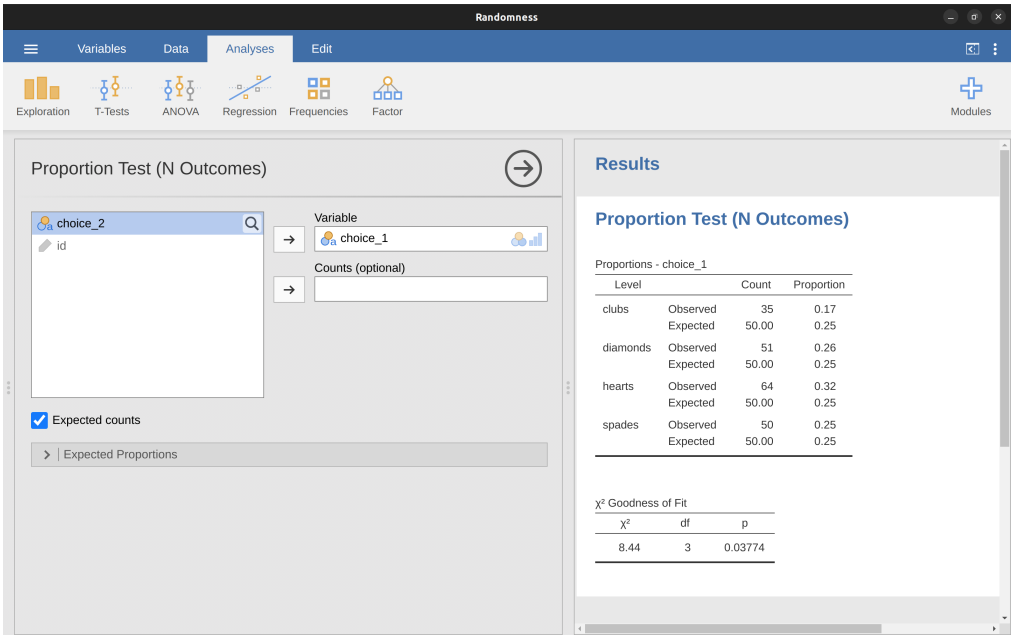


Figure 10.4: Una prueba de proporciones de una muestra de χ^2 en jamovi, con una tabla que muestra las frecuencias y proporciones tanto observadas como esperadas

ción (N resultados)' en Figure 10.4, puedes expandir las opciones para 'Proporciones esperadas'. Si haces esto, hay opciones para introducir diferentes valores de relación para la variable que has seleccionado, en nuestro caso esta es la opción 1. Cambia la relación para reflejar la nueva hipótesis nula, como en Figure 10.5, y fíjate cómo cambian los resultados.

Los recuentos esperados ahora se muestran en Table 10.6.

Table 10.6: recuentos esperados para una hipótesis nula diferente

	♣	♢	♡	♠
expected frequency E_i	40	60	60	40

y el estadístico χ^2 es 4,74, 3 gl, $p = 0,182$. Ahora, los resultados de nuestras hipótesis actualizadas y las frecuencias esperadas son diferentes a las de la última vez. Como consecuencia, nuestra prueba estadística χ^2 es diferente, y nuestro valor p también es diferente. Desgraciadamente, el valor p es $ 0,182 $, por lo que no podemos rechazar la hipótesis nula (consulta Section 9.5 para recordar por qué). Lamentablemente, a pesar de que la hipótesis nula corresponde a una teoría muy tonta, estos datos no aportan pruebas suficientes en su contra.

Figure 10.5: cambiar las proporciones esperadas en la prueba de proporciones de una muestra de
chi^2 en jamovi

10.1.9 Cómo informar los resultados de la prueba

Así que ahora sabes cómo funciona la prueba y sabes cómo hacer la prueba usando una maravillosa caja de informática mágica con sabor a jamovi. Lo siguiente que necesitas saber es cómo escribir los resultados. Después de todo, no tiene sentido diseñar y ejecutar un experimento y luego analizar los datos si no se lo cuentas a nadie. Así que ahora hablemos de lo que debes hacer al informar tu análisis. Sigamos con nuestro ejemplo de palos de cartas. Si quisieras escribir este resultado para un artículo o algo así, entonces la forma convencional de informar esto sería escribir algo como esto:

> De los 200 participantes en el experimento, 64 seleccionaron corazones para su primera opción, 51 seleccionaron diamantes, 50 seleccionaron picas y 35 seleccionaron tréboles. Se realizó una prueba de bondad de ajuste de ji-cuadrado para comprobar si las probabilidades de elección eran idénticas para los cuatro palos. Los resultados fueron significativos ($\chi^2(3) = 8.44, p < .05$), lo que sugiere que las personas no eligieron cartas puramente al azar.

Esto es bastante sencillo y, es de esperar que parezca bastante anodino. Dicho esto, hay algunas cosas que debes tener en cuenta sobre esta descripción:

- *La prueba estadística va precedida por la estadística descriptiva.* Es decir, le he dicho al lector algo sobre el aspecto de los datos antes de pasar a hacer la prueba. En general, esta es una buena práctica. Recuerda siempre que tu lector no conoce tus datos tan bien como tú. Así que, a menos que se los describas correctamente, las pruebas estadísticas no tendrán ningún sentido para ellos y se sentirán frustrados y llorarán.
- *La descripción te dice cuál es la hipótesis nula que se está probando.* A decir verdad, los escritores no siempre lo hacen, pero suele ser una buena idea en situaciones de ambigüedad, o cuando no se puede confiar en que los lectores conozcan a

fondo las herramientas estadísticas que se utilizan. Muy a menudo, es posible que el lector no sepa (o recuerde) todos los detalles de la prueba que estás utilizando, ¡así que es una especie de cortesía "recordárselos"! En lo que respecta a la prueba de bondad de ajuste, generalmente puedes confiar en que una audiencia científica sepa cómo funciona (ya que se trata en la mayoría de las clases de introducción a la estadística). Sin embargo, sigue siendo una buena idea ser explícito al establecer la hipótesis nula (¡brevemente!) porque la hipótesis nula puede ser diferente dependiendo de para qué estés usando la prueba. Por ejemplo, en el ejemplo de las cartas, mi hipótesis nula era que las probabilidades de los cuatro palos eran idénticas (es decir, $P1 = P2 = P3 = P4 = 0,25$), pero esa hipótesis no tiene nada de especial. Podría haber probado fácilmente la hipótesis nula de que $P_1 = 0.7$ y $P2 = P3 = P4 = 0.1$ usando una prueba de bondad de ajuste. Por lo tanto, es útil para el lector que le expliques cuál era tu hipótesis nula. Además, fíjate que describí la hipótesis nula en palabras, no en matemáticas. Eso es perfectamente aceptable. Puedes describirlo en matemáticas si lo deseas, pero dado que la mayoría de los lectores encuentran que las palabras son más fáciles de leer que los símbolos, la mayoría de los escritores tienden a describir la hipótesis nula usando palabras si pueden.

- *Se incluye un "bloque de estadísticos".* Cuando informé los resultados de la prueba en sí, no solo dije que el resultado era significativo, incluí un "bloque de estadísticos" (es decir, la parte densa de aspecto matemático entre paréntesis) que aporta toda la información estadística "clave". Para la prueba de bondad de ajuste ji-cuadrado, la información que se informa es la prueba estadística (que el estadístico de bondad de ajuste fue 8.44), la información sobre la distribución utilizada en la prueba (χ^2 con 3 grados de libertad que normalmente se abrevia a $\chi^2(3)$), y luego la información sobre si el resultado fue significativo (en este caso $p < .05$). La información particular que debe incluirse en el bloque de estadísticos es diferente para cada prueba, por lo que cada vez que presente una nueva prueba, te mostraré cómo debería ser el bloque de estadísticos.[7] Sin embargo, el principio general es que siempre debes proporcionar suficiente información para que el lector pueda verificar los resultados de la prueba por sí mismo si realmente lo desea.
- *Los resultados son interpretados.* Además de indicar que el resultado era significativo, proporcioné una interpretación del resultado (es decir, que la gente no eligió al azar). Esto también es una gentileza para el lector, porque le dice algo sobre lo que debe creer acerca de lo que está pasando en sus datos. Si no incluyes algo como esto, es muy difícil para tu lector entender lo que está pasando.[8]

Como con todo lo demás, tu principal preocupación debe ser explicar las cosas a tu

[7] Bueno, más o menos. Las convenciones sobre cómo se deben presentarse las estadísticas tienden a diferir un poco de una disciplina a otra. He tendido a ceñirme a cómo se hacen las cosas en psicología, ya que es a lo que me dedico. Pero creo que el principio general de proporcionar suficiente información al lector para que pueda comprobar los resultados es bastante universal.

[8] para algunas personas, este consejo puede sonar extraño, o al menos contradictorio con los consejos "habituales" sobre cómo redactar un informe técnico. Por lo general, a los estudiantes se les dice que la sección de "resultados" de un informe sirve para describir los datos e informar del análisis estadístico, y que la sección de "discusión" sirve para interpretarlos. Eso es cierto, pero creo que la gente suele interpretarlo de forma demasiado literal. Yo suelo hacer una interpretación rápida y sencilla de los datos en la sección de resultados, para que el lector entienda lo que nos dicen los datos. Luego, en la discusión, intento contar una historia más amplia sobre cómo mis resultados encajan con el resto de la literatura científica. En resumen, no dejes que el consejo de "la interpretación va en la discusión" convierta tu sección de resultados en una basura incomprensible. Ser entendido por tu lector es mucho más importante.

lector. Recuerda siempre que el objetivo de informar tus resultados es comunicarlo a otro ser humano. No puedo decirte cuántas veces he visto la sección de resultados de un informe o una tesis o incluso un artículo científico que es simplemente un galimatías, porque el escritor se ha centrado únicamente en asegurarse de haber incluido todos los números y se olvidó de realmente comunicarse con el lector humano.

> *Satanás se deleita por igual en las estadísticas y en citar las escrituras*[9] – HG pozos

[9]si has estado leyendo con mucha atención y eres una pedante matemática como yo, hay una cosa sobre la forma en que escribí la prueba de ji-cuadrado en la última sección que podría estar molestándote un poco. Hay algo que no cuadra al escribir "$\chi^2(3) = 8.44$", estarás pensando. Después de todo, es el estadístico de bondad de ajuste lo que equivale a 8,44, así que ¿no debería haber escrito $X^2 = 8,44$ o tal vez $GOF = 8,44$? Esto parece combinar la distribución muestral (es decir, χ^2 con gl = 3) con la prueba estadística (es decir, X^2). Lo más probable es que pensaras que era un error tipográfico, ya que χ y X se parecen bastante. Curiosamente, no lo es. Escribir $\chi^2(3)$= 8,44 es esencialmente una forma muy condensada de escribir "la distribución muestral de la prueba estadística es $\chi^2(3)$. y el valor de la prueba estadística es 8,44". En cierto sentido, esto es algo estúpido. Hay muchas pruebas estadísticas diferentes que resultan tener una distribución muestral de ji-cuadrado. El estadístico X^2 que hemos usado para nuestra prueba de bondad de ajuste es solo uno de muchos (aunque uno de los más comunes). En un mundo sensato y perfectamente organizado, siempre tendríamos un nombre distinto para la prueba estadística y la distribución muestral. De esa manera, el bloque de estadísticos en sí mismo te diría exactamente qué fue lo que calculó el investigador. A veces esto sucede. Por ejemplo, la prueba estadística utilizada en la prueba de bondad de ajuste de Pearson se escribe X^2 , pero hay una prueba estrechamente relacionada conocida como G-test[a] (Sokal & Rohlf, 1994), en la que la prueba estadística se escribe como G. Da la casualidad de que la prueba de bondad de ajuste de Pearson y la prueba G prueban la misma hipótesis nula, y la distribución muestral es exactamente la misma (es decir, ji-cuadrado con $k - 1$ grados de libertad). Si hubieras hecho una prueba G para los datos de las cartas en lugar de una prueba de bondad de ajuste, habrías terminado con una prueba estadística de $G = 8.65$, que es ligeramente diferente del valor $X^ 2 = 8,44 $ que obtuve antes y que produce un valor p ligeramente más pequeño de $p = 0,034 $. Supongamos que la convención fuera informar de la prueba estadística, luego la distribución muestral y luego el valor p. Si eso fuera cierto, estas dos situaciones producirían diferentes bloques de estadísticos: mi resultado original sería $X^2 = 8.44$, $\chi^2(3)$, $p = .038$, mientras que la nueva versión usando la prueba G se escribiría como $G = 8.65$, $\chi^2(3)$, $p = .034$. Sin embargo, la norma de información condensada, el resultado original se escribe $\chi^2(3) = 8.44, p = .038$, y el nuevo se escribe $\chi^2(3) = 8.65, p = .034$, por lo que en realidad no está claro qué prueba realicé. Entonces, ¿por qué no vivimos en un mundo en el que el contenido del bloque de estadísticos especifica de forma única qué pruebas se realizaron? La razón profunda es que la vida es un lío. Nosotras (como usuarias de herramientas estadísticas) queremos que sea agradable, ordenada y organizada. Queremos que esté diseñada, como si fuera un producto, pero no es así como funciona la vida. La estadística es una disciplina intelectual tanto como cualquier otra, y como tal es un proyecto distribuido masivamente, en parte colaborativo y en parte competitivo que nadie realmente entiende por completo. Las cosas que tú y yo usamos como herramientas de análisis de datos no fueron creadas por un acto de los dioses de la estadística. Fueron inventadas por muchas personas diferentes, publicadas como artículos en revistas académicas, implementadas, corregidas y modificadas por muchas otras personas y luego explicadas a los estudiantes en libros de texto por otra persona. Como consecuencia, hay muchas pruebas estadísticas que ni siquiera tienen nombre y, como consecuencia, reciben el mismo nombre que la distribución muestral correspondiente. Como veremos más adelante, cualquier prueba estadística que siga una distribución χ^2 se denomina comúnmente "estadístico ji-cuadrado", cualquier estadístico que siga una distribución t se denomina "estadístico t", etcétera. Pero, como ilustra el ejemplo de χ^2 versus G, dos cosas diferentes con la misma distribución muestral siguen siendo, bueno, diferentes. Como consecuencia, a veces es una buena idea tener claro cuál fue la prueba real que se ejecutó, especialmente si estás haciendo algo inusual. Si solo dices "prueba de ji-cuadrado", en realidad no está claro de qué prueba estás hablando. Aunque, dado que las dos pruebas de ji-cuadrado más comunes son la prueba de bondad de ajuste y la prueba de independencia, la mayoría de los lectores con entrenamiento en estadística probablemente puedan adivinar. Sin embargo, es algo a tener en cuenta. – [a] Para complicar las cosas, la prueba G es un caso especial de toda una clase de pruebas que se conocen como pruebas de razón de verosimilitud. No cubro las pruebas de razón de verosimilitud en este libro, pero es muy útil conocerlas.

10.2 La prueba de independencia (o asociación) χ^2

> GUARDBOT 1: ¡Alto!
> GUARDBOT 2: ¿Eres robot o humano?
> LEELA: Robot… seremos.
> FRY: ¡Ah, sí! ¡Solo dos robots robóticos! ¿eh?
> GUARDBOT 1: Administrar la prueba.
> GUARDBOT 2: ¿Cuál de las siguientes opciones preferirías? ¿A: Un cachorro, B: Una linda flor de tu amorcito, o C: Un gran archivo de datos con el formato adecuado?
> GUARDBOT 1: ¡Elige!
> *Futurama, "Miedo a un planeta bot"*

El otro día estaba viendo un documental animado que examinaba las pintorescas costumbres de los nativos del planeta *Chapek 9*. Al parecer, para acceder a su capital, el visitante debe demostrar que es un robot y no un ser humano. Para determinar si un visitante es humano o no, los nativos le preguntan si prefiere cachorros, flores o archivos de datos grandes y bien formateados. "Muy ingenioso", pensé, "pero, ¿y si los humanos y los robots tienen las mismas preferencias? Entonces probablemente no sería una prueba muy buena, ¿verdad?" Resulta que tengo en mis manos los datos de las prueba que las autoridades civiles de *Chapek 9* utilizaron para comprobarlo. Lo que hicieron fue muy sencillo. Encontraron un grupo de robots y un grupo de humanos y les preguntaron qué preferían. Guardé sus datos en un archivo llamado chapek9.omv, que ahora podemos cargar en jamovi. Además de la variable ID que identifica a cada persona, hay dos variables de texto nominales, especie y elección. En total, hay 180 entradas en el conjunto de datos, una para cada persona (contando tanto a los robots como a los humanos como "personas") a quienes se les pidió que hicieran una elección. En concreto, hay 93 humanos y 87 robots, y la opción preferida por abrumadora mayoría es el archivo de datos. Puedes comprobarlo tú misma pidiéndole a jamovi las tablas de frecuencia, en el botón 'Exploración' - 'Descriptivos'. Sin embargo, este resumen no aborda la pregunta que nos interesa. Para hacerlo, necesitamos una descripción más detallada de los datos. Lo que queremos es ver las opciones desglosadas *por especies*. Es decir, necesitamos tabular los datos de forma cruzada (ver Section 6.1). En jamovi, hacemos esto usando el análisis 'Frecuencias' - 'Tablas de contingencia' - 'Muestras independientes', y deberíamos obtener una tabla parecida a Table 10.7.

Table 10.7: tabulación cruzada de los datos

	Robot	**Human**	**Total**
Puppy	13	15	28
Flower	30	13	43
Data	44	65	109
Total	87	93	180

De ello se desprende claramente que la gran mayoría de los humanos eligieron el archivo de datos, mientras que los robots tendieron a ser mucho más equilibrados en sus preferencias. Dejando a un lado por el momento la pregunta de por qué los humanos son más propensos a elegir el archivo de datos (lo cual parece bastante extraño, hay que reconocerlo), lo primero que tenemos que hacer es determinar si la discrepancia entre las

elecciones de los humanos y las de los robots en el conjunto de datos es estadísticamente significativa.

10.2.1 Construyendo nuestra prueba de hipótesis

¿Cómo analizamos estos datos? En concreto, dado que mi hipótesis de investigación es que "los humanos y los robots responden a la pregunta de forma diferente", ¿cómo puedo construir una prueba de la hipótesis nula de que "los humanos y los robots responden a la pregunta de la misma manera"? Como antes, comenzamos estableciendo una notación para describir los datos (Table 10.8).

Table 10.8: Notación para describir los datos

	Robot	**Human**	**Total**
Puppy	O_{11}	O_{12}	R_1
Flower	O_{21}	O_{22}	R_2
Data	O_{31}	O_{32}	R_3
Total	C_1	C_2	N

En esta notación decimos que O_{ij} es un recuento (frecuencia observada) del número de encuestados que son de la especie j (robots o humanos) que dieron la respuesta i (cachorro, flor o datos) cuando se les pidió que hicieran una elección. El número total de observaciones se escribe N, como de costumbre. Finalmente, he usado R_i para indicar los totales de las filas (p. ej., R_1 es el número total de personas que eligieron la flor) y C_j para indicar los totales de las columnas (p. ej., C_1 es el total número de robots).[10]

Pensemos ahora en lo que dice la hipótesis nula. Si los robots y los humanos responden de la misma manera a la pregunta, significa que la probabilidad de que "un robot diga cachorro" es la misma que la probabilidad de que "un humano diga cachorro", y así sucesivamente para las otras dos posibilidades. Entonces, si usamos P_{ij} para denotar "la probabilidad de que un miembro de la especie j dé una respuesta i", entonces nuestra hipótesis nula es que:

$$
\begin{aligned}
&H_0 : \text{Todo lo siguiente es verdadero:} \\
&\quad P_{11} = P_{12} \text{ (misma probabilidad de decir "cachorro")}, \\
&\quad P_{21} = P_{22} \text{ (misma probabilidad de decir "flor")}, \text{ y} \\
&\quad P_{31} = P_{32} \text{ (misma probabilidad de decir "datos")}.
\end{aligned}
$$

Y en realidad, dado que la hipótesis nula afirma que las probabilidades verdaderas de elección no dependen de la especie de la persona que hace la elección, podemos dejar

[10] Nota técnica. La forma en que describí la prueba supone que los totales de las columnas son fijos (es decir, el investigador tenía la intención de encuestar a 87 robots y 93 humanos) y los totales de las filas son aleatorios (es decir, resulta que 28 personas eligieron el cachorro). Para usar la terminología de mi libro de texto de estadística matemáticas [@ Hogg2005], técnicamente debería referirme a esta situación como una prueba ji-cuadrado de homogeneidad y reservar el término prueba de independencia de ji-cuadrado para la situación en la que tanto los totales de fila como de columna son resultados aleatorios del experimento. En los borradores iniciales de este libro, eso es exactamente lo que hice. Sin embargo, resulta que estas dos pruebas son idénticas, por lo que las he unido.

que Pi se refiera a esta probabilidad, por ejemplo, P1 es la probabilidad verdadera de elegir al cachorro.

A continuación, de la misma manera que hicimos con la prueba de bondad de ajuste, lo que debemos hacer es calcular las frecuencias esperadas. Es decir, para cada uno de los recuentos observados O_{ij} , necesitamos averiguar qué nos diría la hipótesis nula que debemos esperar. Vamos a denotar esta frecuencia esperada por E_{ij}. Esta vez, es un poco más complicado. Si hay un total de C_j personas que pertenecen a la especie j, y la verdadera probabilidad de que cualquiera (independientemente de la especie) elija la opción i es P_i , entonces la frecuencia esperada es simplemente:

$$E_{ij} = C_j \times P_i$$

Ahora bien, todo esto está muy bien, pero tenemos un problema. A diferencia de la situación que tuvimos con la prueba de bondad de ajuste, la hipótesis nula en realidad no especifica un valor particular para Pi.

Es algo que tenemos que estimar (ver Chapter 8) a partir de los datos. Afortunadamente, es bastante fácil. Si 28 de 180 personas seleccionaron las flores, una estimación natural de la probabilidad de elegir flores es $\frac{28}{180}$, que es aproximadamente .16. Si expresamos esto en términos matemáticos, lo que estamos diciendo es que nuestra estimación de la probabilidad de elegir la opción i es solo el total de la fila dividido por el tamaño total de la muestra:

$$\hat{P}_i = \frac{R_i}{N}$$

Por lo tanto, nuestra frecuencia esperada se puede escribir como el producto (es decir, la multiplicación) del total de filas y el total de columnas, dividido por el número total de observaciones:[11]

$$\hat{E}_{ij} = \frac{R_i \times C_j}{N}$$

[Detalle técnico adicional [12]]

Como antes, los valores grandes de X^2 indican que la hipótesis nula proporciona una mala descripción de los datos, mientras que los valores pequeños de X^2 sugieren que hace un buen trabajo al explicar los datos. Por lo tanto, al igual que la última vez, queremos rechazar la hipótesis nula si X^2 es demasiado grande.

[11] Técnicamente, E_{ij} aquí es una estimación, por lo que probablemente debería escribir $\hat{E}_{ij}$. Pero como nadie más lo hace, yo tampoco lo haré.

[12] Ahora que ya sabemos cómo calcular las frecuencias esperadas, es sencillo definir una prueba estadística, siguiendo exactamente la misma estrategia que usamos en la prueba de bondad de ajuste. De hecho, es prácticamente el mismo estadístico. Para una tabla de contingencia con r filas y c columnas, la ecuación que define nuestro estadístico X^2 es

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

La única diferencia es que tengo que incluir dos signos de suma (es decir, $\sum$) para indicar que estamos sumando sobre ambas filas y columnas.

No es sorprendente que este estadístico tenga una distribución χ^2. Todo lo que tenemos que hacer es averiguar cuántos grados de libertad hay, lo que en realidad no es demasiado difícil. Como mencioné antes, se puede pensar (normalmente) que los grados de libertad son iguales al número de puntos de datos que estás analizando, menos el número de restricciones. Una tabla de contingencia con r filas y c columnas contiene un total de r^c frecuencias observadas, por lo que ese es el número total de observaciones. ¿Qué pasa con las restricciones? Aquí, es un poco más complicado. La respuesta es siempre la misma

$$df = (r-1)(c-1)$$

pero la explicación de por qué los grados de libertad toman este valor es diferente dependiendo del diseño experimental. Por ejemplo, supongamos que hubiéramos querido encuestar exactamente a 87 robots y 93 humanos (totales de las columnas fijados por el experimentador), pero hubiéramos dejado que los totales de fila variaran libremente (los totales de fila son variables aleatorias). Pensemos en las restricciones que se aplican en este caso. Bien, puesto que hemos fijado deliberadamente los totales de las columnas por Acto del Experimentador, tenemos restricciones de c allí mismo. Pero, en realidad hay más que eso. ¿Recuerdas que nuestra hipótesis nula tenía algunos parámetros libres (es decir, tuvimos que estimar los valores de Pi)? Esos también importan. No voy a explicar por qué en este libro, pero cada parámetro libre en la hipótesis nula es como una restricción adicional. Entonces, ¿cuántas hay? Bueno, dado que estas probabilidades tienen que sumar 1, solo hay $r-1$ de estas. Así que nuestros grados de libertad totales son:

$$\begin{aligned} df &= \text{(número de observaciones) - (número de restricciones)} \\ &= (r \times c) - (c + (r-1)) \\ &= rc - c - r + 1 \\ &= (r-1)(c-1) \end{aligned}$$

Por otra parte, supongamos que lo único que el experimentador fijó fue el tamaño total de la muestra N. Es decir, quer interrogamos a las primeras 180 personas que vimos y resultó que 87 eran robots y 93 eran humanos. Esta vez, nuestro razonamiento sería ligeramente diferente, pero nos llevaría a la misma respuesta. Nuestra hipótesis nula sigue siendo $r-1$ parámetros libres correspondientes a las probabilidades de elección, pero ahora también tiene $c-1$ parámetros libres correspondientes a las probabilidades de especie, porque también tendríamos que estimar la probabilidad de que una persona muestreada al azar resulte ser un robot.[13] Finalmente, dado que en realidad fijamos el número total de observaciones N, esa es una restricción más. Por lo tanto, ahora tenemos rc observaciones y $(c-1)+(r-1)+1$ restricciones. ¿Cuál es el resultado?

[13] un problema que a muchas nos preocupa en la vida real.

$$
\begin{aligned}
df &= \text{(número de observaciones) - (número de restricciones)} \\
&= (r \times c) - ((c-1) + (r-1) + 1) \\
&= (r-1)(c-1)
\end{aligned}
$$

Increíble.

10.2.2 Haciendo la prueba en jamovi

Bien, ahora que sabemos cómo funciona la prueba, veamos cómo se hace en jamovi. Por muy tentador que sea guiarte a través de los tediosos cálculos para que te veas obligada a aprenderlo por el camino largo, creo que no tiene sentido. Ya te mostré cómo hacerlo de la manera larga para la prueba de bondad de ajuste en la última sección, y como la prueba de independencia no es conceptualmente diferente, no aprenderás nada nuevo haciéndola de la manera larga. Así que en su lugar voy a ir directamente a mostrarte la manera más fácil. Después de ejecutar la prueba en jamovi ('Frecuencias' - 'Tablas de contingencia' - 'Muestras independientes'), todo lo que tienes que hacer es mirar debajo de la tabla de contingencia en la ventana de resultados de jamovi y allí está el estadístico χ^2 para ti. Muestra un valor estadístico χ^2 de 10,72, con 2 gl y valor p = 0,005.

Ha sido fácil, ¿verdad? También puedes pedirle a jamovi que te muestre los recuentos esperados: sólo tienes que hacer clic en la casilla de verificación 'Recuentos' - 'Esperados' en las opciones de 'Celdas' y los recuentos esperados aparecerán en la tabla de contingencia. Y mientras lo haces, sería útil disponer de una medida del tamaño del efecto. Elegiremos la V de Cramer, y puedes especificarlo desde una casilla de verificación en las opciones de 'Estadísticas', y da un valor para la V de Cramer de $0,24$. Ver Figure 10.6. Hablaremos de esto más adelante.

Esta salida nos da suficiente información para escribir el resultado:

> El χ^2 de *Pearson* reveló una asociación significativa entre especie y elección ($\chi^2(2) = 10.7, p < .01$). Los robots parecían más propensos a decir que prefieren las flores, pero los humanos eran más propensos a decir que preferían los datos.

Fíjate en que, una vez más, he dado un poco de interpretación para ayudar al lector humano a entender qué está pasando con los datos. Más adelante, en mi sección de discusión, proporcionaría un poco más de contexto. Para ilustrar la diferencia, esto es lo que probablemente diría más adelante:

> El hecho de que los humanos parezcan preferir más los archivos de datos en bruto que los robots es algo contraintuitivo. Sin embargo, en su contexto tiene cierto sentido, ya que la autoridad civil de Chapek 9 tiene una desafortunada tendencia a matar y diseccionar a los humanos cuando son identificados. Por lo tanto, lo más probable es que los participantes humanos no respondieran honestamente a la pregunta, para evitar consecuencias potencialmente indeseables. Esto debería considerarse una debilidad metodológica importante.

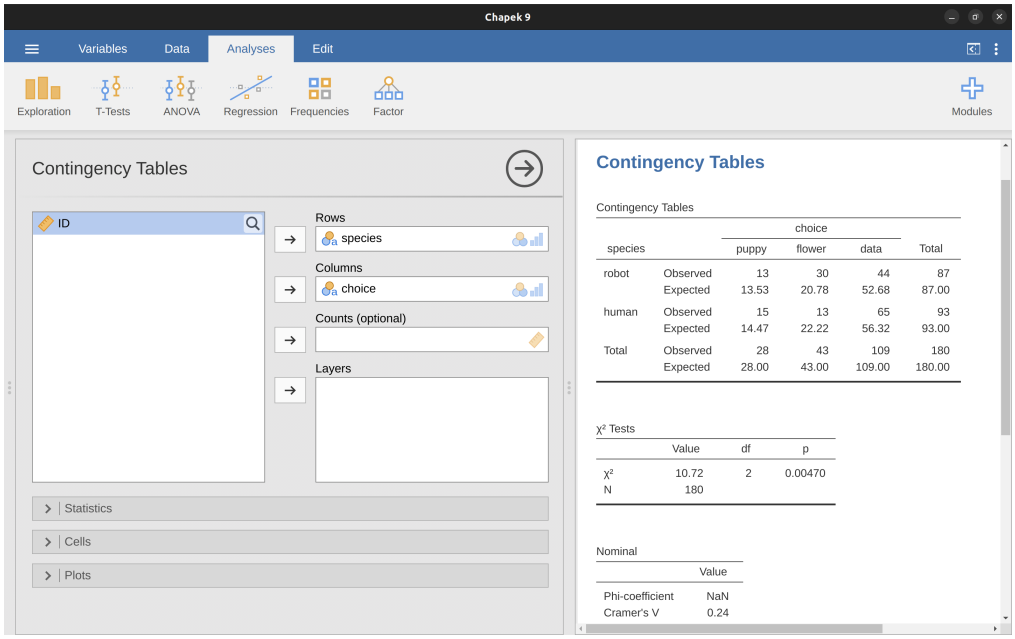


Figure 10.6: Prueba de χ^2 de muestras independientes en jamovi usando los datos de Chapek 9

Esto podría clasificarse como un ejemplo bastante extremo de un efecto de reactividad, supongo. Obviamente, en este caso el problema es lo suficientemente grave como para que el estudio sea más o menos inútil como herramienta para comprender las diferencias de preferencias entre humanos y robots. Sin embargo, espero que esto ilustre la diferencia entre obtener un resultado estadísticamente significativo (nuestra hipótesis nula se rechaza a favor de la alternativa) y encontrar algo de valor científico (los datos no nos dicen nada de interés sobre nuestra hipótesis de investigación debido a un gran problema metodológico).

10.3 La corrección de continuidad

Bien, es hora de una pequeña digresión. Te he estado mintiendo un poco hasta ahora. Hay un pequeño cambio que necesitas hacer en los cálculos cuando sólo tengas 1 grado de libertad. Se llama "corrección de continuidad" o, a veces, **corrección de Yates**. Recuerda lo que señalé antes: la prueba χ^2 se basa en una aproximación, concretamente en el supuesto de que la distribución binomial empieza a parecerse a una distribución normal para N grandes. Uno de los problemas de esto es que a menudo no funciona del todo bien, especialmente cuando solo se tiene 1 grado de libertad (por ejemplo, cuando se realiza una prueba de independencia en una tabla de contingencia de 2×2). La razón principal principal es que la verdadera distribución muestral para el estadístico X^2 es en realidad discreta (¡porque se trata de datos categóricos!) pero la distribución χ^2 es continua. Esto puede introducir problemas sistemáticos. En concreto, cuando N es pequeño y cuando $df = 1$, el estadístico de bondad de ajuste tiende a ser "demasiado grande", lo que significa que en realidad tiene un valor mayor de lo que piensas (o, de

manera equivalente, los valores p son un poco demasiado pequeño).

Por lo que he podido leer en el artículo de Yates[14], la corrección es básicamente un truco. No se deriva de ninguna teoría basada en principios. Más bien, se basa en un examen del comportamiento de la prueba y en la observación de que la versión corregida parece funcionar mejor. Puedes especificar esta corrección en jamovi desde una casilla de verificación en las opciones de 'Estadísticas', donde se llama 'corrección de continuidad χ^2'.

10.4 Tamaño del efecto

Como ya hemos comentado en Section 9.8, cada vez es más habitual pedir a los investigadores que informen sobre alguna medida del tamaño del efecto. Supongamos que hemos realizado la prueba de ji-cuadrado, que resulta ser significativa. Ahora sabes que existe alguna asociación entre las variables (prueba de independencia) o alguna desviación de las probabilidades especificadas (prueba de bondad de ajuste). Ahora deseas informar una medida del tamaño del efecto. Es decir, dado que hay una asociación o desviación, ¿cuán fuerte es?

Hay varias medidas diferentes que puedes elegir para informar y varias herramientas diferentes que puedes usar para calcularlas. No voy a hablar de todas ellas, sino que me centraré en las medidas del tamaño del efecto que se informan con más frecuencia.

Por defecto, las dos medidas que la gente tiende a informar con más frecuencia son el estadístico ϕ y la versión algo superior, conocida como V de Cramer.

[Detalle técnico adicional [15]]

Y ya está. Esta parece ser una medida bastante popular, presumiblemente porque es fácil de calcular y da respuestas que no son completamente tontas. Con V de Cramer, se sabe que el valor realmente oscila entre 0 (ninguna asociación) a 1 (asociación perfecta).

[14] Yates (1934) sugirió una solución simple, en la que redefine el estadístico de bondad de ajuste como:

$$\chi^2 = \sum_i \frac{(|E_i - O_i| - 0.5)^2}{E_i}$$

Básicamente, solo resta 0.5 en todas partes.

[15] Matemáticamente, son muy sencillos. Para calcular el estadístico ϕ, basta con dividir el valor de X^2 por el tamaño de la muestra y sacar la raíz cuadrada:

$$\phi = \sqrt{\frac{X^2}{N}}$$

La idea es que el estadístico ϕ oscila entre 0 (ninguna asociación) y 1 (asociación perfecta), pero no siempre lo hace cuando la tabla de contingencia es mayor que \$2 ×2 \$, lo que es un auténtico incordio. Para tablas más grandes, es posible obtener $\phi > 1$, lo cual es bastante insatisfactorio. Así que, para corregir esto, la gente suele preferir informar el estadístico V propuesto por Cramer (1946). Es un ajuste bastante simple de ϕ. Si tienes una tabla de contingencia con r filas y c columnas, defines $k = min(r, c)$ como el menor de los dos valores. Si es así, entonces el estadístico V de Cramer es

$$V = \sqrt{\frac{X^2}{N(k-1)}}$$

10.5 Supuestos de la(s) prueba(s)

Todas las pruebas estadísticas se basan en supuestos y suele ser una buena idea comprobar que se cumplen. En el caso de las pruebas de ji-cuadrado analizadas hasta ahora en este capítulo, los supuestos son:

- *Las frecuencias esperadas son suficientemente grandes.* ¿Recuerdas que en la sección anterior vimos que la distribución muestral χ^2 surge porque la distribución binomial es bastante parecida a una distribución normal? Pues bien, como comentamos en Chapter 7, esto solo es cierto cuando el número de observaciones es suficientemente grande. En la práctica, esto significa es que todas las frecuencias esperadas deben ser razonablemente grandes. ¿Cómo de razonablemente grandes? Las opiniones difieren, pero el supuesto por defecto parece ser que, en general, te gustaría ver todas las frecuencias esperadas mayores de 5, aunque para tablas más grandes, probablemente estaría bien si al menos el 80% de las frecuencias esperadas están por encima de 5 y ninguna está por debajo de 1. Sin embargo, por lo que he podido descubrir (p. ej., Cochran (1954)), estos parecen haber sido propuestos como pautas generales, no reglas estrictas y rápidas, y parecen ser algo conservadoras (Larntz, 1978) .
- *Los datos son independientes entre sí.* Un supuesto algo oculto de la prueba de ji-cuadrado es que tienes que creer de verdad que las observaciones son independientes. Esto es lo que quiero decir. Supongamos que estoy interesada en la proporción de bebés nacidos en un hospital en particular que son niños. Me paseo por las salas de maternidad y observo a 20 niñas y solo 10 niños. Parece una diferencia bastante convincente, ¿verdad? Pero más tarde, resulta que en realidad había entrado en la misma sala 10 veces y, en realidad, solo había visto a 2 niñas y 1 niño. No es tan convincente, ¿verdad? Mis 30 observaciones originales no eran en absoluto independientes y, de hecho, solo equivalían a 3 observaciones independientes. Obviamente, este es un ejemplo extremo (y muy tonto), pero ilustra la cuestión básica. La no independencia "estropea las cosas". A veces, hace que rechace falsamente la hipótesis nula, como ilustra el ejemplo tonto del hospital, pero también puede ocurrir al contrario. Para dar un ejemplo un poco menos estúpido, consideremos lo que pasaría si hubiera hecho el experimento con las cartas de forma ligeramente diferente. En lugar de pedir a 200 personas que imaginen la selección de una carta al azar, supongamos que pido a 50 personas que seleccionen 4 cartas. Una posibilidad sería que *todos* seleccionen un corazón, un trébol, un diamante y una pica (de acuerdo con la "heurística de la representatividad" (Tversky & Kahneman, 1974). Se trata de un comportamiento muy poco aleatorio de las personas, pero en este caso obtendría una frecuencia observada de 50 para los cuatro palos. Para este ejemplo, el hecho de que las observaciones no sean independientes (porque las cuatro cartas que elija estarán relacionadas entre sí) en realidad conduce al efecto opuesto, manteniendo falsamente la hipótesis nula.

Si te encuentras en una situación en la que se viola la independencia, puedes utilizar la prueba de McNemar (de la que hablaremos) o la prueba de Cochran (de la que no hablaremos). Del mismo modo, si los recuentos esperados son demasiado pequeños, consulta la prueba exacta de Fisher. A continuación abordaremos estos temas.

10.6 La prueba exacta de Fisher

¿Qué hacer si los recuentos en las celdas son demasiado pequeños, pero aún así quieres probar la hipótesis nula de que las dos variables son independientes? Una respuesta sería "recopilar más datos", pero eso es demasiado simplista. Hay muchas situaciones en las que sería inviable o poco ético hacerlo. Si es así, los estadísticos tienen una especie de obligación moral de proporcionar a los científicos mejores pruebas. En este caso, Fisher (1922a) proporcionó amablemente la respuesta correcta a la pregunta. Para ilustrar la idea básica, supongamos que estamos analizando los datos de un experimento de campo que analiza el estado emocional de las personas que han sido acusadas de brujería, algunas de las cuales están siendo quemadas en la hoguera.[16] Desafortunadamente para el científico (pero afortunadamente para la población en general), en realidad es bastante difícil encontrar personas en el proceso de ser prendidas fuego, por lo que los recuentos son terriblemente pequeños en algunos casos. Una tabla de contingencia de los datos de salem.csv ilustra el punto (Table 10.9).

Table 10.9: tabla de contingencia de los datos de salem.csv

	happy	**FALSE**	**TRUE**
on.fire	FALSE	3	10
	TRUE	3	0

Observando estos datos, sería difícil no sospechar que las personas que no están en llamas tienen más probabilidades de ser felices que las que están en llamas. Sin embargo, la prueba de ji-cuadrado hace que esto sea muy difícil de comprobar debido al pequeño tamaño de la muestra. Así que, hablando como alguien que no quiere que le prendan fuego, *realmente* me gustaría poder obtener una respuesta mejor que esta. Aquí es donde la **prueba exacta de Fisher** (Fisher, 1922a) es muy útil.

La prueba exacta de Fisher funciona de forma algo diferente a la prueba de ji-cuadrado (o, de hecho, a cualquiera de las otras pruebas de hipótesis de las que hablo en este libro) en la medida en que no tiene una prueba estadística, pero calcula el valor p "directamente". Explicaré los fundamentos de cómo funciona la prueba para una tabla de contingencia de 2×2. Como antes, vamos a tener un poco de notación (Table 10.10).

Table 10.10: Notación para la prueba exacta de Fisher

	Happy	**Sad**	**Total**
Set on fire	O_{11}	O_{12}	R_1
Not set on fire	O_{21}	O_{22}	R_2
Total	C_1	C_2	N

Para construir la prueba, Fisher trata los totales de fila y columna (R_1, R_2, C_1 y C_2) como cantidades fijas conocidas y luego calcula la probabilidad de que hubiéramos obtenido las frecuencias observadas que obtuvimos (O_{11}, O_{12}, O_{21} and O_{22}) dados esos totales. En la notación que desarrollamos en Chapter 7 esto se escribe:

[16] Este ejemplo se basa en un artículo de broma publicado en el *Journal of Irreproducible Results*

$$P(O_{11}, O_{12}, O_{21}, O_{22} \mid R_1, R_2, C_1, C_2)$$

y, como puedes imaginar, es un ejercicio un poco complicado averiguar cuál es esta probabilidad. Pero resulta que esta probabilidad viene descrita por una distribución conocida como distribución hipergeométrica. Lo que tenemos que hacer para calcular nuestro valor p es calcular la probabilidad de observar esta tabla en particular o una tabla *"más extrema"*. [17] En la década de 1920, calcular esta suma era desalentador incluso en las situaciones más simples, pero hoy en día es bastante fácil siempre que las tablas no sean demasiado grandes y el tamaño de la muestra no sea demasiado grande. La cuestión conceptualmente complicada es averiguar qué significa decir que una tabla de contingencia es más "extrema" que otra. La solución más sencilla es decir que la tabla con la probabilidad más baja es la más extrema. Esto nos da el valor p.

Puedes especificar esta prueba en jamovi desde una casilla de verificación en las opciones de 'Estadísticas' del análisis de 'Tablas de contingencia'. Cuando se hace esto con los datos del archivo salem.csv, el estadístico de la prueba exacta de Fisher se muestra en los resultados. Lo que más nos interesa aquí es el valor p, que en este caso es lo suficientemente pequeño (p = 0,036) para justificar el rechazo de la hipótesis nula de que las personas que se queman son tan felices como las que no. Ver Figure 10.7.

10.7 La prueba de McNemar

Supongamos que te han contratado para trabajar para el *Partido Político Genérico Australiano* (PPGA), y parte de tu trabajo consiste en averiguar la eficacia de los anuncios políticos del PPGA. Así que decides reunir una muestra de $N = 100$ personas y pedirles que vean los anuncios de AGPP. Antes de que vean nada, les preguntas si tienen intención de votar al PPGA, y después de ver los anuncios, les vuelves a preguntar para ver si alguien ha cambiado de opinión. Obviamente, si eres buena en tu trabajo, también harías muchas otras cosas, pero consideremos sólo este sencillo experimento. Una forma de describir los datos es mediante la tabla de contingencia que se muestra en Table 10.11.

Table 10.11: Tabla de contingencia con datos de anuncios políticos del PPGA

	Before	**After**	**Total**
Yes	30	10	40
No	70	90	160
Total	100	100	200

A primera vista, se podría pensar que esta situación se presta a la prueba de independencia χ^2 de Pearson (según La prueba de independencia (o asociación) χ^2). Sin embargo, un poco de reflexión revela que tenemos un problema. Tenemos 100 participantes, pero 200 observaciones. Esto se debe a que cada persona nos ha proporcionado una respuesta tanto en la columna del antes como en la del después. Esto significa que las 200 observaciones no son independientes entre sí. Si el votante A dice "sí" la primera vez y el votante B dice "no", entonces es de esperar que el votante A tenga más

[17] No es sorprendente que la prueba exacta de Fisher esté motivada por la interpretación de Fisher de un valor p, ¡no por la de Neyman! Consulta Section 9.5.

Contingency Tables

	on.fire		
happy	FALSE	TRUE	Total
FALSE	3	3	6
TRUE	10	0	10
Total	13	3	16

χ^2 Tests

	Value	df	p
χ^2	6.15	1	0.01311
Fisher's exact test			0.03571
N	16		

Figure 10.7: análisis de prueba exacta de Fisher en jamovi

probabilidades de decir "sí" la segunda vez que el votante B. La consecuencia de esto es que la prueba habitual χ^2 no dará respuestas fiables debido a la violación del supuesto de independencia. Ahora bien, si esta fuera una situación realmente poco común, no me molestaría en hacerte perder el tiempo hablando de ella. Pero no es poco común en absoluto. Este es un diseño estándar de medidas repetidas, y ninguna de las pruebas que hemos considerado hasta ahora puede manejarlo.

La solución al problema fue publicada por McNemar (1947). El truco consiste en comenzar tabulando los datos de una forma ligeramente distinta (Table 10.12).

Table 10.12: tabula los datos de una manera diferente cuando tienes datos de medidas repetidas

	Before: Yes	**Before: No**	**Total**
After: Yes	5	5	10
After: No	25	65	90
Total	30	70	100

A continuación, pensemos en cuál es nuestra hipótesis nula: es que la prueba del "antes" y la prueba del "después" tienen la misma proporción de personas que dicen "Sí, votaré por PPGA". Debido a la forma en que hemos reescrito los datos, significa que ahora estamos probando la hipótesis de que los totales de fila y los totales de columna provienen de la misma distribución. Así, la hipótesis nula en la prueba de McNemar es que tenemos "homogeneidad marginal". Es decir, que los totales de fila y los totales de columna tienen la misma distribución: $P_a + P_b = P_a + P_c$ y de manera similar que $P_c + P_d = P_b + P_d$. Observa que esto significa que la hipótesis nula en realidad se simplifica a Pb = Pc. En otras palabras, en lo que respecta a la prueba de McNemar, ¡solo importan las entradas fuera de la diagonal de esta tabla (es decir, b y c)! Después de observar esto, la **prueba de homogeneidad marginal de McNemar** no es diferente a una prueba habitual de χ^2. Después de aplicar la corrección de Yates, nuestra prueba estadística se convierte en:

$$\chi^2 = \frac{(|bc| - 0.5)^2}{b + c}$$

o, para volver a la notación que usamos anteriormente en este capítulo:

$$\chi^2 = \frac{(|O_{12} - O_{21}| - 0.5)^2}{O_{12} + O_{21}}$$

y este estadístico tiene un χ^2 (aproximadamente) con gl = 1. Sin embargo, recuerda que, al igual que las otras pruebas de χ^2, es solo una aproximación, por lo que debes tener un número de recuentos esperados razonablemente grande para que funcione.

10.7.1 Haciendo la prueba de McNemar en jamovi

Ahora que ya sabes en qué consiste la prueba de McNemar, hagamos una. El archivo agpp.csv contiene los datos sin procesar de los que he hablado anteriormente. El conjunto de datos de agpp contiene tres variables, una variable de id que etiqueta a cada participante en el conjunto de datos (veremos por qué es útil en un momento), una

variable **response_before** que registra la respuesta de la persona cuando se le hizo la pregunta la primera vez, y una variable response_after que muestra la respuesta que dio cuando se le hizo la misma pregunta por segunda vez. Ten en cuenta que cada participante aparece solo una vez en este conjunto de datos. Ves a 'Análisis' - 'Frecuencias' - 'Tablas de contingencia' - Análisis de 'Muestras emparejadas' en jamovi, y mueve **response_before** al cuadro 'Rows' y **response_after** al cuadro 'Columns'. Obtendrás entonces una tabla de contingencia en la ventana de resultados, con el estadístico de la prueba de McNemar justo debajo, ver Figure 10.8.

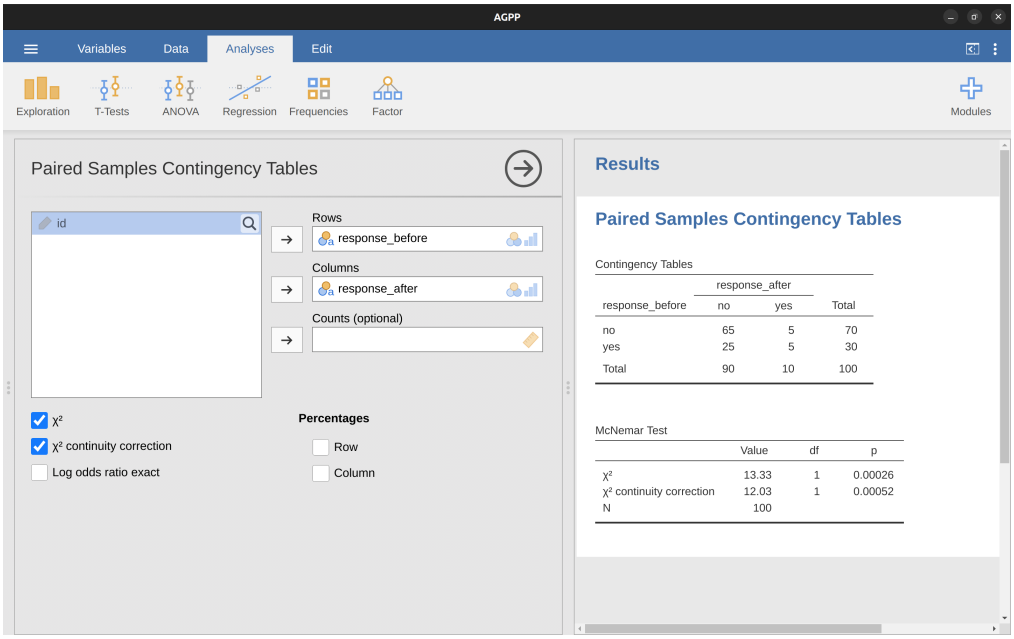


Figure 10.8: salida de prueba de McNemar en jamovi

Y hemos terminado. Acabamos de realizar una prueba de McNemar para determinar si las personas tenían la misma probabilidad de votar PPGA después de los anuncios que antes. La prueba fue significativa ($\chi^2(1) = 12.03, p < .001$), lo que sugiere que no lo fueron. Y, de hecho, parece que los anuncios tuvieron un efecto negativo: era menos probable que las personas votaran por el PPGA después de ver los anuncios. Lo cual tiene mucho sentido si consideras la calidad de un típico anuncio político.

10.8 ¿Cuál es la diferencia entre McNemar y la independencia?

Volvamos al principio del capítulo y examinemos de nuevo el conjunto de datos de las cartas. Si recuerdas, el diseño experimental real que describí implicaba que las personas hicieran dos elecciones. Como tenemos información sobre la primera elección y la segunda elección que todos hicieron, podemos construir la siguiente tabla de contingencia que compara la primera elección con la segunda elección (Table 10.13).

Supongamos que quisiera saber si la elección que haces la segunda vez depende de la

Table 10.13: tabulación cruzada de la primera contra la segunda opción con los datos de Randomness.omv (cartas)

	Before: Yes	**Before: No**	**Total**
After: Yes	a	b	$a + b$
After: No	c	d	$c + d$
Total	$a + c$	$b + d$	n

elección que hiciste la primera vez. Aquí es donde es útil una prueba de independencia, y lo que estamos tratando de hacer es ver si hay alguna relación entre las filas y las columnas de esta tabla.

Supongamos que quisiera saber si, en promedio, las frecuencias de las elecciones de palo fueron diferentes la segunda vez que la primera vez. En esa situación, lo que realmente estoy intentando ver es si los totales de las filas son diferentes de los totales de las columnas. Es entonces cuando se utiliza la prueba de McNemar.

En Figure 10.9 se muestran los diferentes estadísticos producidos por esos distintos análisis. ¡Observe que los resultados son diferentes! No se trata de la misma prueba.

10.9 Resumen

Las ideas clave discutidas en este capítulo son:

- La prueba de bondad de ajuste χ^2 (ji-cuadrado) se usa cuando tienes una tabla de frecuencias observadas de diferentes categorías, y la hipótesis nula te da un conjunto de probabilidades "conocidas" para compararlas.
- La prueba de independencia (o asociación) χ^2 se usa cuando se tiene una tabla de contingencia (tabulación cruzada) de dos variables categóricas. La hipótesis nula es que no existe relación o asociación entre las variables.
- Tamaño del efecto para una tabla de contingencia se puede medir de varias maneras. En particular, observamos el estadístico V de Cramer.
- Ambas versiones de la prueba de Pearson se basan en dos supuestos: que las frecuencias esperadas son suficientemente grandes y que las observaciones son independientes (Supuestos de la(s) prueba(s). La prueba exacta de Fisher se puede usar cuando las frecuencias esperadas son pequeñas La prueba de McNemar se puede utilizar para algunos tipos de violaciones de la independencia.

Si estás interesada en obtener más información sobre el análisis de datos categóricos, una buena primera opción sería Agresti (1996) que, como sugiere el título, ofrece una Introducción al análisis de datos categóricos. Si el libro introductorio no es suficiente para ti (o no puedes resolver el problema en el que estás trabajando), podrías considerar Agresti (2002), Análisis de datos categóricos. Este último es un texto más avanzado, por lo que probablemente no sea prudente pasar directamente de este libro a aquel.

Contingency Tables

Contingency Tables

	choice_2				
choice_1	clubs	diamonds	hearts	spades	Total
clubs	10	9	10	6	35
diamonds	20	4	13	14	51
hearts	20	18	3	23	64
spades	18	13	15	4	50
Total	68	44	41	47	200

χ^2 Tests

	Value	df	p
χ^2	29.24	9	0.00059
N	200		

Paired Samples Contingency Tables

Contingency Tables

	choice_2				
choice_1	clubs	diamonds	hearts	spades	Total
clubs	10	9	10	6	35
diamonds	20	4	13	14	51
hearts	20	18	3	23	64
spades	18	13	15	4	50
Total	68	44	41	47	200

McNemar Test

	Value	df	p
χ^2	16.03	6	0.01358
N	200		

Figure 10.9: Independiente vs. Emparejado (McNemar) con los datos de Randomness.omv (cartas)

Chapter 11

Comparar dos medias

En Chapter 10 cubrimos la situación en la que tanto la variable de resultado como la variable predictora estaban en una escala nominal. Muchas situaciones del mundo real presentan esa característica, por lo que encontrarás que las pruebas de ji-cuadrado en particular se usan bastante. Sin embargo, es mucho más probable que te encuentres en una situación en la que tu variable de resultado esté en una escala de intervalo o mayor, y lo que te interese es si el valor promedio de la variable de resultado es mayor en un grupo u otro. Por ejemplo, una psicóloga podría querer saber si los niveles de ansiedad son más altos entre los padres que entre los que no son padres, o si la capacidad de la memoria de trabajo se reduce al escuchar música (en relación con no escuchar música). En un contexto médico, podríamos querer saber si un nuevo medicamento aumenta o disminuye la presión arterial. Un científico agrícola podría querer saber si agregar fósforo a las plantas nativas australianas las matará.[1] En todas estas situaciones, nuestra variable de resultado es una variable continua, con escala de intervalo o de razón, y nuestro predictor es una variable de "agrupación" binaria. En otras palabras, queremos comparar las medias de los dos grupos.

Para comparar medias se utiliza una prueba t, de la cual hay diversas variedades dependiendo exactamente de qué pregunta quieras resolver. Como consecuencia, la mayor parte de este capítulo se centra en diferentes tipos de pruebas t: pruebas t de una muestra, pruebas t de muestras independientes y pruebas t de muestras relacionadas. Luego hablaremos sobre las pruebas unilaterales y, después, hablaremos un poco sobre la d de Cohen, que es la medida estándar del tamaño del efecto para una prueba t. Las últimas secciones del capítulo se centran en los supuestos de las pruebas t y los posibles remedios si se violan. Sin embargo, antes de discutir cualquiera de estas cosas, comenzaremos con una discusión sobre la prueba z.

[1] La experimentación informal en mi jardín sugiere que sí. Los nativos australianos están adaptados a niveles bajos de fósforo en relación con cualquier otro lugar de la Tierra, por lo que si compraste una casa con un montón de plantas exóticas y quieres plantar nativas, manténlas separadas; los nutrientes para las plantas europeas son veneno para las australianas.

11.1 La prueba z de una muestra

En esta sección describiré una de las pruebas más inútiles de toda la estadística: la prueba z. En serio, esta prueba casi nunca se usa en la vida real. Su único propósito real es que, al enseñar estadística, es un trampolín muy conveniente en el camino hacia la prueba t, que es probablemente la herramienta más (sobre)utilizada en estadística.

11.1.1 El problema de inferencia que aborda la prueba

Para presentar la idea que subyace a la prueba z, usemos un ejemplo sencillo. Un amigo mío, el Dr. Zeppo, califica su clase de introducción a la estadística en una curva. Supongamos que la calificación promedio en su clase es de 67.5 y la desviación estándar es de 9.5. De sus muchos cientos de estudiantes, resulta que 20 de ellos también reciben clases de psicología. Por curiosidad, me pregunto si los estudiantes de psicología tienden a obtener las mismas calificaciones que todos los demás (es decir, $67.5 $ promedio) o si tienden a obtener una puntuación más alta o más baja. Me envía por correo electrónico el archivo zeppo.csv, que uso para ver las calificaciones de esos estudiantes, en la vista de hoja de cálculo jamovi, y luego calculo la media en 'Exploración' - 'Descriptivos' [2]. El valor medio es 72.3.

50 60 60 64 66 66 67 69 70 74 76 76 77 79 79 79 81 82 82 89

Mmm. Puede ser que los estudiantes de psicología estén sacando notas un poco más altas de lo normal. Esa media muestral de $\bar{X} = 72,3$ es un poco más alta que la media hipotética de la población de $\mu = 67,5$ pero, por otro lado, un tamaño muestral de $N = 20$ no es tan grande. Tal vez sea pura casualidad.

Para responder a la pregunta, ayuda poder escribir qué es lo que creo que sé. En primer lugar, sé que la media muestral es $\bar{X} = 72,3$. Si estoy dispuesta a asumir que el alumnado de psicología tiene la misma desviación estándar que el resto de la clase, entonces puedo decir que la desviación estándar de la población es $\sigma = 9.5$. También supondré que dado que el Dr. Zeppo está calificando en una curva, las calificaciones de los y las estudiantes de psicología se distribuyen normalmente.

A continuación, ayuda tener claro lo que quiero aprender de los datos. En este caso mi hipótesis de investigación se relaciona con la media poblacional μ para las calificaciones del alumnado de psicología, la cual se desconoce. Específicamente quiero saber si $\mu = 67.5$ o no. Dado que esto es lo que sé, ¿podemos idear una prueba de hipótesis para resolver nuestro problema? Los datos, junto con la distribución hipotética de la que se cree que surgen, se muestran en Figure 11.1. No es del todo obvio cuál es la respuesta correcta, ¿verdad? Para ello, vamos a necesitar algunos estadísticos.

11.1.2 Construyendo la prueba de hipótesis

El primer paso para construir una prueba de hipótesis es tener claro cuáles son las hipótesis nula y alternativa. Esto no es demasiado difícil. Nuestra hipótesis nula, H_0, es que la verdadera media poblacional μ para las calificaciones de los estudiantes de psicología es $67,5\%$, y nuestra hipótesis alternativa es que la media poblacional no es $67,5\%$. Si escribimos esto en notación matemática, estas hipótesis se convierten en:

[2] para hacer esto, tuve que cambiar el nivel de medida de X a 'Continuo', ya que durante la apertura/importación del archivo csv jamovi lo convirtió en una variable de nivel nominal, que no es adecuada para mi análisis

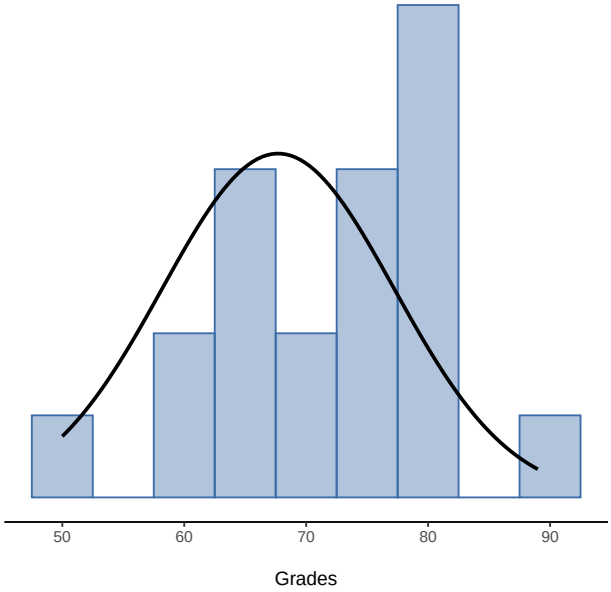


Figure 11.1: La distribución teórica (línea sólida) a partir de la cual se supone que se generaron las calificaciones (barras) de los estudiantes de psicología

$$H_0 : \mu = 67.5$$

$$H_1 : \mu \neq 67.5$$

aunque, para ser sincera, esta notación no añade mucho a nuestra comprensión del problema, es solo una forma compacta de escribir lo que estamos tratando de aprender de los datos. Las hipótesis nulas H_0 alternativa H_1 para nuestra prueba se ilustran en Figure 11.2. Además de ofrecernos estas hipótesis, el escenario descrito anteriormente nos proporciona una buena cantidad de conocimientos previos que podrían ser útiles. En concreto, hay dos datos especiales que podemos añadir:

1. Las calificaciones de psicología se distribuyen normalmente.
2. Se sabe que la verdadera desviación estándar de estas puntuaciones σ es 9,5.

Por el momento, actuaremos como si estos fueran hechos absolutamente fiables. En la vida real, este tipo de conocimiento de fondo absolutamente fiable no existe, por lo que si queremos confiar en estos hechos, solo tendremos que *suponer* que estas cosas son ciertas. Sin embargo, dado que estas suposiciones pueden o no estar justificadas, es posible que debamos verificarlas. Sin embargo, por ahora, mantendremos las cosas simples.

El siguiente paso es averiguar cuál sería una buena opción para la prueba estadística, algo que nos ayudará a discriminar entre H_0 y H_1. Dado que todas las hipótesis se refieren a la media de la población μ, la media de la muestra $\bar{X}$ sería un punto de partida muy útil. Lo que podríamos hacer es observar la diferencia entre la media muestral $\bar{X}$ y el valor que predice la hipótesis nula para la media poblacional. En nuestro ejemplo, eso significaría que calculamos $\bar{X} - 67.5$. De forma más general, si hacemos que μ_0 se

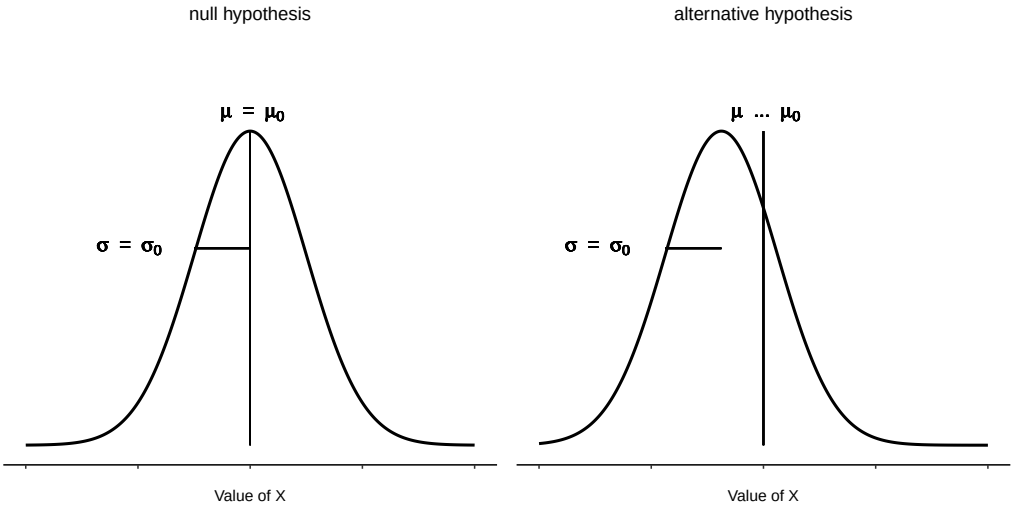


Figure 11.2: Ilustración gráfica de las hipótesis nula y alternativa asumidas por la prueba z de una muestra (es decir, la versión de dos colas). Tanto la hipótesis nula como la alternativa suponen que la distribución de la población es normal y, además, suponen que se conoce la desviación estándar de la población (fijada en algún valor $\$sigma_0$). La hipótesis nula (izquierda) es que la media poblacional μ es igual a algún valor especificado μ_0. La hipótesis alternativa (derecha) es que la media poblacional difiere de este valor, $\mu \neq \mu_0$

refiera al valor que la hipótesis nula afirma que es nuestra media poblacional, entonces querríamos calcular

$$\bar{X} - \mu_0$$

Si esta cantidad es igual o está muy cerca de 0, las cosas pintan bien para la hipótesis nula. Si esta cantidad está muy lejos de 0, entonces parece menos probable que valga la pena mantener la hipótesis nula. Pero, ¿a qué distancia de cero debería estar para que rechacemos H0?

Para averiguarlo necesitaremos confiar en esos dos conocimientos previos que mencioné anteriormente; es decir, que los datos sin procesar se distribuyen normalmente y que conocemos el valor de la desviación estándar de la población σ. Si la hipótesis nula es realmente verdadera y la media verdadera es μ_0, entonces estos hechos juntos significan que conocemos la distribución completa de la población de los datos: una distribución normal con media μ_0 y desviación estándar σ .[3]

Bien, si eso es cierto, ¿qué podemos decir sobre la distribución de $\bar{X}$? Bueno, como discutimos anteriormente (ver Section 8.3.3), la distribución muestral de la media $\bar{X}$ también es normal, y tiene media μ. Pero la desviación estándar de esta distribución

[3]adoptando la notación de Section 7.5, un estadístico podría escribir esto como:

$$X \sum Normal(\mu_0, \sigma^2)$$

muestral
$se(\bar{X})$, que se denomina error estándar de la media, es [4]

$$se(\bar{X}) = \frac{\sigma}{\sqrt{N}}$$

Ahora viene el truco. Lo que podemos hacer es convertir la media muestral $\bar{X}$ en una puntuación estándar (ver Section 4.5). Esto se escribe convencionalmente como z, pero por ahora me referiré a él como $z_{\bar{X}}$. Uso esta notación expandida para ayudarte a recordar que estamos calculando una versión estandarizada de una media muestral, no una versión estandarizada de una sola observación, que es a lo que generalmente se refiere una puntuación z. Cuando lo hacemos, la puntuación z para nuestra media muestral es

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{SE(\bar{X})}$$

o, equivalentemente

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{N}}}$$

Esta puntuación z es nuestra prueba estadística. Lo bueno de usar esto como nuestra prueba estadística es que, como todas las puntuaciones z, tiene una distribución normal estándar:[5]

En otras palabras, independientemente de la escala en la que se encuentren los datos originales, el estadístico z siempre tiene la misma interpretación: es igual a la cantidad de errores estándar que separan la media muestral observada $\bar{X}$ de la media poblacional μ_0 predicha por la hipótesis nula. Mejor aún, independientemente de cuáles sean realmente los parámetros poblacionales para las puntuaciones sin procesar, las regiones críticas del 5% para la prueba z son siempre las mismas, como se ilustra en Figure 11.3. Y lo que esto significaba, en los tiempos en que la gente hacía todos sus cálculos a mano, es que alguien podía publicar una tabla como Table 11.1. Esto, a su vez, significó que los investigadores pudieran calcular su estadístico z a mano y luego buscar el valor crítico en un libro de texto.

$$z_{\bar{X}} \sum Normal(0, 1)$$

[4] En otras palabras, si la hipótesis nula es verdadera, entonces la distribución muestral de la media se puede escribir de la siguiente manera:

$$\bar{X} \sum Normal(\mu_0, ES(\bar{X}))$$

[5] Nuevamente, puedes ver Section 4.5 si has olvidado por qué esto es cierto.

Table 11.1: valores críticos para diferentes niveles alfa

	critical z value	
desired α level	two-sided test	one-sided test
.1	1.644854	1.281552
.05	1.959964	1.644854
.01	2.575829	2.326348
.001	3.290527	3.090232

11.1.3 Un ejemplo práctico, a mano

Ahora, como mencioné anteriormente, la prueba z casi nunca se usa en la práctica. Se usa tan raramente en la vida real que la instalación básica de jamovi no tiene una función integrada para ello. Sin embargo, la prueba es tan increíblemente simple que es muy fácil hacerla manualmente. Volvamos a los datos de la clase del Dr. Zeppo. Habiendo cargado los datos de calificaciones, lo primero que debo hacer es calcular la media de la muestra, lo cual ya hice (72.3). Ya tenemos la desviación estándar poblacional conocida ($\sigma = 9.5$), y el valor de la media poblacional que especifica la hipótesis nula ($\mu_0 = 67.5$), y conocemos el tamaño de la muestra ($N = 20$).

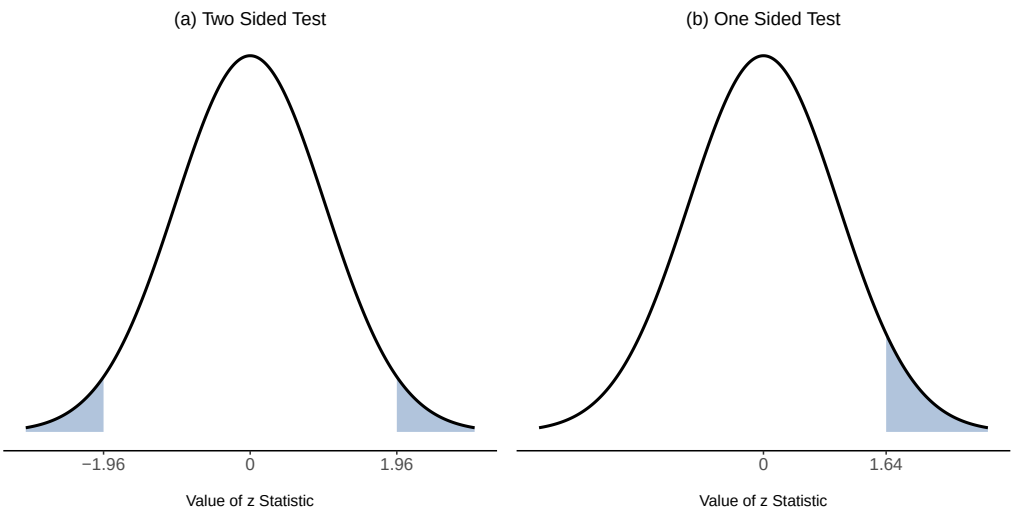


Figure 11.3: Regiones de rechazo para la prueba z de dos colas (panel (a)) y la prueba z de una cola (panel (b))

A continuación, calculemos el error estándar (verdadero) de la media (fácil de hacer con una calculadora):

$$sem.true = \frac{sd.true}{\sqrt{N}}$$

$$= \frac{9.5}{\sqrt{20}}$$

$$= 2.124265$$

Y finalmente, calculamos nuestra puntuación z:

$$z.score = \frac{sample.mean - mu.null}{sem.true}$$

$$= \frac{(72.3 - 67.5)}{2.124265}$$

$$= 2.259606$$

En este punto, tradicionalmente buscaríamos el valor 2.26 en nuestra tabla de valores críticos. Nuestra hipótesis original era de dos colas (realmente no teníamos ninguna teoría sobre si los estudiantes de psicología serían mejores o peores en estadística que otros estudiantes), por lo que nuestra prueba de hipótesis también es bilateral (o de dos colas). Mirando la pequeña tabla que mostré anteriormente, podemos ver que 2.26 es mayor que el valor crítico de 1.96 que se requeriría para ser significativo en $\alpha = .05$, pero menor que el valor de 2.58 que se requeriría para que sea significativo a un nivel de $\alpha = .01$. Por lo tanto, podemos concluir que tenemos un efecto significativo, que podríamos escribir diciendo algo como esto:

> *Con una nota media de* $73,2$ en la muestra de estudiantes de psicología, y asumiendo una desviación estándar poblacional real de $9,5$, podemos concluir que los y las estudiantes de psicología tienen puntuaciones en estadística significativamente diferentes a la media de la clase ($z = 2,26, N = 20, p < .05$).

11.1.4 Supuestos de la prueba z

Como he dicho antes, todas las pruebas estadísticas tienen supuestos. Algunas pruebas tienen supuestos razonables, mientras que otras no. La prueba que acabo de describir, la prueba z de una muestra, hace tres suposiciones básicas. Estas son:

- *Normalidad*. Como suele describirse, la prueba z supone que la verdadera distribución de la población es normal.[6] Suele ser un supuesto bastante razonable, y es un supuesto que podemos verificar si nos preocupa (consulta la Sección sobre [Comprobación de la normalidad de una muestra]).

[6] En realidad, esto es demasiado. Estrictamente hablando, la prueba z solo requiere que la distribución muestral de la media se distribuya normalmente. Si la población es normal, necesariamente se deduce que la distribución muestral de la media también es normal. Sin embargo, como vimos al hablar

- *Independencia.* El segundo supuesto de la prueba es que las observaciones en su conjunto de datos no están correlacionadas entre sí, o relacionadas entre sí de alguna manera divertida. Esto no es tan fácil de verificar estadísticamente, depende un poco de un buen diseño experimental. Un ejemplo obvio (y estúpido) de algo que viola este supuesto es un conjunto de datos en el que "copias" la misma observación una y otra vez en tu archivo de datos para que termines con un "tamaño de muestra" masivo, que consiste solo en una observación genuina. De manera más realista, debes preguntarte si es realmente plausible imaginar que cada observación es una muestra completamente aleatoria de la población que te interesa. En la práctica, este supuesto nunca se cumple, pero hacemos todo lo posible para diseñar estudios que minimicen los problemas de los datos correlacionados.
- *Desviación estándar conocida.* El tercer supuesto de la prueba z es que el investigador conoce la verdadera desviación estándar poblacional. Esto es una estupidez. En ningún problema de análisis de datos del mundo real conoces la desviación estándar de alguna población pero ignoras por completo la media μ. En otras palabras, este supuesto siempre es incorrecto.

En vista de la estupidez de suponer que se conoce α, veamos si podemos vivir sin ello. ¡Esto nos saca del lúgubre dominio de la prueba z y nos lleva al reino mágico de la prueba t, con unicornios, hadas y duendes!

11.2 La prueba t de una muestra

Después de pensarlo un poco, decidí que no sería seguro asumir que las calificaciones de los estudiantes de psicología necesariamente tienen la misma desviación estándar que los otros estudiantes en la clase del Dr. Zeppo. Después de todo, si considero la hipótesis de que no tienen la misma media, ¿por qué debería creer que tienen la misma desviación estándar? En vista de esto, debería dejar de asumir que conozco el verdadero valor de σ. Esto viola los supuestos de mi prueba z, por lo que, en cierto sentido, vuelvo al punto de partida. Sin embargo, no es que me falten opciones. Después de todo, todavía tengo mis datos sin procesar, y esos datos sin procesar me dan una estimación de la desviación estándar de la población, que es 9,52. En otras palabras, aunque no puedo decir que sé que $\sigma = 9,5$, puedo decir que $\hat{\sigma} = 9,52$.

Está bien, genial. Lo más obvio que podría hacer es ejecutar una prueba z, pero usando la desviación estándar estimada de 9.52 en lugar de confiar en mi suposición de que la verdadera desviación estándar es 9.5. Y probablemente no te sorprenda saber que esto aún nos daría un resultado significativo. Este enfoque está cerca, pero no es del todo correcto. Debido a que ahora confiamos en una estimación de la desviación estándar poblacional, necesitamos hacer algunos ajustes por el hecho de que tenemos cierta incertidumbre sobre cuál es realmente la desviación estándar poblacional real. Tal vez nuestros datos sean solo una casualidad… tal vez la verdadera desviación estándar poblacional sea 11, por ejemplo. Pero si eso fuera realmente cierto, y ejecutamos la prueba z asumiendo $\sigma = 11$, entonces el resultado terminaría siendo no significativo. Esto es un problema, y es uno que vamos a tener que abordar.

---

del teorema central del límite, es muy posible (incluso habitual) que la distribución muestral sea normal incluso si la distribución poblacional en sí misma no es normal. Sin embargo, a la luz de lo ridículo que resulta suponer que se conoce la verdadera desviación estándar, no tiene mucho sentido entrar en detalles al respecto.

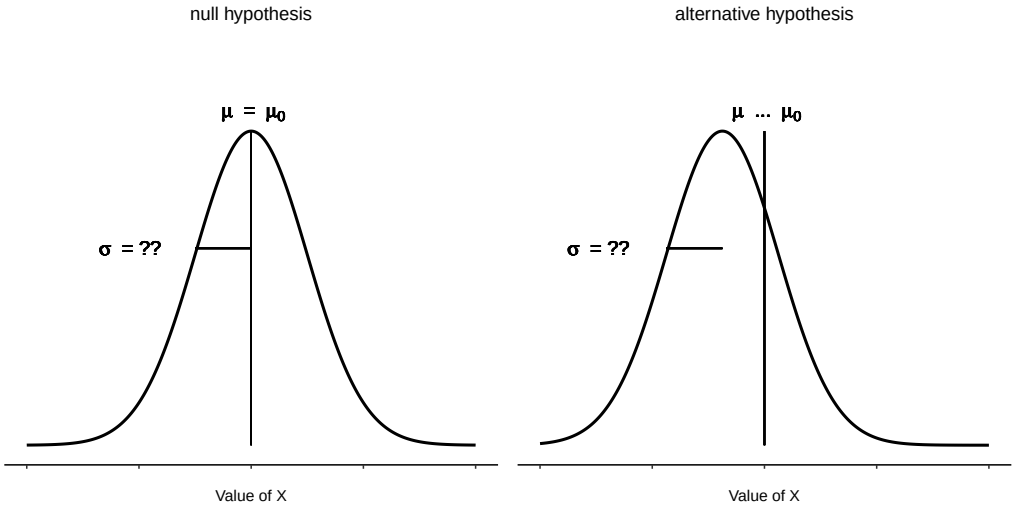


Figure 11.4: Ilustración gráfica de las hipótesis nula y alternativa asumidas por la prueba t de una muestra (bilateral). Ten en cuenta la similitud con la prueba z (Figure 11.2). La hipótesis nula es que la media poblacional μ es igual a algún valor especificado μ_0, y la hipótesis alternativa es que no lo es. Al igual que la prueba z, asumimos que los datos se distribuyen normalmente, pero no asumimos que la desviación estándar poblacional σ se conoce de antemano

11.2.1 Introducción a la prueba t

Esta ambigüedad es molesta y fue resuelta en 1908 por un tipo llamado William Sealy Gosset (Student, 1908), que en ese momento trabajaba como químico para la cervecería Guinness (ver J. F. Box (1987)). Debido a que Guinness vio con malos ojos que sus empleados publicaran análisis estadísticos (aparentemente sintieron que era un secreto comercial), publicó el trabajo bajo el seudónimo de "Un estudiante" y, hasta el día de hoy, el nombre completo de la prueba t es en realidad **Prueba t de Student**. Lo más importante que descubrió Gosset es cómo debemos tener en cuenta el hecho de que no estamos completamente seguras de cuál es la verdadera desviación estándar.[7] La respuesta es que cambia sutilmente la distribución muestral. En la prueba t, nuestra prueba estadística, ahora llamada estadístico t, se calcula exactamente de la misma manera que mencioné anteriormente. Si nuestra hipótesis nula es que la verdadera media es μ, pero nuestra muestra tiene una media $\bar{X}$ y nuestra estimación de la desviación estándar poblacional es $\hat{\sigma}$, entonces nuestro estadístico t es :

$$t = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}}{\sqrt{N}}}$$

Lo único que ha cambiado en la ecuación es que en lugar de usar el valor verdadero conocido σ, usamos la estimación $\hat{\sigma}$. Y si esta estimación se ha construido a partir de N observaciones, entonces la distribución muestral se convierte en una distribución t con

[7] Bueno, más o menos. Según entiendo la historia, Gosset solo proporcionó una solución parcial; Sir Ronald Fisher proporcionó la solución general al problema.

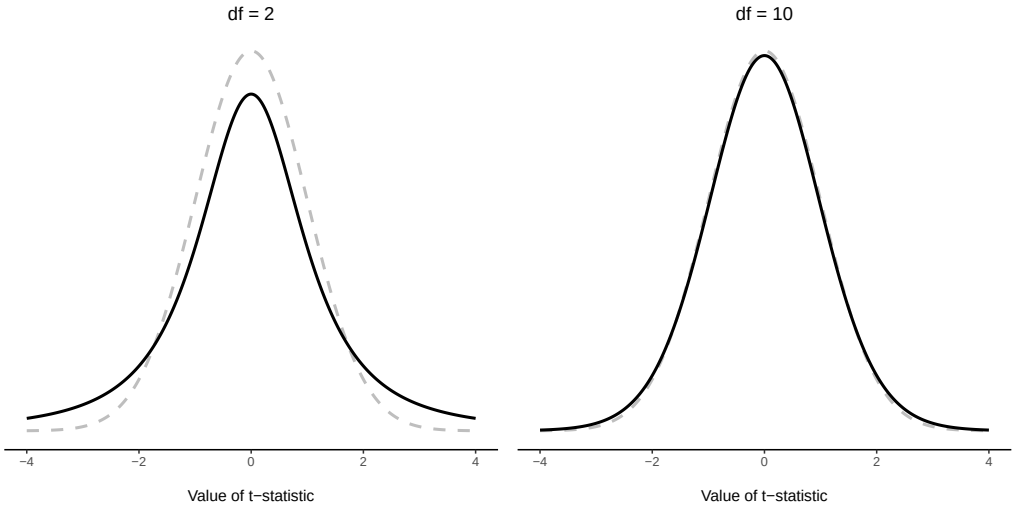


Figure 11.5: Distribución t con 2 grados de libertad (izquierda) y 10 grados de libertad (derecha), con una distribución normal estándar (es decir, media 0 y desviación estándar 1) representada como líneas de puntos a efectos comparativos. Observa que la distribución t tiene colas más pesadas (leptocúrticas, mayor curtosis) que la distribución normal; este efecto es bastante exagerado cuando los grados de libertad son muy pequeños, pero despreciable para valores más grandes. En otras palabras, para df grandes, la distribución t es esencialmente idéntica a una distribución normal.

$N-1$ **grados de libertad** (gl). La distribución t es muy similar a la distribución normal, pero tiene colas "más pesadas", como se explicó anteriormente en Section 7.6 y se ilustró en Figure 11.5. Ten en cuenta, sin embargo, que a medida que gl aumenta, la distribución t empieza a ser idéntica a la distribución normal estándar. Así es como debería ser: si tienes un tamaño de muestra de $N = 70 000 000 $, entonces tu "estimación" de la desviación estándar sería bastante perfecta, ¿verdad? Por lo tanto, debes esperar que para N grandes, la prueba t se comporte exactamente de la misma manera que una prueba z. Y eso es exactamente lo que sucede.

11.2.2 Haciendo la prueba en jamovi

Como era de esperar, la mecánica de la prueba t es casi idéntica a la mecánica de la prueba z. Así que no tiene mucho sentido pasar por el tedioso ejercicio de mostrarte cómo hacer los cálculos usando comandos de bajo nivel. Es prácticamente idéntico a los cálculos que hicimos anteriormente, excepto que usamos la desviación estándar estimada y luego probamos nuestra hipótesis usando la distribución t en lugar de la distribución normal. Entonces, en lugar de pasar por los cálculos en detalle por segunda vez, pasaré directamente a mostraros cómo se realizan realmente las pruebas t. jamovi viene con un análisis dedicado para pruebas t que es muy flexible (se pueden ejecutar muchos tipos diferentes de pruebas t). Es bastante sencillo de usar; todo lo que tienes que hacer es especificar 'Análisis' - 'T-Tests' - 'One Sample T-Test', mover la variable que te interesa (X) al cuadro 'Variables' y escribir el valor medio para la hipótesis nula ('67.5') en el cuadro 'Hipótesis' - 'Valor de la prueba'. Bastante fácil. Consulta Figure 11.6, que,

entre otras cosas a las que llegaremos en un momento, te da una prueba t = 2.25, con 19 grados de libertad y un valor p asociado de $ 0.036 $.

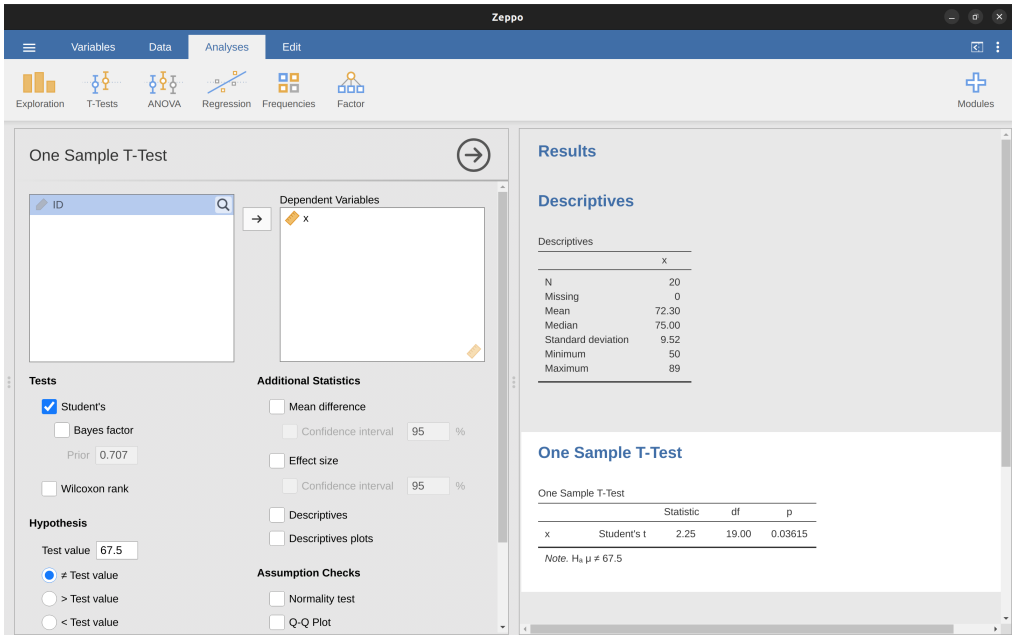


Figure 11.6: jamovi hace la prueba t de una muestra

También se informan otras dos cosas que podrían interesarte: el intervalo de confianza del 95% y una medida del tamaño del efecto (hablaremos más sobre los tamaños del efecto más adelante). Eso parece bastante sencillo. Ahora, ¿qué hacemos con este resultado? Bueno, ya que estamos fingiendo que realmente nos importa mi ejemplo de juguete, nos alegramos al descubrir que el resultado es estadísticamente significativo (es decir, un valor de p por debajo de 0,05). Podríamos informar del resultado diciendo algo así:

> *Con una nota media de* 72,3, los estudiantes de psicología obtuvieron una puntuación ligeramente superior a la nota media de 67,5 ($t(19) = 2,25$, $p = 0,036$); la diferencia de medias fue de 4,80 y el intervalo de confianza de 95% fue de 0,34 a 9,26.

...donde $t(19)$ es la notación abreviada de un estadístico t que tiene 19 grados de libertad. Dicho esto, a menudo no se informa el intervalo de confianza, o se hace usando una forma mucho más reducida que la que he utilizado aquí. Por ejemplo, no es raro ver el intervalo de confianza incluido como parte del bloque de estadísticos después de informar la diferencia media, así:

$$t(19) = 2.25, p = .036, CI_{95} = [0.34, 9.26]$$

Con tanta jerga metida en media línea, sabes que debes ser muy inteligente.[8]

[8] Más en serio, tiendo a pensar que lo contrario es cierto. Desconfío mucho de los informes técnicos

11.2.3 Supuestos de la prueba t de una muestra

Bien, entonces, ¿qué supuestos hace la prueba t de una muestra? Bueno, dado que la prueba t es básicamente una prueba z con el supuesto de desviación estándar conocida eliminada, no deberías sorprenderte al ver que hace los mismos supuestos que la prueba z, menos la desviación estándar conocida. Eso es:

- Normalidad. Seguimos suponiendo que la distribución poblacional es normal[9] y, como se indicó anteriormente, existen herramientas estándar que puedes usar para comprobar si se cumple este supuesto ([Comprobar la normalidad de una muestra]), y otras pruebas que puedes hacer en su lugar si se viola este supuesto ([Prueba de datos no normales]).
- Independencia. Una vez más, debemos suponer que las observaciones de nuestra muestra se generan independientemente unas de otras. Consulta la discusión anterior sobre la prueba z para obtener información específica (Supuestos de la prueba z).

En general, estos dos supuestos son razonables y, como consecuencia, la prueba t de una muestra se usa bastante en la práctica como una forma de comparar una media muestral con una media poblacional hipotética.

11.3 La prueba t de muestras independientes (prueba de Student)

Aunque la prueba t de una muestra tiene sus usos, no es el ejemplo más típico de una prueba t[10]. Una situación mucho más común surge cuando tienes dos grupos diferentes de observaciones. En psicología, esto tiende a corresponder a dos grupos diferentes de participantes, donde cada grupo corresponde a una condición diferente en tu estudio. Para cada persona en el estudio, mides alguna variable de resultado de interés, y la pregunta de investigación que estás haciendo es si los dos grupos tienen o no la misma media poblacional. Esta es la situación para la que está diseñada la prueba t de muestras independientes.

11.3.1 Los datos

Supongamos que tenemos 33 estudiantes asistiendo a las clases de estadística del Dr. Harpo, y el Dr. Harpo no califica con una curva. En realidad, la calificación del Dr. Harpo es un poco misteriosa, por lo que realmente no sabemos nada sobre cuál es

que llenan sus secciones de resultados con nada más que números. Puede que sea sólo que soy una idiota arrogante, pero a menudo siento que un autor que no intenta explicar e interpretar su análisis al lector, o bien no lo entiende por sí mismo o bien es un poco perezoso. Tus lectores son inteligentes, pero no infinitamente pacientes. No les molestes si puedes evitarlo.

[9] un comentario técnico. De la misma manera que podemos debilitar los supuestos de la prueba z para que solo estemos hablando de la distribución muestral, podemos debilitar los supuestos de la prueba t para que no tengamos que asumir la normalidad poblacional. Sin embargo, para la prueba t es más complicado hacer esto. Como antes, podemos reemplazar el supuesto de normalidad de la población con el supuesto de que la distribución muestral de $\bar{X}$ es normal. Sin embargo, recuerda que también confiamos en una estimación muestral de la desviación estándar, por lo que también requerimos que la distribución muestral de $\hat{\sigma}$ sea ji-cuadrado. Eso hace que las cosas sean más difíciles, y esta versión rara vez se usa en la práctica. Afortunadamente, si la distribución poblacional es normal, entonces se cumplen estos dos supuestos.

[10] Aunque es el más simple, por eso empecé con él.

la calificación promedio para la clase en general. En la clase hay dos tutores, Anastasia y Bernadette. Hay $N_1 = 15$ estudiantes en las tutorías de Anastasia y $N_2 = 18$ en las tutorías de Bernadette. La pregunta de investigación que me interesa es si Anastasia o Bernadette son mejores tutoras, o si no hay mucha diferencia. El Dr. Harpo me envía por correo electrónico las calificaciones del curso en el archivo harpo.csv. Como de costumbre, cargaré el archivo en jamovi y veré qué variables contiene: hay tres variables, ID, calificación y tutor. La variable de calificación contiene la calificación de cada estudiante, pero no se importa a jamovi con el atributo de nivel de medición correcto, por lo que necesito cambiar esto para que se considere una variable continua (ver Section 3.6). La variable tutor es un factor que indica quién fue la tutora de cada estudiante, ya sea Anastasia o Bernadette.

Podemos calcular las medias y las desviaciones estándar, utilizando el análisis 'Exploración' - 'descriptivo', y aquí hay un pequeño cuadro resumen (Table 11.2).

Table 11.2: cuadro resumen de descriptivos

	mean	**std dev**	**N**
Anastasia's students	74.53	9.00	15
Bernadette's students	69.06	5.77	18

Para darte una idea más detallada de lo que está pasando aquí, he trazado diagramas de caja y violín en jamovi, con puntuaciones medias agregadas al diagrama con un pequeño cuadrado sólido. Estos gráficos muestran la distribución de calificaciones para ambas tutoras (Figure 11.7),

11.3.2 Introducción a la prueba

La **prueba t de muestras independientes** se presenta de dos formas diferentes, la de Student y la de Welch. La prueba t de Student original, que es la que describiré en esta sección, es la más simple de las dos pero se basa en supuestos mucho más restrictivos que la prueba t de Welch. Suponiendo por el momento que deseas ejecutar una prueba bilateral, el objetivo es determinar si se extraen dos "muestras independientes" de datos de poblaciones con la misma media (la hipótesis nula) o diferentes medias (la hipótesis alternativa). Cuando decimos muestras "independientes", lo que realmente queremos decir aquí es que no existe una relación especial entre las observaciones en las dos muestras. Esto probablemente no tenga mucho sentido en este momento, pero estará más claro cuando hablemos más adelante sobre la prueba t de muestras relacionadas. Por ahora, señalemos que si tenemos un diseño experimental en el que los participantes se asignan aleatoriamente a uno de dos grupos, y queremos comparar el rendimiento medio de los dos grupos en alguna medida de resultado, entonces una prueba t de muestras independientes (en lugar de una prueba t de muestras pareadas) es lo que buscamos.

Bien, dejemos que μ_1 denote la media poblacional real para el grupo 1 (p. ej., los estudiantes de Anastasia), y μ_2 será la media poblacional real para el grupo 2 (p. ej., los estudiantes de Bernadette),[11] y, como de costumbre, dejaremos que $\bar{X}_1$ y $\bar{X}_2$ denoten

[11] Casi siempre surge una pregunta divertida en este punto: ¿a qué diablos se refiere la población

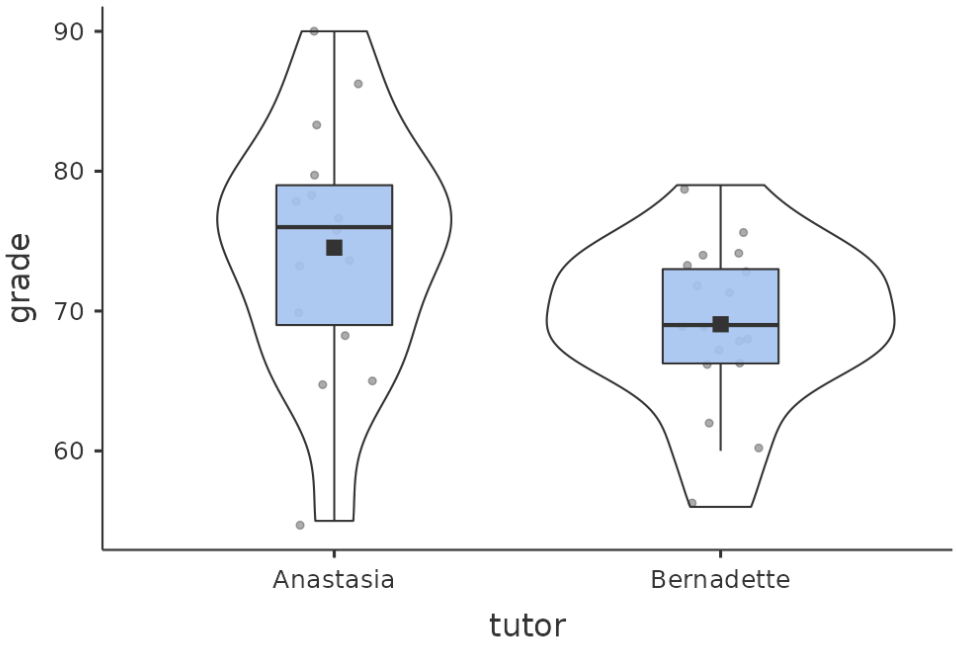


Figure 11.7: diagramas de caja y violín que muestran la distribución de calificaciones de los estudiantes en las clases de Anastasia y Bernadette. Visualmente, esto sugiere que los estudiantes de la clase de Anastasia pueden estar obteniendo calificaciones ligeramente mejores en promedio, aunque también parecen un poco más variables.

las medias muestrales observadas para ambos grupos. Nuestra hipótesis nula establece que las medias de las dos poblaciones son idénticas ($\mu_1 = \mu_2$) y la alternativa a esto es que no lo son ($\mu_1 \neq \mu_2$) (Figure 11.8). Escrito en lenguaje matemático, esto es:

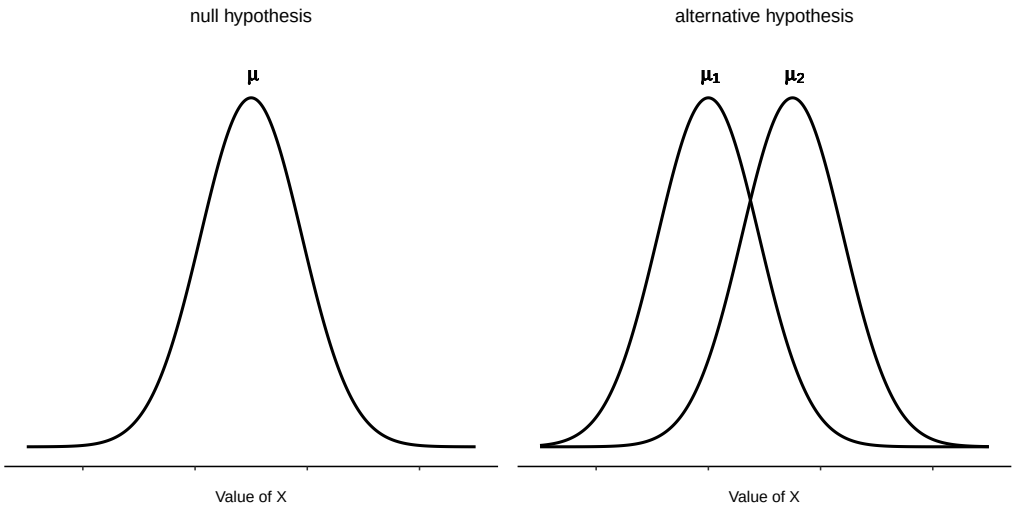


Figure 11.8: Ilustración gráfica de las hipótesis nula y alternativa asumidas por la prueba t de Student. La hipótesis nula supone que ambos grupos tienen la misma media μ, mientras que la alternativa supone que tienen medias diferentes μ_1 y μ_2. Observa que se supone que las distribuciones de la población son normales y que, aunque la hipótesis alternativa permite que el grupo tenga diferentes medias, se supone que tienen la misma desviación estándar.

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 \neq \mu_2$$

Para construir una prueba de hipótesis que maneje este escenario, comenzamos observando que si la hipótesis nula es verdadera, entonces la diferencia entre las medias poblacionales es *exactamente* cero, $\mu_1 - \mu_2 = 0$. Como consecuencia, una prueba estadística se basará en la diferencia entre las medias de las dos muestras. Porque si la hipótesis nula es verdadera, esperaríamos que $\bar{X}_1 - \bar{X}_2$ sea bastante cercano a cero. Sin embargo, tal como vimos con nuestras pruebas de una muestra (es decir, la prueba z de una muestra y la prueba t de una muestra), debemos ser precisos acerca de la proximidad a cero de esta diferencia. Y la solución al problema es más o menos la misma.

en este caso? ¿Es el grupo de estudiantes que realmente recibe la clase del Dr. Harpo (los 33)? ¿El conjunto de personas que podrían recibir la clase (un número desconocido de ellos)? ¿O algo mas? ¿Importa cuál de estos escojamos? Como mi alumnado me hace esta pregunta todos los años, daré una respuesta breve. Técnicamente sí, sí importa. Si cambias tu definición de lo que realmente es la población del "mundo real", entonces la distribución muestral de tu media observada $\bar{X}$ también cambia. La prueba t se basa en el supuesto de que las observaciones se muestrean al azar de una población infinitamente grande y, en la medida en que la vida real no sea así, entonces la prueba t puede ser incorrecta. En la práctica, sin embargo, esto no suele ser un gran problema. Aunque el supuesto casi siempre es incorrecto, no conduce a una gran cantidad de comportamiento patológico de la prueba, por lo que tendemos a ignorarlo.

Calculamos una estimación del error estándar (SE), como la última vez, y luego dividimos la diferencia entre las medias por esta estimación. Por tanto nuestro **estadístico t** será:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

Solo necesitamos averiguar cuál es realmente esta estimación del error estándar. Esto es un poco más complicado que en el caso de cualquiera de las dos pruebas que hemos visto hasta ahora, por lo que debemos analizarlo con mucho más cuidado para comprender cómo funciona.

11.3.3 Una "estimación conjunta" de la desviación estándar

En la "prueba t de Student" original, asumimos que los dos grupos tienen la misma desviación estándar poblacional. Es decir, sin importar si las medias de la población son las mismas, asumimos que las desviaciones estándar de la población son idénticas, $\sigma_1 = \sigma_2$. Dado que asumimos que las dos desviaciones estándar son iguales, quitamos los subíndices y nos referimos a ambos como σ. ¿Cómo debemos estimar esto? ¿Cómo debemos construir una única estimación de una desviación estándar cuando tenemos dos muestras? La respuesta es, básicamente, las promediamos. Bueno, más o menos. En realidad, lo que hacemos es calcular un promedio *ponderado* de las estimaciones de *varianza*, que usamos como nuestra **estimación combinada de la varianza**. El peso asignado a cada muestra es igual al número de observaciones en esa muestra, menos 1.

[Detalle técnico adicional [12]]

[12] Matemáticamente, podemos escribir esto como

$$w_1 = N_1 - 1$$

$$w_2 = N_2 - 1$$

Ahora que hemos asignado pesos a cada muestra calculamos la estimación combinada de la varianza cogiendo el promedio ponderado de las dos estimaciones de varianza, $\hat{sigma}_1^2$ y $\hat{sigma}_2^2$

$$\hat{\sigma}_p^2 = \frac{w_1 \hat{\sigma}_1^2 + w_2 \hat{\sigma}_2^2}{w_1 + w_2}$$

Finalmente, convertimos la estimación de la varianza agrupada a una estimación de desviación estándar agrupada, haciendo la raíz cuadrada.

$$\hat{\sigma}_p = \sqrt{\frac{w_1 \hat{\sigma}_1^2 + w_2 \hat{\sigma}_2^2}{w_1 + w_2}}$$

Y si mentalmente sustituyes ($w_1 = N_1 - 1$) y $w_2 = N_2 - 1$ en esta ecuación obtendrá una fórmula muy fea. Una fórmula muy fea que en realidad parece ser la forma "estándar" de describir la estimación de la desviación estándar agrupada. Sin embargo, no es mi forma favorita de pensar en las desviaciones estándar agrupadas. Prefiero pensarlo así. Nuestro conjunto de datos en realidad corresponde a un conjunto de N observaciones que se clasifican en dos grupos. Así que usemos la notación X_{ik} para referirnos a la calificación recibida por el i-ésimo estudiante en el k-ésimo grupo de tutoría. Es decir, (X_{11} es la calificación que recibió el primer estudiante en la clase de Anastasia, X_{21} es su segundo estudiante, y así sucesivamente. Y tenemos dos medias grupales separadas $\bar{X}_1$ y $\bar{X}_2$, a las que podríamos referirnos "genéricamente" usando la notación $\bar{X}_k$, es decir, la calificación media para el k-ésimo grupo de tutoría. Hasta ahora, todo bien. Ahora, dado que cada estudiante cae en una de las dos tutorías, podemos describir su desviación de la media del grupo como la diferencia

$$X_{ik} - \bar{X}_k$$

11.4 Completando la prueba

Independientemente de cómo quieras pensarlo, ahora tenemos nuestra estimación agrupada de la desviación estándar. De ahora en adelante, quitaré el subíndice p y me referiré a esta estimación como $\hat{\sigma}$. Excelente. Ahora volvamos a pensar en la prueba de hipótesis, ¿de acuerdo? La razón principal para calcular esta estimación agrupada era que sabíamos que sería útil a la hora de calcular la estimación del error estándar. Pero ¿error estándar de qué? En la prueba t de una muestra, fue el error estándar de la media muestral, $se(\bar{X})$, y dado que $se(\bar{X}) = \frac{\sigma}{\sqrt{N}}$ ESTE ERA el denominador de nuestra estadística t. Esta vez, sin embargo, tenemos dos medias muestrales. Y lo que nos interesa, específicamente, es la diferencia entre las dos $\bar{X}_1 - \bar{X}_2$ Como consecuencia, el error estándar por el que debemos dividir es de hecho el **error estándar de la diferencia** entre medias.

[Detalle técnico adicional [13]]

Tal como vimos con nuestra prueba de una muestra, la distribución muestral de este estadístico t es una distribución t (sorprendente, ¿verdad?) siempre que la hipótesis nula sea verdadera y se cumplan todos los supuestos de la prueba. Los grados de libertad, sin embargo, son ligeramente diferentes. Como de costumbre, podemos pensar que los grados de libertad son iguales al número de puntos de datos menos el número de restricciones. En este caso, tenemos N observaciones (N_1 en la muestra 1 y N_2 en la muestra 2) y 2 restricciones (las medias de la muestra). Entonces, los grados de libertad totales para esta prueba son $N - 2$.

Entonces, ¿por qué no usar estas desviaciones (es decir, ¿en qué medida la calificación de cada estudiante difiere de la calificación media en su tutoría?). Recuerda, una varianza es solo el promedio de un montón de desviaciones al cuadrado, así que hagamos eso. Matemáticamente, podríamos escribirlo así

$$frac\sum_{ik}(X_{ik} - \bar{X}_k)^2 N$$

donde la notación "$\sum_{ik}$" es una forma perezosa de decir "calcular una suma mirando a todos los estudiantes en todas las tutorías", ya que cada "$_{ik}$" corresponde a un estudiante.[a] Pero, como vimos en Section 8.5, calcular la varianza dividiendo por N produce una estimación sesgada de la varianza de la población. Y previamente necesitábamos dividir por $(N-1)$ para arreglar esto. Sin embargo, como mencioné en ese momento, la razón por la que existe este sesgo es que la estimación de la varianza se basa en la media muestral y, en la medida en que la media muestral no es igual a la media poblacional, puede sesgar sistemáticamente nuestra estimación de la media. ¡Pero esta vez nos basamos en dos medias muestrales! ¿Significa esto que tenemos más sesgos? Sí, eso significa. ¿Significa esto que ahora debemos dividir por $(N-2)$ en lugar de $(N-1)$, para calcular nuestra estimación de la varianza agrupada? Pues sí

$$\hat{\sigma}_p^2 = \frac{\sum_{ik}(X_{ik} - \bar{X}_k)^2}{N-2}$$

Ah, y si sacas la raíz cuadrada de esto entonces obtienes $\hat{\sigma}_p$, la estimación de la desviación estándar agrupada. En otras palabras, el cálculo de la desviación estándar agrupada no es nada especial. No es muy diferente al cálculo de la desviación estándar normal. — [a] Se introducirá una notación más correcta en Chapter 13.

[13] Siempre que las dos variables realmente tengan la misma desviación estándar, nuestra estimación del error estándar es

$$SE(\bar{X}_1 - \bar{X}_2) = \hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

y nuestro estadístico t es por lo tanto

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$$

11.4.1 Haciendo la prueba en jamovi

No es sorprendente que puedas ejecutar una prueba t de muestras independientes fácilmente en jamovi. La variable de resultado de nuestra prueba es la nota del alumnado, y los grupos se definen en función de la tutora de cada clase. Así que probablemente no te sorprendas mucho de que todo lo que tienes que hacer en jamovi es ir al análisis relevante ("Análisis" - "T-Tests" - "Independent Samples T-Test") y mover la variable de calificación al cuadro 'Variables dependientes' y la variable de la tutora en el cuadro 'Variable de agrupación', como se muestra en Figure 11.9.

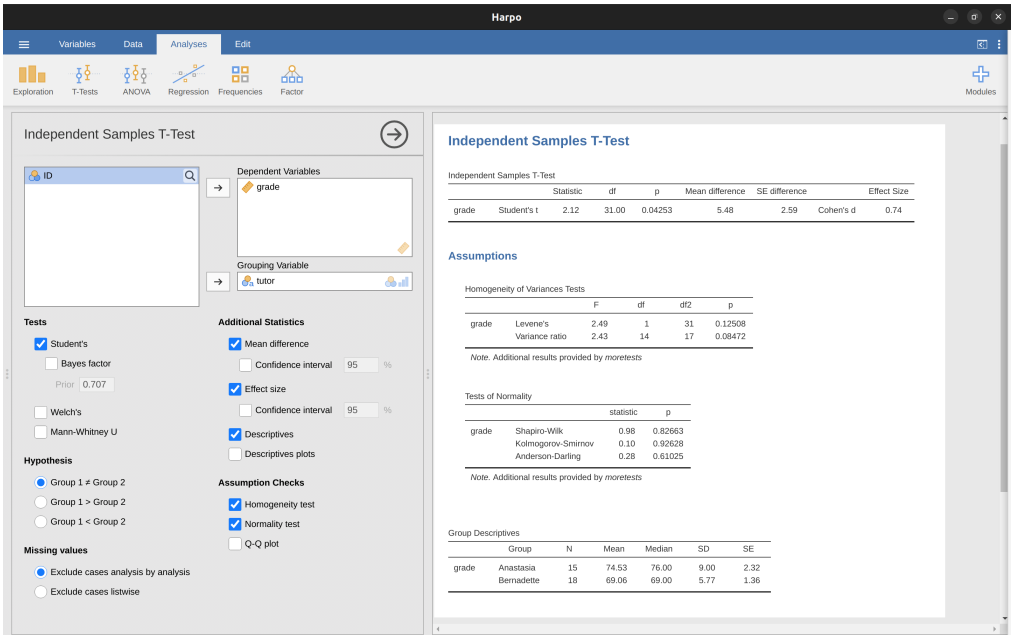


Figure 11.9: prueba t para muestras independientes en jamovi, con opciones verificadas para obtener resultados útiles

La salida tiene una forma muy familiar. Primero, te dice qué prueba se ejecutó y te dice el nombre de la variable dependiente que usaste. Luego informa los resultados de la prueba. Al igual que la última vez, los resultados de la prueba consisten en un estadístico t, los grados de libertad y el *valor p*. La sección final indica dos cosas: el intervalo de confianza y el tamaño del efecto. Hablaré sobre los tamaños del efecto más adelante. Del intervalo de confianza, sin embargo, debería hablar ahora.

Es muy importante tener claro a qué se refiere realmente este intervalo de confianza. Es un intervalo de confianza para la *diferencia* entre las medias de los grupos. En nuestro ejemplo, los y las estudiantes de Anastasia obtuvieron una calificación promedio de $74,53 $ y el alumnado de Bernadette tuvieron una calificación promedio de $69,06 $, por lo que la diferencia entre las medias de las dos muestras es $5,48 $. Pero, por supuesto, la diferencia entre las medias de la población puede ser mayor o menor que esto. El intervalo de confianza informado en Figure 11.10 te dice que si replicamos este estudio una y otra vez, entonces $ 95 % $ del tiempo, la verdadera diferencia en las medias estaría entre $ 0.20 $ y $ 10.76 $. Consulta Section 8.5 para recordar qué

significan los intervalos de confianza.

En cualquier caso, la diferencia entre los dos grupos es significativa (apenas), por lo que podríamos escribir el resultado usando un texto como este:

> La nota media en la clase de Anastasia fue de $74,5\%$ (desviación estándar $= 9,0$), mientras que la media en la clase de Bernadette fue de $69,1\%$ (desviación estándar $= 5,8$). La prueba t de Student de muestras independientes mostró que esta diferencia de 5.4% fue significativa ($t(31) = 2.1, p < .05, CI_{95} = [0.2, 10.8], d = .74$), lo que sugiere que se ha producido una diferencia genuina en los resultados del aprendizaje.

Observa que he incluido el intervalo de confianza y el tamaño del efecto en el bloque de estadísticos. La gente no siempre lo hace. Como mínimo, esperarías ver el estadístico t, los grados de libertad y el valor p. Entonces deberías incluir algo como esto como mínimo: $t(31) = 2.1, p < .05$. Si los estadísticos se salieran con la suya, todos también informarían el intervalo de confianza y probablemente también la medida del tamaño del efecto, porque son cosas útiles que hay que saber. Pero la vida real no siempre funciona de la forma en que los estadísticos quieren que lo haga, por lo que debes tomar una decisión en función de si crees que ayudará a tus lectores y, si estás escribiendo un artículo científico, el estándar editorial de la revista en cuestión. Algunas revistas esperan que informes los tamaños del efecto, otras no. Dentro de algunas comunidades científicas es una práctica estándar informar intervalos de confianza, en otras no lo es. Tendrás que averiguar qué espera tu audiencia. Pero, para que quede claro, si estás en mi clase, mi posición por defecto es que normalmente merce la pena incluir tanto el tamaño del efecto como el intervalo de confianza.

11.4.2 Valores t positivos y negativos

Antes de pasar a hablar de los supuestos de la prueba t, hay un punto adicional que quiero señalar sobre el uso de las pruebas t en la práctica. El primero se relaciona con el signo del estadístico t (es decir, si es un número positivo o negativo). Una preocupación muy común que tiene el alumnado cuando empieza a realizar su primera prueba t es que a menudo se obtienen valores negativos para el estadístico t y no saben cómo interpretarlo. De hecho, no es nada raro que dos personas que trabajan de forma independiente terminen con resultados casi idénticos, excepto que una persona tiene un valor de t negativo y la otra tiene un valor de t positivo. Si estás ejecutando una prueba bilateral, los valores p serán idénticos. En una inspección más detallada, los estudiantes notarán que los intervalos de confianza también tienen signos opuestos. Está bien. Siempre que esto suceda, encontrarás que las dos versiones de los resultados surgen de formas ligeramente diferentes de ejecutar la prueba t. Lo que está pasando aquí es muy sencillo. El estadístico t que calculamos aquí presenta la siguiente estructura

$$t = \frac{\text{media 1-media 2}}{SE}$$

Si "media 1" es mayor que "media 2", el estadístico t será positivo, mientras que si "media 2" es mayor, el estadístico t será negativo. De manera similar, el intervalo de confianza que informa jamovi es el intervalo de confianza para la diferencia "(media 1) menos (media 2)", que será la inversa de lo que obtendrías si estuvieras calculando el intervalo de confianza para la diferencia "(media 2) menos (media 1)".

De acuerdo, eso es bastante sencillo si lo piensas, pero ahora considera nuestra prueba t que compara la clase de Anastasia con la clase de Bernadette. ¿Cuál debería ser "media 1" y cuál "media 2"?. Es arbitrario. Sin embargo, necesitas designar uno de ellos como "media 1" y el otro como "media 2". No es sorprendente que la forma en que jamovi maneja esto también sea bastante arbitraria. En versiones anteriores del libro, solía tratar de explicarlo, pero después de un tiempo me di por vencida, porque en realidad no es tan importante y, para ser honesta, nunca puedo acordarme. Cada vez que obtengo un resultado significativo en la prueba t y quiero averiguar cuál es la media más grande, no trato de averiguarlo mirando el estadístico t. ¿Por qué me molestaría en hacer eso? Es una tontería. Es más fácil simplemente mirar las medias del grupo real ya que la salida de jamovi las muestra.

Esto es lo importante. Debido a que realmente no importa lo que te muestre jamovi, generalmente intento informar el estadístico t de tal manera que los números coincidan con el texto. Supongamos que lo que quiero escribir en mi informe es: *La clase de Anastasia tuvo calificaciones más altas que la clase de Bernadette*. El enunciado aquí implica que el grupo de Anastasia es el primero, por lo que tiene sentido informar del estadístico t como si la clase de Anastasia correspondiera al grupo 1. Si es así, escribiría *La clase de Anastasia tuvo* calificaciones más altas que la clase de Bernadette ($t(31) = 2.1, p = .04$).

(En realidad, no subrayaría la palabra "más alto" en la vida real, solo lo hago para enfatizar el punto de que "más alto" corresponde a valores t positivos). Por otro lado, supongamos que la frase que quiero usar tiene la clase de Bernadette en primer lugar. Si es así, tiene más sentido tratar a su clase como el grupo 1, y si es así, la redacción sería así: *La clase de Bernadette tenía* calificaciones más bajas que la clase de Anastasia ($t(31) = -2.1, p = .04$).

Debido a que estoy hablando de un grupo que tiene puntuaciones "más bajas" esta vez, es más sensato usar la forma negativa del estadístico t. Simplemente hace que se lea de manera más limpia.

Una última cosa: ten en cuenta que no puedes hacer esto para otros tipos de pruebas estadísticas. Funciona para las pruebas t, pero no tendría sentido para las pruebas de ji-cuadrado, las pruebas F o, de hecho, para la mayoría de las pruebas de las que hablo en este libro. Así que no generalices demasiado este consejo. Solo estoy hablando de pruebas t y nada más.

11.4.3 Supuestos de la prueba

Como siempre, la prueba de hipótesis se basa en algunos supuestos. Para la prueba t de Student hay tres supuestos, algunos de los cuales vimos anteriormente en el contexto de la prueba t de una muestra (consulta Supuestos de la prueba t de una muestra):

- *Normalidad.* Al igual que la prueba t de una muestra, se supone que los datos se distribuyen normalmente. Específicamente, asumimos que ambos grupos están normalmente distribuidos[14]. En la sección sobre [Comprobación de la normalidad

[14]Estrictamente hablando, es la **diferencia** en las medias lo que debería distribuirse normalmente, pero si ambos grupos tienen datos normalmente distribuidos, entonces la diferencia en las medias también estará normalmente repartida. En la práctica, el teorema central del límite nos asegura que, en general, las distribuciones de las medias de las dos muestras que se prueban se aproximarán a las distribuciones normales a medida que los tamaños de las muestras aumentan, independientemente de

de una muestra], analizaremos cómo probar la normalidad, y en [Prueba de datos no normales], analizaremos las posibles soluciones.

- *Independencia.* Una vez más, se supone que las observaciones se muestrean de forma independiente. En el contexto de la prueba de Student, esto se refiere a dos aspectos. En primer lugar, suponemos que las observaciones dentro de cada muestra son independientes entre sí (exactamente lo mismo que para la prueba de una muestra). Sin embargo, también asumimos que no hay dependencias entre muestras. Si, por ejemplo, resulta que incluiste a algunos participantes en ambas condiciones experimentales del estudio (por ejemplo, al permitir accidentalmente que la misma persona se inscribiera en diferentes condiciones), entonces hay algunas dependencias de muestras cruzadas que necesitarías tener en cuenta.
- *Homogeneidad de varianzas* (también llamada "homocedasticidad"). El tercer supuesto es que la desviación estándar de la población es la misma en ambos grupos. Se Puede probar este supuesto usando la prueba de Levene, de la que hablaré más adelante en el libro (en Section 13.6.1). Sin embargo, hay una solución muy simple para este supuesto, de la cual hablaré en la siguiente sección.

11.5 La prueba t de muestras independientes (prueba de Welch) {##sec-the-independent-samples-t-test-welch-test}

El mayor problema con el uso de la prueba de Student en la práctica es el tercer supuesto enumerado en la sección anterior. Se supone que ambos grupos tienen la misma desviación estándar. Esto rara vez es cierto en la vida real. Si dos muestras no tienen las mismas medias, ¿por qué deberíamos esperar que tengan la misma desviación estándar? Realmente no hay razón para esperar que este supuesto sea cierto. Más adelante hablaremos de cómo se puede verificar este supuesto, ya que aparece en varios lugares, no solo en la prueba t. Pero ahora hablaré sobre una forma diferente de la prueba t [@ Welch1947] que no se basa en este supuesto. En Figure 11.10 se muestra una ilustración gráfica de lo que asume la prueba t de Welch sobre los datos, para proporcionar un contraste con la versión de la prueba de Student en Figure 11.8. Admito que es un poco extraño hablar sobre la cura antes de hablar sobre el diagnóstico, pero da la casualidad de que la prueba de Welch se puede especificar como una de las opciones de 'Prueba T de muestras independientes' en jamovi, así que éste es probablemente el mejor lugar para hablar de ello.

La prueba de Welch es muy similar a la prueba de Student. Por ejemplo, el estadístico t que usamos en la prueba de Welch se calcula de la misma manera que para la prueba de Student. Es decir, calculamos la diferencia entre las medias muestrales y luego la dividimos por alguna estimación del error estándar de esa diferencia.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$$

La principal diferencia es que los cálculos del error estándar son diferentes. Si las dos poblaciones tienen diferentes desviaciones estándar, entonces es una tontería tratar de

las distribuciones de los datos subyacentes.

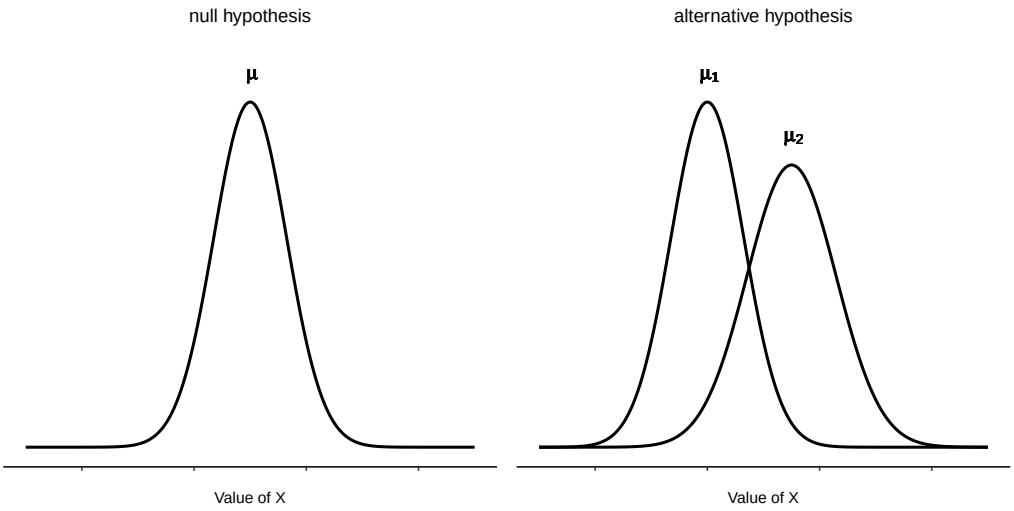


Figure 11.10: Ilustración gráfica de las hipótesis nula y alternativa asumidas por la prueba t de Welch. Al igual que la prueba de Student (Figure 11.9), asumimos que ambas muestras se extraen de una población normal; pero la hipótesis alternativa ya no requiere que las dos poblaciones tengan la misma varianza

calcular una estimación de la desviación estándar agrupada, porque está promediando manzanas y naranjas.[15]

[Detalle técnico adicional [16]]

La segunda diferencia entre Welch y Student es que los grados de libertad se calculan de forma muy diferente. En la prueba de Welch, los "grados de libertad" ya no tienen que ser un número entero, y no se corresponde del todo con la heurística "número de puntos de datos menos el número de restricciones" que he estado utilizando ahora.

11.5.1 Haciendo la prueba de Welch en jamovi

Si marcas la casilla de verificación de la prueba de Welch en el análisis que hicimos anteriormente, esto es lo que te da (Figure 11.11).

La interpretación de esta salida debería ser bastante obvia. Lee el resultado de la prueba de Welch de la misma manera que lo harías con la prueba de Student. Tienes tus estadísticos descriptivos, los resultados de las pruebas y alguna otra información. Así que todo eso es bastante fácil.

[15] Bueno, supongo que puedes promediar manzanas y naranjas, y lo que obtienes al final es un delicioso batido de frutas. Pero nadie piensa realmente que un batido de frutas sea una buena manera de describir las frutas originales, ¿verdad?

[16] pero aún se puede estimar el error estándar de la diferencia entre las medias muestrales, sólo que tiene un aspecto diferente

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$$

La razón por la que se calcula de esta manera va más allá del alcance de este libro. Lo que importa para nuestros propósitos es que el estadístico t que surge de la prueba t de Welch es en realidad algo diferente al que surge de la prueba t de Student.

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference		Effect Size
grade	Student's t	2.12	31.00	0.04253	5.48	2.59	Cohen's d	0.74
	Welch's t	2.03	23.02	0.05361	5.48	2.69	Cohen's d	0.72

Figure 11.11: resultados que muestran la prueba de Welch junto con la prueba t de Student predeterminada en jamovi

Excepto, excepto… nuestro resultado ya no es significativo. Cuando ejecutamos la prueba de Student, obtuvimos un efecto significativo, pero la prueba de Welch en el mismo conjunto de datos no lo es ($t(23.02) = 2.03, p = .054$). ¿Qué significa esto? ¿Debería cundir el pánico? Probablemente no. El hecho de que una prueba sea significativa y la otra no, no significa gran cosa, sobre todo porque he manipulado los datos para que esto sucediera. Como regla general, no es una buena idea esforzarse por intentar interpretar o explicar la diferencia entre un valor p de $ 0,049 y un valor p de $ 0,051. Si esto sucede en la vida real, la *diferencia* en estos valores p se debe casi con seguridad al azar. Lo que importa es que tengas un poco de cuidado al pensar qué prueba usas. La prueba de Student y la prueba de Welch tienen diferentes fortalezas y debilidades. Si las dos poblaciones realmente tienen varianzas iguales, entonces la prueba de Student es un poco más potente (menor tasa de error de tipo II) que la prueba de Welch. Sin embargo, si *no* tienen las mismas varianzas, entonces se violan los supuestos de la prueba de Student y es posible que no puedas confiar en ella; podrías terminar con una tasa de error Tipo I más alta. Así que es un intercambio. Sin embargo, en la vida real tiendo a preferir la prueba de Welch, porque casi nadie cree que las varianzas de la población sean idénticas.

11.5.2 Supuestos de la prueba

Los supuestos de la prueba de Welch son muy similares a los realizados por la prueba t de Student (ver Supuestos de la prueba, excepto que la prueba de Welch no asume homogeneidad de varianzas. Esto deja solo el supuesto de normalidad y el supuesto de independencia. Los detalles de estos supuestos son los mismos para la prueba de Welch que para la prueba de Student.

11.6 La prueba t de muestras pareadas

Independientemente de si estamos hablando de la prueba de Student o la prueba de Welch, una prueba t de muestras independientes está diseñada para usarse en una situación en la que tienes dos muestras que son, bueno, independientes entre sí. Esta situación surge naturalmente cuando los participantes se asignan aleatoriamente a una de dos condiciones experimentales, pero proporciona una aproximación muy pobre a otros tipos de diseños de investigación. En particular, un diseño de medidas repetidas, en el que se mide a cada participante (con respecto a la misma variable de resultado) en ambas condiciones experimentales, no es adecuado para el análisis mediante pruebas

t de muestras independientes. Por ejemplo, podríamos estar interesadas en saber si escuchar música reduce la capacidad de la memoria de trabajo de las personas. Con ese fin, podríamos medir la capacidad de la memoria de trabajo de cada persona en dos condiciones: con música y sin música. En un diseño experimental como este, [17] cada participante aparece en *ambos* grupos. Esto requiere que abordemos el problema de una manera diferente, usando la **prueba t de muestras pareadas**.

11.6.1 Los datos

El conjunto de datos que usaremos esta vez proviene de la clase de la Dra. Chico.[18] En su clase, los estudiantes realizan dos pruebas importantes, una al principio del semestre y otra más tarde. Por lo que cuenta, ella dirige una clase muy dura, que la mayoría de los estudiantes consideran un gran reto. Pero ella argumenta que al establecer evaluaciones difíciles, se alienta a los estudiantes a trabajar más duro. Su teoría es que la primera prueba es un poco como una "llamada de atención" para los estudiantes. Cuando se den cuenta de lo difícil que es realmente su clase, trabajarán más duro para la segunda prueba y obtendrán una mejor calificación. ¿Tiene razón? Para probar esto, importemos el archivo chico.csv a jamovi. Esta vez jamovi hace un buen trabajo durante la importación de atribuir correctamente los niveles de medida. El conjunto de datos chico contiene tres variables: una variable id que identifica a cada estudiante en la clase, la variable grade_test1 que registra la calificación del estudiante para la primera prueba y la variable grade_test2 que tiene las calificaciones para la segunda prueba.

Si miramos la hoja de cálculo de jamovi, parece que la clase es difícil (la mayoría de las calificaciones están entre 50% y 60%), pero parece que hay una mejora desde la primera prueba hasta la segunda.

Si echamos un vistazo rápido a los estadísticos descriptivos, en Figure 11.12, vemos que esta impresión parece confirmarse. Entre los 20 estudiantes, la calificación media para la primera prueba es del 57 %, pero aumenta al 58 % para la segunda prueba. Aunque, dado que las desviaciones estándar son 6,6 % y 6,4 % respectivamente, se empieza a sentir que tal vez la mejora es simplemente ilusoria; tal vez solo una variación aleatoria. Esta impresión se refuerza cuando ves las medias y los intervalos de confianza trazados en Figure 11.13a. Si nos basáramos únicamente en este gráfico y observáramos la amplitud de esos intervalos de confianza, tendríamos la tentación de pensar que la aparente mejora del rendimiento de los estudiantes es pura casualidad.

Sin embargo, esta impresión es incorrecta. Para ver por qué, echa un vistazo al gráfico de dispersión de las calificaciones de la prueba 1 frente a las calificaciones de la prueba 2, que se muestra en Figure 11.13b. En este gráfico, cada punto corresponde a las dos calificaciones de un estudiante determinado. Si su calificación para la prueba 1 (coordenada x) es igual a su calificación para la prueba 2 (coordenada y), entonces el punto cae en la línea. Los puntos que caen por encima de la línea son los estudiantes que tuvieron mejor rendimiento en la segunda prueba. Desde un punto de vista crítico, casi todos los puntos de datos se sitúan por encima de la línea diagonal: casi todos los estudiantes parecen haber mejorado su calificación, aunque solo sea un poco. Esto

[17] este diseño es muy similar al que motivó la prueba de McNemar (Section 10.7). Esto no debería ser una sorpresa. Ambos son diseños estándar de medidas repetidas que involucran dos medidas. La única diferencia es que esta vez nuestra variable de resultado está en una escala de intervalo (capacidad de la memoria de trabajo) en lugar de una variable de escala nominal binaria (una pregunta de sí o no).

[18] En este punto tenemos a los Drs. Harpo, Chico y Zeppo. No hay premios por adivinar quién es el Dr. Groucho.

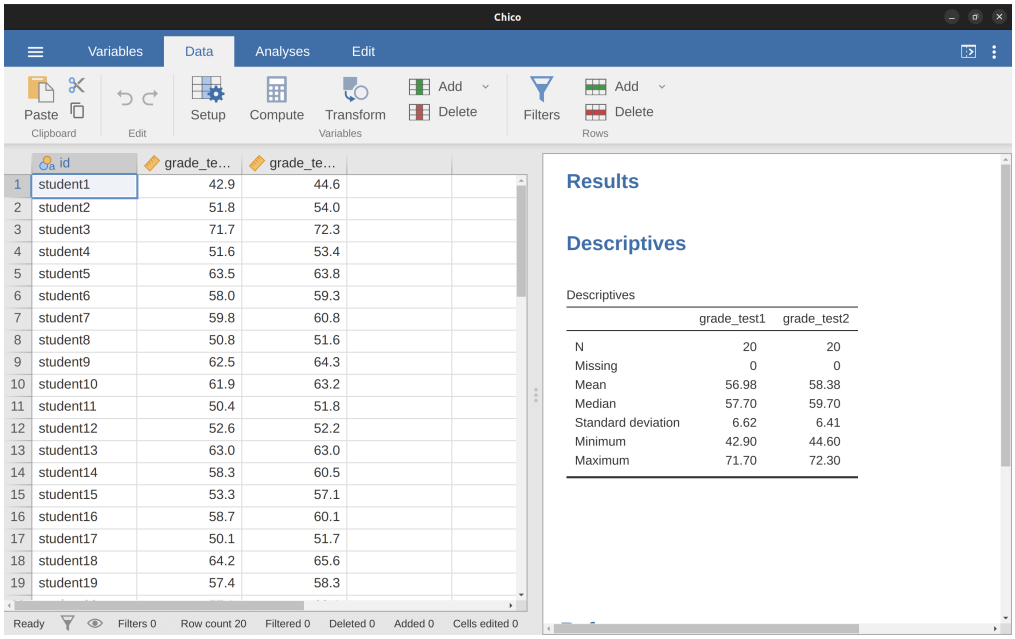


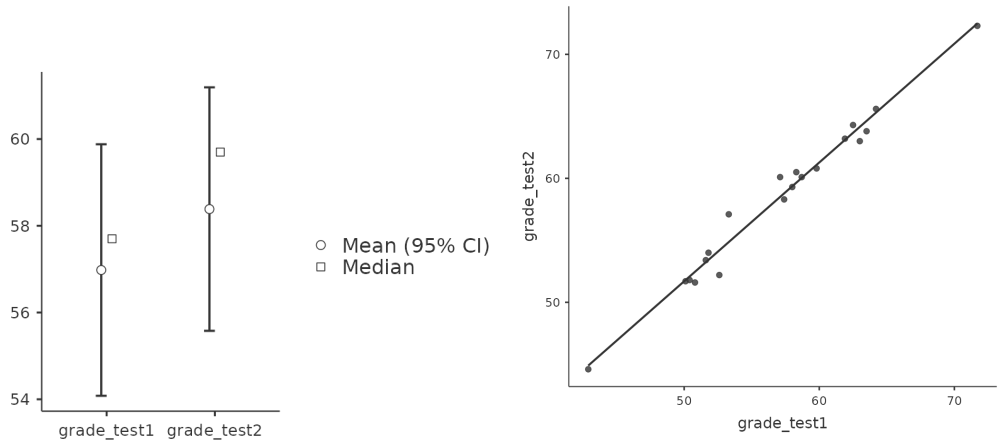
Figure 11.12: descriptivos para las dos variables de prueba de grado en el conjunto de datos de chico

sugiere que deberíamos observar la mejora realizada por cada estudiante de una prueba a la siguiente y tratarla como nuestros datos brutos. Para hacer esto, necesitaremos crear una nueva variable para la mejora que hace cada estudiante y agregarla al conjunto de datos de chico. La forma más sencilla de hacer esto es calcular una nueva variable, con la expresión calificación prueba2 - calificación prueba1.

Una vez que hayamos calculado esta nueva variable de mejora, podemos dibujar un histograma que muestre la distribución de estas puntuaciones de mejora, que se muestra en Figure 11.14. Si nos fijamos en el histograma, está muy claro que hay una mejora real aquí. La gran mayoría de los estudiantes obtuvo una puntuación más alta en la prueba 2 que en la prueba 1, lo que se refleja en el hecho de que casi todo el histograma está por encima de cero.

11.6.2 ¿Qué es la prueba t de muestras pareadas?

A la luz de la exploración anterior, pensemos en cómo construir una prueba t apropiada. Una posibilidad sería intentar ejecutar una prueba t de muestras independientes utilizando grade_test1 y grade_test2 como las variables de interés. Sin embargo, esto es claramente lo incorrecto, ya que la prueba t de muestras independientes asume que no existe una relación particular entre las dos muestras. Sin embargo, está claro que esto no es cierto en este caso debido a la estructura de medidas repetidas de los datos. Para usar el lenguaje que introduje en la última sección, si intentáramos hacer una prueba t de muestras independientes, estaríamos fusionando las diferencias **dentro del sujeto** (que es lo que nos interesa probar) con la variabilidad **entre sujetos** (que no nos interesa).



(a) (b)

Figure 11.13: Nota media para la prueba 1 y la prueba 2, con intervalos de confianza del 95% asociados (a). Diagrama de dispersión que muestra las calificaciones individuales para la prueba 1 y la prueba 2 (b).

La solución al problema es obvia, espero, ya que ya hicimos todo el trabajo duro en la sección anterior. En lugar de ejecutar una prueba t de muestras independientes en grade_test1 y grade_test2, ejecutamos una prueba t de una muestra en la variable de diferencia dentro del sujeto, mejora. Para formalizar esto un poco, si X_{i1} es la puntuación que obtuvo el i-ésimo participante en la primera variable, y X_{i2} es la puntuación que obtuvo la misma persona en la segunda, entonces la puntuación de diferencia es:

$$D_i = X_{i1} - X_{i2}$$

Ten en cuenta que las puntuaciones de diferencia son la variable 1 menos la variable 2 y no al revés, por lo que si queremos que la mejora corresponda a una diferencia de valor positivo, en realidad queremos que la "prueba 2" sea nuestra "variable 1". Igualmente, diríamos que $\mu_D = \mu_1 - \mu_2$ es la media poblacional para esta variable diferencia. Entonces, para convertir esto en una prueba de hipótesis, nuestra hipótesis nula es que esta diferencia de medias es cero y la hipótesis alternativa es que no lo es.

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

Asumiendo que estamos hablando de una prueba bilateral. Esto es más o menos idéntico a la forma en que describimos las hipótesis para la prueba t de una muestra. La única diferencia es que el valor específico que predice la hipótesis nula es 0. Por lo tanto, nuestro estadístico t también se define más o menos de la misma manera. Si hacemos que $\bar{D}$ denote la media de las puntuaciones de diferencia, entonces

$$t = \frac{\bar{D}}{SE(\bar{D})}$$

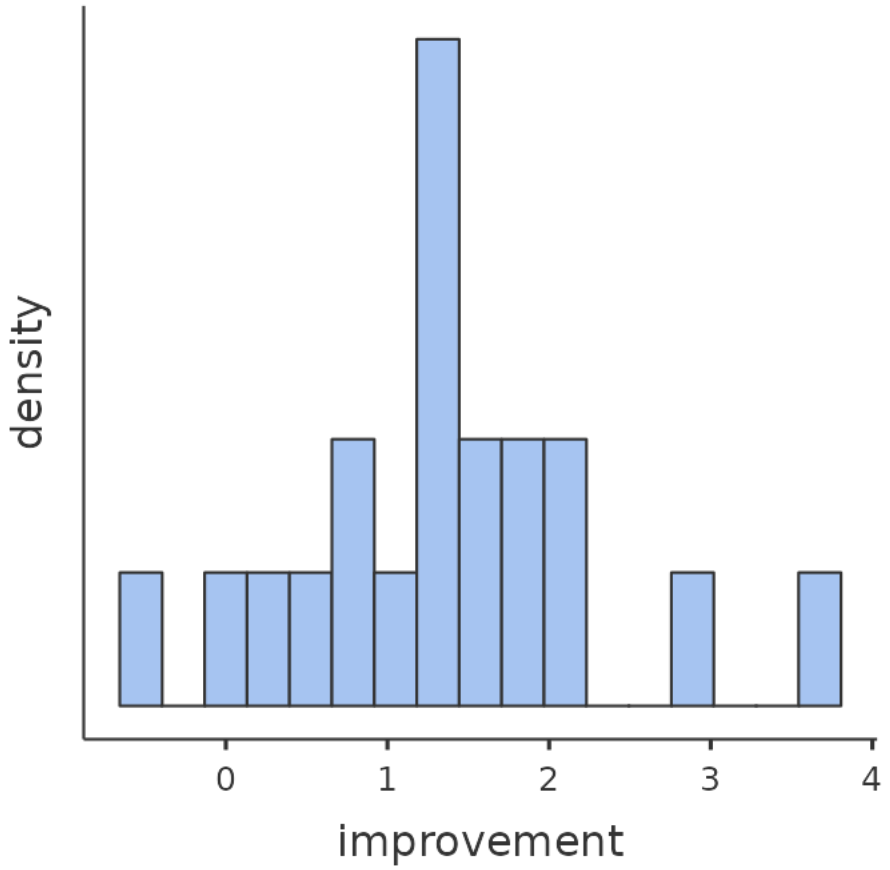


Figure 11.14: Histograma que muestra la mejora realizada por cada estudiante en la clase del Dr. Chico. Ten en cuenta que casi toda la distribución está por encima de cero: la gran mayoría de los estudiantes mejoraron su rendimiento des la primera prueba a la segunda

que es

$$t = \frac{\bar{D}}{\frac{\hat{\sigma}_D}{\sqrt{N}}}$$

donde $\hat{\sigma}_D$ es la desviación estándar de las puntuaciones de diferencia. Dado que esta es solo una prueba t ordinaria de una muestra, sin nada especial, los grados de libertad siguen siendo $N - 1$. Y eso es todo. La prueba t de muestras pareadas realmente no es una prueba nueva en absoluto. Es una prueba t de una muestra, pero aplicada a la diferencia entre dos variables. En realidad es muy simple. La única razón por la que merece una discusión tan larga como la que acabamos de ver es que debes poder reconocer *cuándo* una prueba de muestras pareadas es apropiada y comprender *por qué* es mejor que una prueba t de muestras independientes.

11.6.3 Haciendo la prueba en jamovi

¿Cómo se hace una prueba t de muestras pareadas en jamovi? Una posibilidad es seguir el proceso que describí anteriormente. Es decir, crea una variable de "diferencia" y luego ejecutas una prueba t de una muestra sobre eso. Como ya hemos creado una variable llamada mejora, hagámoslo y veamos qué obtenemos, Figure 11.15.

One Sample T-Test

One Sample T-Test

						95% Confidence Interval			
		Statistic	df	p	Mean difference	Lower	Upper		Effect Size
improvement	Student's t	6.48	19.00	< .00001	1.40	0.95	1.86	Cohen's d	1.45

Figure 11.15: resultados que muestran una prueba t de una muestra en puntajes de diferencia emparejados

El resultado que se muestra en Figure 11.15 tiene (obviamente) el mismo formato que tenía la última vez que usamos el análisis de prueba t de una muestra (Section 11.2), y confirma nuestra intuición. Hay una mejora promedio de 1.4% de la prueba 1 a la prueba 2, y esto es significativamente diferente de 0 ($t(19) = 6.48, p < .001$).

Sin embargo, supongamos que eres perezosa y no quieres hacer todo el esfuerzo de crear una nueva variable. O tal vez solo quieras mantener clara la diferencia entre las pruebas de una muestra y muestras pareadas. Si es así, puedes usar el análisis 'Prueba T de muestras emparejadas' de jamovi, obteniendo los resultados que se muestran en Figure 11.16.

Las cifras son idénticas a las de la prueba de una muestra, lo que, por supuesto, tiene que ser así, dado que la prueba t de muestras pareadas no es más que una prueba de una muestra.

Paired Samples T-Test

Paired Samples T-Test

								95% Confidence Interval			
			statistic	df	p	Mean difference	SE difference	Lower	Upper		Effect Size
grade_test2	grade_test1	Student's t	6.48	19.00	< .00001	1.40	0.22	0.95	1.86	Cohen's d	1.45

Figure 11.16: resultados que muestran una prueba t de muestra pareada. Comparar con Figure 11.15

11.7 Pruebas unilaterales

Al presentar la teoría de las pruebas de hipótesis nulas, mencioné que hay algunas situaciones en las que es apropiado especificar una prueba unilateral (ver Section 9.4.3). Hasta ahora, todas las pruebas t han sido pruebas bilaterales. Por ejemplo, cuando especificamos una prueba t de una muestra para las calificaciones en la clase del Dr. Zeppo, la hipótesis nula fue que la verdadera media era 67.5%. La hipótesis alternativa era que la verdadera media era mayor o menor que 67.5%. Supongamos que solo nos interesa saber si la media real es mayor que $67,5\%$ y no tenemos ningún interés en probar si la media real es menor que $67,5\%$. Si es así, nuestra hipótesis nula sería que la verdadera media es $67,5\%$ o menos, y la hipótesis alternativa sería que la verdadera media es mayor que $67,5\%$. En jamovi, para el análisis 'Prueba T de una muestra', puedes especificar esto haciendo clic en la opción '> Valor de prueba', en 'Hipótesis'. Cuando hayas hecho esto, obtendrás los resultados que se muestran en Figure 11.17.

One Sample T-Test

One Sample T-Test

						95% Confidence Interval			
		Statistic	df	p	Mean difference	Lower	Upper		Effect Size
x	Student's t	2.25	19.00	0.01807	4.80	1.12	Inf	Cohen's d	0.50

Note. H_a μ > 67.5

Figure 11.17: resultados de jamovi que muestran una 'Prueba T de una muestra' donde la hipótesis real es unilateral, es decir, que la media real es mayor que 67.5%

Ten en cuenta que hay algunos cambios con respecto a la salida que vimos la última vez. Lo más importante es el hecho de que la hipótesis real ha cambiado, para reflejar la prueba diferente. La segunda cosa a tener en cuenta es que aunque el estadístico t y los grados de libertad no han cambiado, el valor p sí lo ha hecho. Esto se debe a que la prueba unilateral tiene una región de rechazo diferente de la prueba bilateral. Si has olvidado por qué es esto y qué significa, puede que te resulte útil volver a leer Chapter 9 y Section 9.4.3 en particular. La tercera cosa a tener en cuenta es que el intervalo de confianza también es diferente: ahora informa un intervalo de confianza "unilateral" en lugar de uno bilateral. En un intervalo de confianza de dos colas, estamos tratando de encontrar los números a y b de modo que estemos seguros de que, si tuviéramos que repetir el estudio muchas veces, entonces 95% del tiempo la media estaría entre a y b.

En un intervalo de confianza unilateral, estamos tratando de encontrar un solo número a tal que estemos seguros de que 95% del tiempo la verdadera media sería mayor que a (o menor que a si seleccionaste la Medida 1 < Medida 2 en la sección 'Hipótesis').

Así es como se hace una prueba t unilateral de una muestra. Sin embargo, todas las versiones de la prueba t pueden ser unilaterales. Para una prueba t de muestras independientes, podrías tener una prueba unilateral si solo estás interesada en probar si el grupo A tiene puntuaciones más altas que el grupo B, pero no tienes interés en averiguar si el grupo B tiene puntuaciones más altas que el grupo R. Supongamos que, para la clase del Dr. Harpo, quisieras ver si los estudiantes de Anastasia tenían calificaciones más altas que las de Bernadette. Para este análisis, en las opciones de 'Hipótesis', especifica que 'Grupo 1 > Grupo2'. Deberías obtener los resultados que se muestran en Figure 11.18.

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference		Effect Size
grade	Student's t	2.12	31.00	0.02126	5.48	2.59	Cohen's d	0.74
	Welch's t	2.03	23.02	0.02680	5.48	2.69	Cohen's d	0.72

Note. H_a $\mu_{Anastasia} > \mu_{Bernadette}$

Figure 11.18: resultados de jamovi que muestran una 'Prueba t de muestras independientes' donde la hipótesis real es unilateral, es decir, que los estudiantes de Anastasia obtuvieron calificaciones más altas que los de Bernadette

Una vez más, la salida cambia de forma predecible. La definición de la hipótesis alternativa ha cambiado, el valor p ha cambiado y ahora informa un intervalo de confianza unilateral en lugar de uno bilateral.

¿Qué pasa con la prueba t de muestras pareadas? Supongamos que quisiéramos probar la hipótesis de que las calificaciones suben de la prueba 1 a la prueba 2 en la clase del Dr. Zeppo y no estamos preparados para considerar la idea de que las calificaciones bajan. En jamovi, harías esto especificando, en la opción 'Hipótesis', que grade_test2 ('Medida 1' en jamovi, porque copiamos esto primero en el cuadro de pares de variables) > grade test1 ('Medida 2' en jamovi). Deberías obtener los resultados que se muestran en Figure 11.19.

Una vez más, la salida cambia de forma predecible. La hipótesis ha cambiado, el valor p ha cambiado y el intervalo de confianza ahora es unilateral.

11.8 Tamaño del efecto

La medida del tamaño del efecto más utilizada para una prueba t es la **d de Cohen** (Cohen, 1988). En principio, se trata de una medida muy sencilla, pero que presenta algunos inconvenientes cuando se profundiza en los detalles. El propio Cohen la definió principalmente en el contexto de una prueba t de muestras independientes, específicamente la prueba de Student. En ese contexto, una forma natural de definir el tamaño

Paired Samples T-Test

Paired Samples T-Test

								95% Confidence Interval			
			statistic	df	p	Mean difference	SE difference	Lower	Upper		Effect Size
grade_test2	grade_test1	Student's t	6.48	19.00	< .00001	1.40	0.22	1.03	Inf	Cohen's d	1.45

Note. H_a $\mu_{\text{Measure 1 - Measure 2}} > 0$

Figure 11.19: resultados de jamovi que muestran una 'Prueba T de muestras emparejadas' donde la hipótesis real es unilateral, es decir, calificación prueba2 ('Medida 1') > calificación prueba1 ('Medida 2')

del efecto es dividir la diferencia entre las medias por una estimación de la desviación estándar. En otras palabras, estamos buscando calcular algo similar a esto:

$$d = \frac{\text{(media 1)} - \text{(media 2)}}{\text{desviación estándar}}$$

y sugirió una guía aproximada para interpretar d en Table 11.3.

Table 11.3: Una guía (muy) aproximada para interpretar la d de Cohen. Mi recomendación personal es no usarlos a ciegas. El estadístico d tiene una interpretación natural en sí mismo. Vuelve a describir la diferencia de medias como el número de desviaciones estándar que separa esas medias. Por lo tanto, generalmente es una buena idea pensar en lo que eso significa en términos prácticos. En algunos contextos, un efecto 'pequeño' podría ser de gran importancia práctica. En otras situaciones, un efecto 'grande' puede no ser tan interesante

d-value	**rough interpretation**
about 0.2	"small" effect
about 0.5	"moderate" effect
about 0.8	"large" effect

Se podría pensar que no hay ninguna ambigüedad, pero no es así. Esto se debe en gran parte a que Cohen no fue demasiado específico sobre lo que pensó que debería usarse como la medida de la desviación estándar (en su defensa, estaba tratando de hacer un punto más amplio en su libro, no criticar los pequeños detalles). Como comenta @ McGrath2006, hay varias versiones diferentes de uso común, y cada autor tiende a adoptar una notación ligeramente diferente. En aras de la simplicidad (en oposición a la precisión), usaré d para referirme a cualquier estadístico que calcule a partir de la muestra, y usaré δ para referirme a un teórico efecto poblacional. Obviamente, eso significa que hay varias cosas diferentes, todas llamadas d.

Mi sospecha es que el único momento en el que querrías la d de Cohen es cuando estás ejecutando una prueba t, y jamovi tiene una opción para calcular el tamaño del efecto para todos los diferentes tipos de prueba t que proporciona.

11.8.1 d de Cohen de una muestra

La situación más sencilla de considerar es la correspondiente a una prueba t de una muestra. En este caso, se trata de una media muestral $\bar{X}$ y una media poblacional (hipotética) μ_0 con la que compararla. No solo eso, en realidad solo hay una forma sensata de estimar la desviación estándar de la población. Simplemente usamos nuestra estimación habitual $\hat{\sigma}$. Por lo tanto, terminamos con la siguiente como la única forma de calcular d

$$d = \frac{\bar{X} - \mu_0}{\hat{\sigma}}$$

Cuando volvemos a mirar los resultados en Figure 11.6, el valor del tamaño del efecto es $d = 0,50$ de Cohen. Entonces, en general, los estudiantes de psicología de la clase del Dr. Zeppo obtienen calificaciones ($media = 72,3\%$) que son alrededor de 0,5 desviaciones estándar más altas que el nivel que esperarías ($67,5\%$) si tuvieran un rendimiento igual que otros estudiantes. A juzgar por la guía aproximada de Cohen, este es un tamaño de efecto moderado.

11.8.2 d de Cohen a partir de una prueba t de Student

La mayoría de los debates sobre la d de Cohen se centran en una situación que es análoga a la prueba t de Student de muestras independientes, y es en este contexto que la historia se vuelve más complicada, ya que hay varias versiones diferentes de d que es posible que quieras utilizar en esta situación. Para entender por qué hay múltiples versiones de d, es útil tomarse el tiempo para escribir una fórmula que corresponda al verdadero tamaño del efecto poblacional δ. Es bastante sencilla,

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

donde, como es habitual, μ_1 y μ_2 son las medias poblacionales correspondientes al grupo 1 y al grupo 2 respectivamente, y σ es la desviación estándar (igual para ambas poblaciones). La forma obvia de estimar δ es hacer exactamente lo mismo que hicimos en la prueba t, es decir, usar las medias muestrales como la línea superior y una desviación estándar combinada para la línea inferior.

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}_p}$$

donde $\hat{\sigma}_p$ es exactamente la misma medida de desviación estándar agrupada que aparece en la prueba t. Esta es la versión más utilizada de la d de Cohen cuando se aplica al resultado de una prueba t de Student, y es la que se proporciona en jamovi. A veces se la denomina estadístico g de Hedges (Hedges, 1981).

Sin embargo, hay otras posibilidades que describiré brevemente. En primer lugar, es posible que tengas razones para querer usar solo uno de los dos grupos como base para calcular la desviación estándar. Este enfoque (a menudo llamado $\triangle$ de Glass, pronunciado delta) solo tiene sentido cuando tienes una buena razón para tratar a uno de los dos grupos como un reflejo más puro de la "variación natural" del otro. Esto

puede suceder si, por ejemplo, uno de los dos grupos es un grupo de control. En segundo lugar, recuerda que en el cálculo habitual de la desviación estándar agrupada dividimos entre $N - 2$ para corregir el sesgo en la varianza de la muestra. En una versión de la d de Cohen se omite esta corrección y en su lugar se divide por N. Esta versión tiene sentido principalmente cuando intentas calcular el tamaño del efecto muestral en lugar de estimar el tamaño del efecto poblacional. Finalmente, hay una versión llamada g de Hedge, basada en Hedges & Olkin (1985), que señala que existe un pequeño sesgo en la estimación habitual (agrupada) para la d de Cohen.[19]

En cualquier caso, ignorando todas aquellas variaciones que podrías utilizar si quisieras, echemos un vistazo a la versión por defecto en jamovi. En Figure 11.10 la de de Cohen es $d = 0.74$, lo que indica que las calificaciones de los estudiantes en la clase de Anastasia son, en promedio, 0.74 desviaciones estándar más altas que las calificaciones de los estudiantes en la clase de Bernadette. Para una prueba de Welch, el tamaño del efecto estimado es el mismo (Figure 11.12).

11.8.3 d de Cohen a partir de una prueba de muestras pareadas

Finalmente, ¿qué debemos hacer para una prueba t de muestras pareadas? En este caso, la respuesta depende de lo que estés tratando de hacer. jamovi asume que deseas medir los tamaños de su efecto en relación con la distribución de las puntuaciones de diferencia, y la medida de d que calcula es:

$$d = \frac{\bar{D}}{\hat{\sigma}_D}$$

donde $\hat{\sigma}_D$ es la estimación de la desviación estándar de las diferencias. En Figure 11.16 la de de Cohen es $d = 1,45$, lo que indica que las puntuaciones de la calificación en el momento 2 son, en promedio, $1,45$ desviaciones estándar más altas que las puntuaciones de la calificación en el momento 1.

Esta es la versión de d de Cohen que se informa en el análisis 'Prueba T de muestras emparejadas' de jamovi. La única pega es averiguar si esta es la medida que deseas o no. En la medida en que te importen las consecuencias prácticas de tu investigación, a menudo querrás medir el tamaño del efecto en relación con las variables *originales*, no las *puntuaciones de diferencia* (p. ej., la mejora del 1% en la clase del Dr. Chico con el tiempo es bastante pequeña cuando se compara con la cantidad de variación entre estudiantes en las calificaciones), en cuyo caso usa las mismas versiones de la d de Cohen que usarías para una prueba de Student o Welch. No es tan sencillo hacer esto en jamovi; básicamente, debes cambiar la estructura de los datos en la vista de hoja de cálculo, por lo que no entraré en eso aquí [20], pero la d de Cohen para esta situación es bastante diferente: es $ 0.22 $ que es bastante pequeña cuando se evalúa en la escala de las variables originales.

[19] Introducen una pequeña corrección al multiplicar el valor habitual de d por $\frac{(N-3)}{(N-2.25)}$.

[20] si estás interesada, puedes ver cómo se hizo esto en el archivo chico2.omv

11.9 Comprobando la normalidad de una muestra

Todas las pruebas que hemos discutido hasta ahora en este capítulo han asumido que los datos están normalmente distribuidos. Este supuesto suele ser bastante razonable, porque el teorema cntral del límite (ver Section 8.3.3) tiende a garantizar que muchas cantidades del mundo real se distribuyan normalmente. Cada vez que sospeches que tu variable es *en realidad* un promedio de muchas cosas diferentes, existe una gran probabilidad de que se distribuya normalmente, o al menos lo suficientemente cerca de lo normal como para que puedas usar pruebas t. Sin embargo, la vida no viene con garantías y, además, hay muchas formas en las que puedes terminar con variables que son muy anormales. Por ejemplo, cada vez que pienses que tu variable es en realidad el mínimo de muchas cosas diferentes, es muy probable que termine bastante sesgada. En psicología, los datos de tiempo de respuesta (TR) son un buen ejemplo de esto. Si supones que hay muchas cosas que podrían desencadenar una respuesta de un participante humano, entonces la respuesta real ocurrirá la primera vez que ocurra uno de estos eventos desencadenantes.[21] Esto significa que los datos de TR son sistemáticamente no normales. De acuerdo, entonces, si todas las pruebas asumen la normalidad, y la mayor parte, pero no siempre, la satisfacen (al menos aproximadamente) los datos del mundo real, ¿cómo podemos verificar la normalidad de una muestra? En esta sección analizo dos métodos: gráficos QQ y la prueba de Shapiro-Wilk.

11.9.1 Gráficos QQ

Una forma de verificar si una muestra viola el supuesto de normalidad es dibujar un **"Gráfico QQ"** (Gráfico Cuantil-Cuantil). Esto te permite verificar visualmente si estás viendo alguna infracción sistemática. En un gráfico QQ, cada observación se representa como un solo punto. La coordenada x es el cuantil teórico en el que debería caer la observación si los datos se distribuyeran normalmente (con la media y la varianza estimadas a partir de la muestra), y en la coordenada y está el cuantil real de los datos dentro de la muestra. Si los datos son normales, los puntos deben formar una línea recta. Por ejemplo, veamos qué sucede si generamos datos tomando muestras de una distribución normal y luego dibujando un gráfico QQ. Los resultados se muestran en Figure 11.20.

Como puedes ver, estos datos forman una línea bastante recta; ¡lo cual no es una sorpresa dado que los cogimos como muestra de una distribución normal! Por el contrario, echa un vistazo a los dos conjuntos de datos que se muestran en Figure 11.21. Los paneles superiores muestran el histograma y un gráfico QQ para un conjunto de datos que está muy sesgado: el gráfico QQ se curva hacia arriba. Los paneles inferiores muestran los mismos gráficos para un conjunto de datos de cola pesada (es decir, alta curtosis): en este caso, el gráfico QQ se aplana en el medio y se curva bruscamente en cada extremo.

11.9.2 Gráficos QQ para pruebas t independientes y pareadas

En nuestros análisis anteriores mostramos cómo realizar en jamovi una prueba t independiente (Figure 11.10) y una prueba t de muestras pareadas (Figure 11.16). Y para estos análisis, jamovi proporciona una opción para mostrar un gráfico QQ para las puntuaciones de diferencia (que jamovi llama "residuales"), que es una mejor manera de verificar el supuesto de normalidad. Cuando seleccionamos esta opción para estos

[21] esta es una simplificación excesiva.

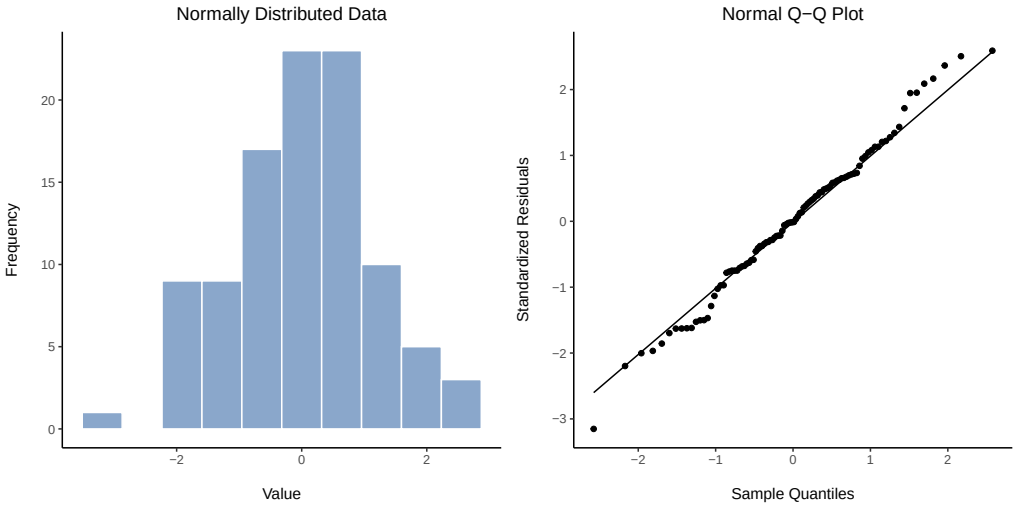


Figure 11.20: Histograma (panel (a)) y gráfico QQ normal (panel (b)) de normal.data, una muestra distribuida normalmente con 100 observaciones. El estadístico de Shapiro-Wilk asociado con estos datos es $W = .99$, lo que indica que no se detectaron desviaciones significativas de la normalidad ($p = .54$)

análisis, obtenemos los gráficos QQ que se muestran en Figure 11.22 y Figure 11.23, respectivamente. Mi interpretación es que estos gráficos muestran que las puntuaciones de diferencia están razonablemente distribuidas normalmente, ¡así que estamos listos para comenzar!

11.9.3 Pruebas de Shapiro-Wilk

Los diagramas QQ ofrecen una buena manera de verificar informalmente la normalidad de tus datos, pero a veces querrás hacer algo un poco más formal y la **prueba de Shapiro-Wilk** [@ Shapiro1965] es probablemente lo que estás buscando .[22] Como era de esperar, la hipótesis nula que se prueba es que un conjunto de N observaciones se distribuye normalmente.

[Detalle técnico adicional [23]]

Para obtener el estadístico de Shapiro-Wilk en las pruebas t de jamovi, marca la opción

[22] O eso, o la prueba de Kolmogorov-Smirnov, que probablemente sea más tradicional que la de Shapiro-Wilk. Aunque la mayoría de las cosas que he leído parecen sugerir que Shapiro-Wilk es la mejor prueba de normalidad, Kolomogorov Smirnov es una prueba de propósito general de equivalencia distribucional que se puede adaptar para manejar otros tipos de pruebas de distribución. En jamovi se prefiere la prueba de Shapiro-Wilk.

[23] la prueba estadística que calcula se denota convencionalmente como W y se calcula de la siguiente manera. Primero, clasificamos las observaciones en orden creciente y dejamos que $\bar{X}_1$ sea el valor más pequeño de la muestra, X_2 el segundo más pequeño y así sucesivamente. Entonces el valor de W viene dado por

$$W = \frac{(\sum_{i=1}^{N} a_i X_i)^2}{\sum_{i=1}^{N} (X_i - \bar{X})^2}$$

donde $\bar{X}$ es la media de las observaciones, y los valores de a_i son … algo complicado que está fuera del alcance de un texto introductorio.

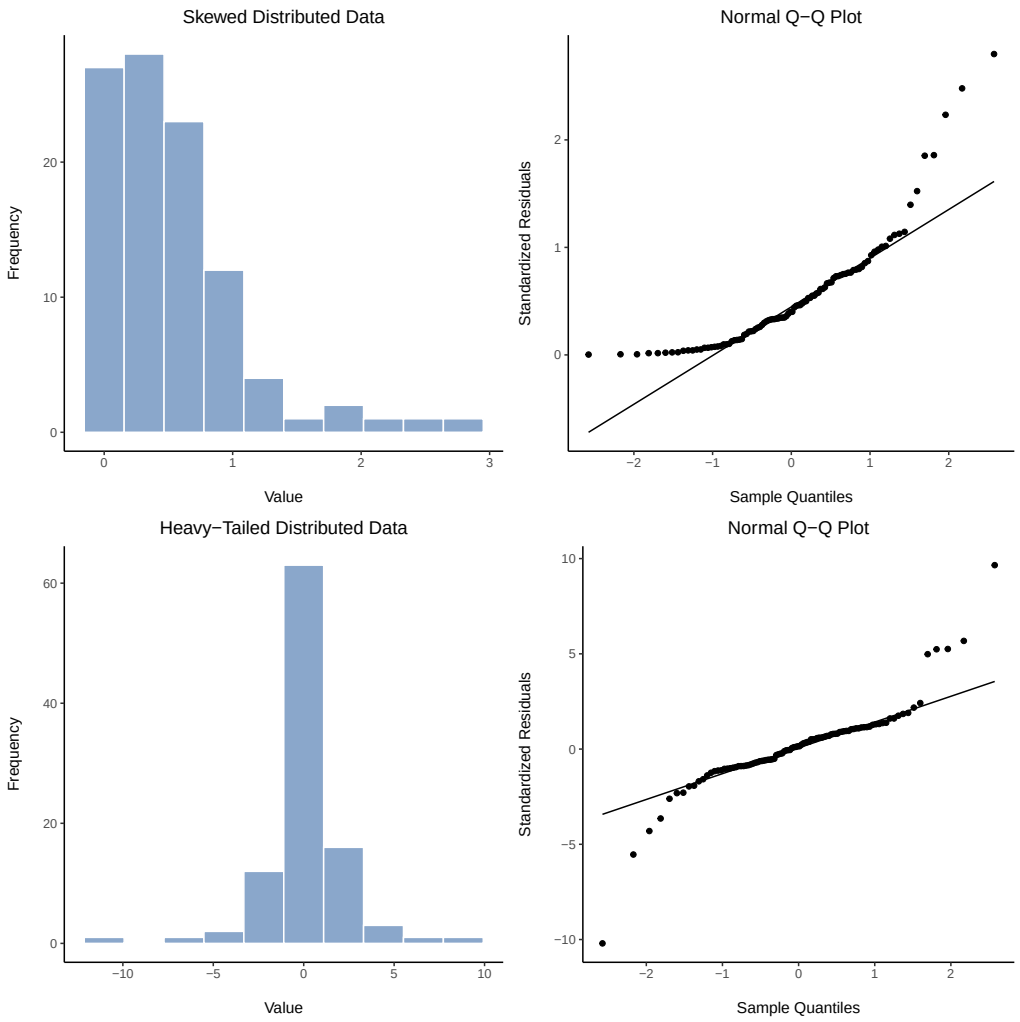


Figure 11.21: en la fila superior, un histograma y un gráfico QQ normal de las observaciones de 100 en un conjunto de datos sesgados. La asimetría de los datos aquí es de 1.88 y se refleja en un gráfico QQ que se curva hacia arriba. Como consecuencia, el estadístico de Shapiro-Wilk es $W = .80$, lo que refleja una desviación significativa de la normalidad ($p < .001$). La fila inferior muestra los mismos gráficos para un conjunto de datos de cola pesada, que nuevamente consta de 100 observaciones. En este caso, las colas pesadas en los datos producen una curtosis alta ($ 6,57 $) y hacen que el gráfico QQ se aplane en el medio y se curve bruscamente a ambos lados. El estadístico resultante de Shapiro-Wilk es $W = .75$, lo que nuevamente refleja una falta de normalidad significativa ($p < .001$)

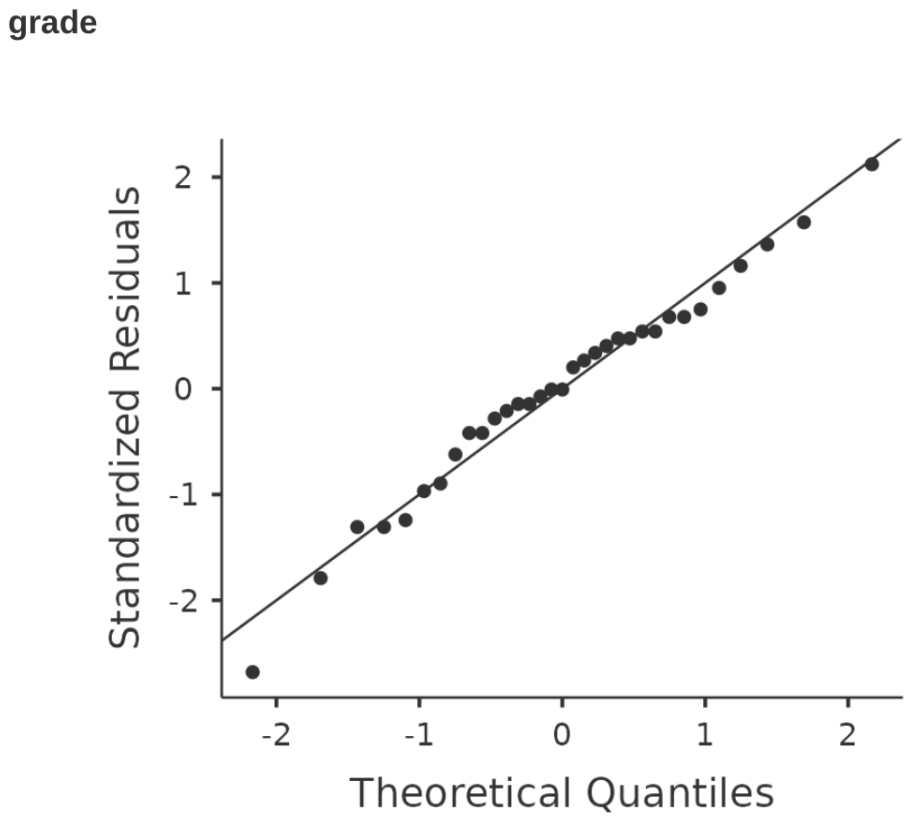


Figure 11.22: diagrama QQ para el análisis de prueba t independiente que se muestra en Figure 11.10

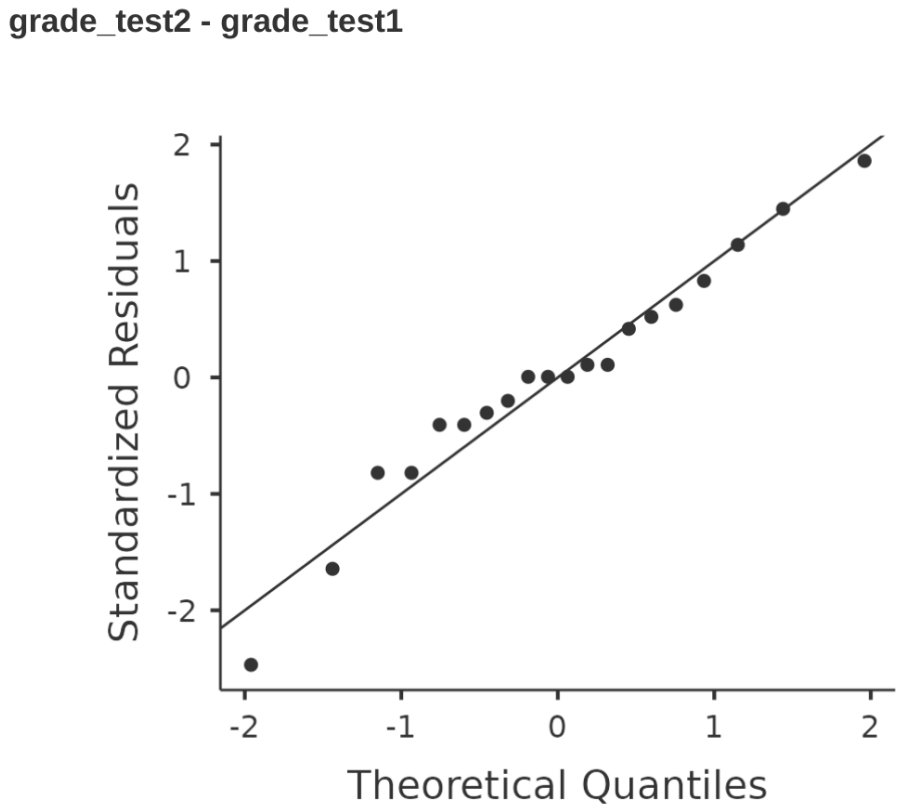


Figure 11.23: gráfico QQ para el análisis de prueba t de muestras emparejadas que se muestra en Figure 11.16

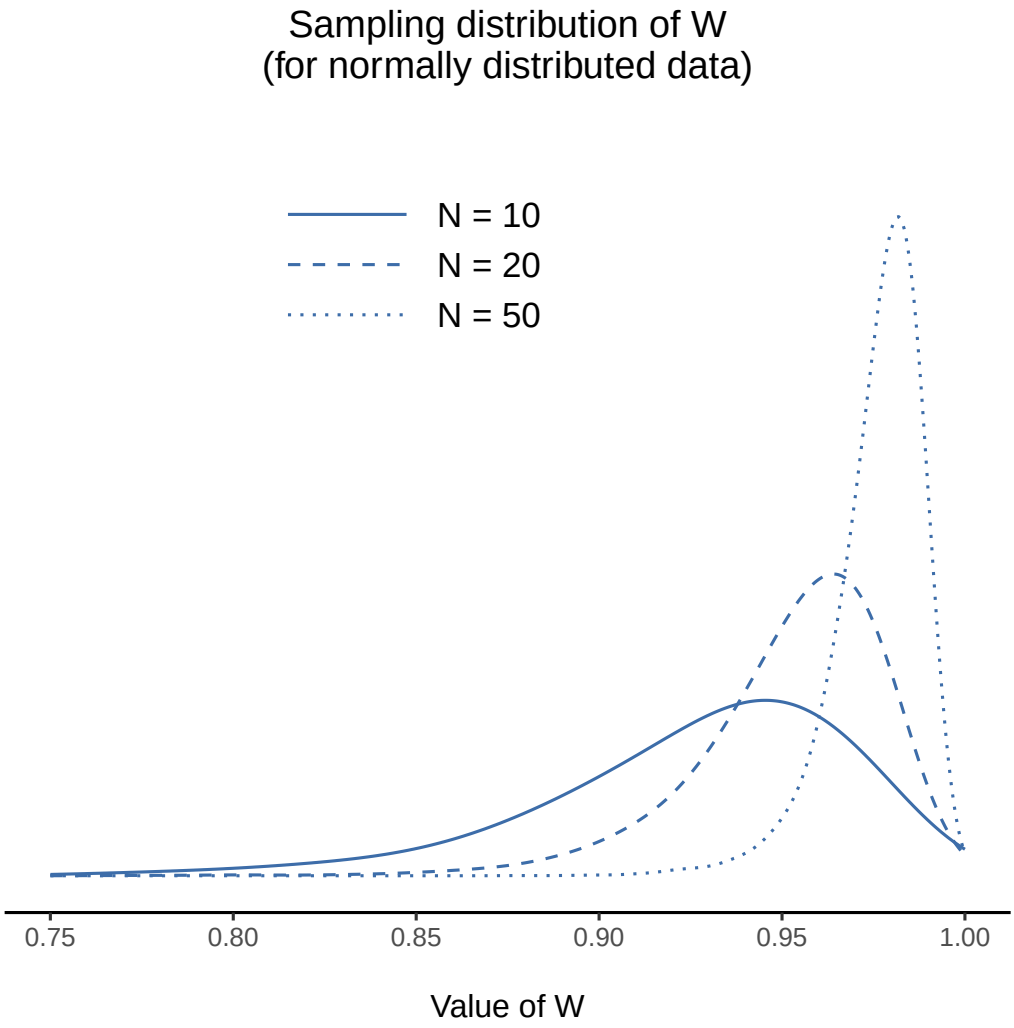


Figure 11.24: Distribución muestral del estadístico W de Shapiro-Wilk, bajo la hipótesis nula de que los datos se distribuyen normalmente, para muestras de tamaño 10, 20 y 50. Nótese que valores pequeños de W indican desviación de la normalidad

de 'Normalidad' que se encuentra en 'Supuestos'. En los datos muestreados aleatoriamente ($N = 100$) que usamos para el gráfico QQ, el valor del estadístico de la prueba de normalidad de Shapiro-Wilk fue $W = 0,99$ con un valor p de $0,54$. Entonces, como es lógico, no tenemos evidencia de que estos datos se aparten de la normalidad. Al informar los resultados de una prueba de Shapiro-Wilk, debes (como de costumbre) asegurarte de incluir la prueba estadística W y el valor p, aunque dado que la distribución muestral depende tanto de N, probablemente estaría bien incluir N también.

11.9.4 Ejemplo

Mientras tanto, probablemente valga la pena mostrarte un ejemplo de lo que sucede con el gráfico QQ y la prueba de Shapiro-Wilk cuando los datos no son normales. Para eso, veamos la distribución de nuestros datos de márgenes ganadores de la AFL, que si recuerdas Chapter 4, no parecían provenir de una distribución normal en absoluto. Esto es lo que sucede con el gráfico QQ (Figure 11.25).

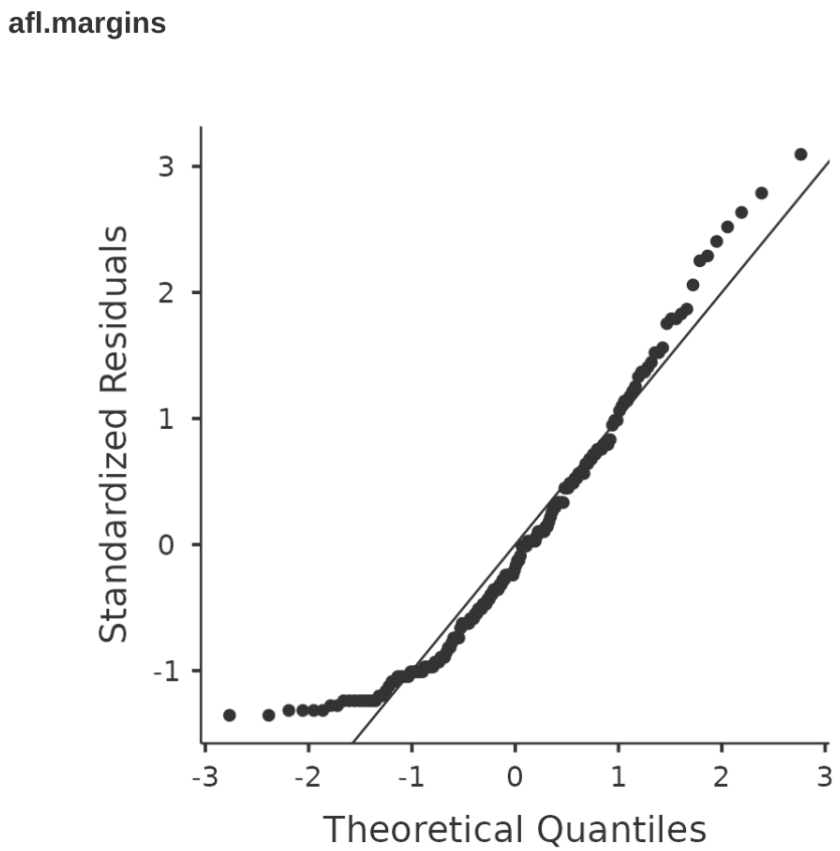


Figure 11.25: gráfico QQ que muestra la no normalidad de los datos de márgenes ganadores de la AFL

Y cuando ejecutamos la prueba de Shapiro-Wilk en los datos de márgenes de AFL, obtenemos un valor para la prueba estadística de normalidad de Shapiro-Wilk de $W = 0.94$ y valor $p = 9.481 \text{x} 10^{-07}$. ¡Claramente un efecto significativo!

11.10 Comprobación de datos no normales

Bien, supongamos que los datos resultan ser sustancialmente no normales, pero aún así queremos ejecutar algo como una prueba t. Esta situación se da a menudo en la vida real. Para los datos de los márgenes ganadores de la AFL, por ejemplo, la prueba de Shapiro-Wilk dejó muy claro que se viola el supuesto de normalidad. Esta es la situación en la que conviene utilizar las pruebas de Wilcoxon.

Al igual que la prueba t, la prueba de Wilcoxon viene en dos formas, de una muestra y de dos muestras, y se utilizan más o menos en las mismas situaciones que las pruebas t correspondientes. A diferencia de la prueba t, la prueba de Wilcoxon no asume normalidad, lo cual es bueno. De hecho, no hace ninguna suposición sobre qué tipo de distribución está involucrada. En la jerga estadística, esto lo convierte en **pruebas no paramétricas**. Aunque evitar el supuesto de normalidad es bueno, existe un inconveniente: la prueba de Wilcoxon suele ser menos potente que la prueba t (es decir, una mayor tasa de error de tipo II). No discutiré las pruebas de Wilcoxon con tanto detalle como las pruebas t, pero os daré una breve descripción general.

11.10.1 Prueba U de Mann-Whitney de dos muestras

Comenzaré describiendo la **prueba U de Mann-Whitney**, ya que en realidad es más simple que la versión de una muestra. Supongamos que estamos viendo las puntuaciones de 10 personas en algún examen. Como mi imaginación ahora me ha fallado por completo, supongamos que es una "prueba de asombro" y que hay dos grupos de personas, "A" y "B". Tengo curiosidad por saber qué grupo es más impresionante. Los datos están incluidos en el archivo awesome.csv, y hay dos variables además de la variable ID habitual: puntuaciones y grupo.

Mientras no haya vínculos (es decir, personas con exactamente la misma puntuación de genialidad), la prueba que queremos hacer es muy simple. Todo lo que tenemos que hacer es construir una tabla que compare cada observación del grupo A con cada observación del grupo B. Siempre que el dato del grupo A sea más grande, colocamos una marca de verificación en la tabla (Table 11.4).

Table 11.4: Comparación de observaciones por grupo para una prueba U de Mann-Whitney de dos muestras

		group B				
		14.5	10.4	12.4	11.7	13.0
group A	6.4	.	.	.	.	.
	10.7	.	✓	.	.	.
	11.9	.	✓	.	✓	.
	7.3	.	.	.	.	.
	10	.	.	.	.	.

Luego contamos el número de marcas de verificación. Esta es nuestra prueba estadística, W. [24] La distribución muestral real para W es algo complicada y me saltearé los detalles. Para nuestros propósitos, es suficiente notar que la interpretación de W es cualitativamente la misma que la interpretación de t o z. Es decir, si queremos una prueba bilateral, rechazamos la hipótesis nula cuando W es muy grande o muy pequeño, pero si tenemos una hipótesis unidireccional (es decir, unilateral), entonces solo usamos una u otra.

En jamovi, si ejecutamos una 'Prueba T de muestras independientes' con puntuaciones como variable dependiente y grupo como la variable de agrupación, y luego en las opciones de 'pruebas' marcas la opción de U Mann-Whitney, obtendremos resultados que muestran que $U = 3$ (es decir, el mismo número de marcas de verificación que se muestra arriba), y un valor p = 0.05556. Ver Figure 11.26.

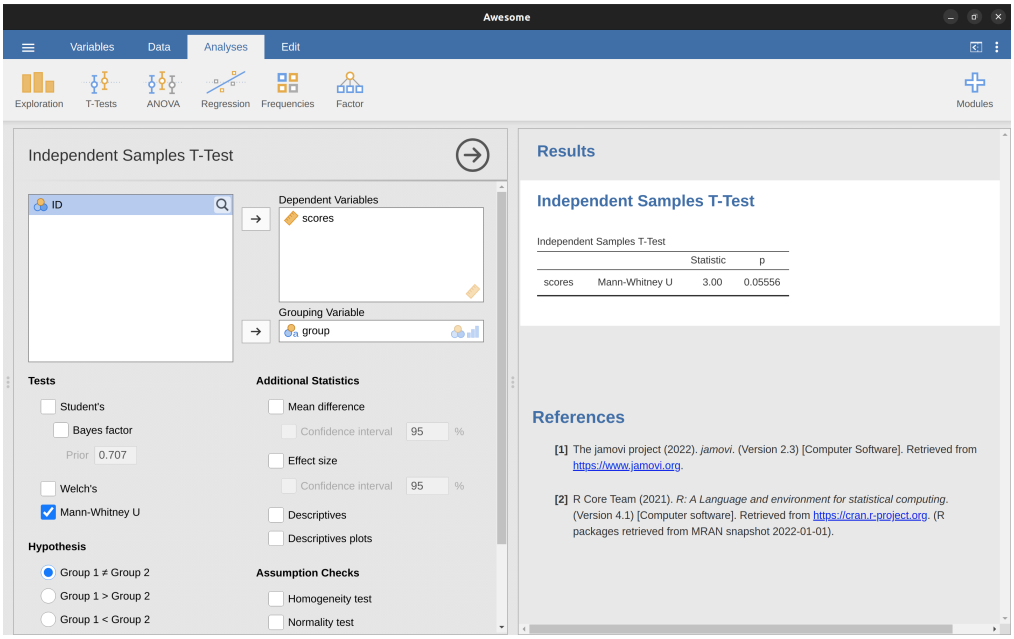


Figure 11.26: pantalla jamovi que muestra los resultados de la prueba U Mann-Whitney

11.10.2 Prueba de Wilcoxon de una muestra

¿Qué pasa con la **prueba de Wilcoxon de una muestra** (o de manera equivalente, la prueba de Wilcoxon de muestras pareadas)? Supongamos que estoy interesada en saber si recibir una clase de estadística tiene algún efecto sobre la felicidad de los estudiantes. Mis datos están en el archivo luck.csv. Lo que he medido aquí es la felicidad de cada estudiante antes de recibir la clase y después de recibir la clase, y la puntuación de cambio es la diferencia entre los dos. Tal como vimos con la prueba t, no hay una diferencia fundamental entre hacer una prueba de muestras pareadas usando antes y después, versus hacer una prueba de una muestra usando las puntuaciones de cambio.

[24] en realidad, hay dos versiones diferentes de la prueba estadística que difieren entre sí por un valor constante. La versión que he descrito es la que calcula jamovi.

Como antes, la forma más sencilla de pensar en la prueba es construir una tabulación. La forma de hacerlo esta vez es tomar esas puntuaciones de cambio que son diferencias positivas y tabularlas con respecto a toda la muestra completa. Lo que termina es una tabla que se parece a Table 11.5.

Table 11.5: Comparación de observaciones por grupo para una prueba U de Wilcoxon de una muestra

		all differences									
positive differences		−24	−14	−10	7	−6	−38	2	−35	−30	5
	7	.	.	.	✓	✓	.	✓	.	.	✓
	2	.	.	.	.	.	.	✓	.	.	.
	5	.	.	.	.	.	.	✓	.	.	✓

Contando las marcas de verificación esta vez obtenemos una una prueba estadística de $W = 7$. Como antes, si nuestra prueba es bilateral, rechazamos la hipótesis nula cuando W es muy grande o muy pequeña. En cuanto a ejecutarlo en jamovi, es más o menos lo que cabría esperar. Para la versión de una muestra, especifica la opción 'Clasificación de Wilcoxon' en 'Pruebas' en la ventana de análisis 'Prueba T de una muestra'. Esto te da Wilcoxon $W = 7$, valor p = 0.03711. Como esto demuestra, tenemos un efecto significativo. Evidentemente, tomar una clase de estadística tiene un efecto en tu felicidad. Cambiar a una versión de la prueba con muestras emparejadas no nos dará una respuesta diferente, por supuesto; ver Figure 11.27.

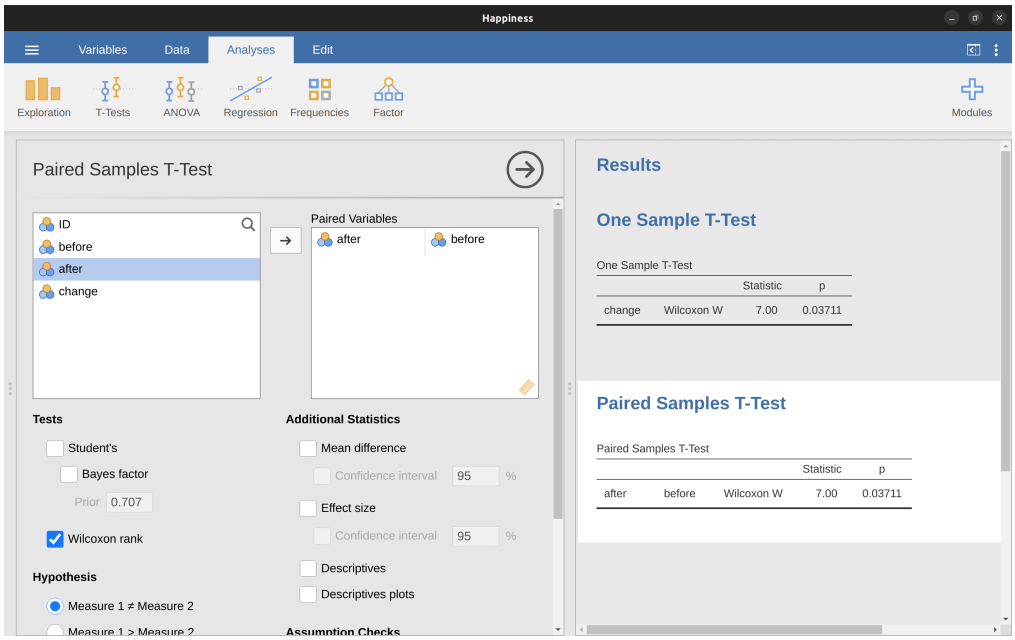


Figure 11.27: pantalla jamovi que muestra los resultados de las pruebas no paramétricas de Wilcoxon de una muestra y muestras emparejadas

11.11 Resumen

- La prueba t de una muestra se utiliza para comparar la media de una sola muestra con un valor hipotético para la media poblacional.
- Se utiliza una prueba t de muestras independientes para comparar las medias de dos grupos y prueba la hipótesis nula de que tienen la misma media. Viene en dos formas: La prueba t de muestras independientes (prueba de Student) (#sec-the-independent-samples-t-test-student-test) asume que los grupos tienen la misma desviación estándar, Las muestras independientes prueba t (prueba de Welch) no lo hace.
- [La prueba t de muestras relacionadas] se usa cuando tienes dos puntuaciones de cada persona y deseas probar la hipótesis nula de que las dos puntuaciones tienen la misma media. Es equivalente a tomar la diferencia entre las dos puntuaciones para cada persona y luego ejecutar una prueba t de una muestra en las puntuaciones de diferencia.
- [Las pruebas unilaterales] son perfectamente legítimas siempre que estén planificadas previamente (¡como todas las pruebas!).
- Los cálculos de Tamaño del efecto para la diferencia entre las medias se pueden calcular a través del estadístico d de Cohen.
- [Comprobación de la normalidad de una muestra] mediante gráficos QQ y la prueba de Shapiro-Wilk.
- Si tus datos no son normales, puedes usar las pruebas de Mann-Whitney o Wilcoxon en lugar de las pruebas t para [Prueba de datos no normales].

Chapter 12

Correlación y regresión lineal

El objetivo de este capítulo es presentar la **correlación** y la **regresión lineal**. Estas son las herramientas estándar en las que confían los estadísticos para analizar la relación entre factores predictores continuos y resultados continuos.

12.1 Correlaciones

En esta sección hablaremos sobre cómo describir las relaciones entre variables en los datos. Para ello, hablaremos principalmente de la **correlación** entre variables. Pero primero, necesitamos algunos datos (Table 12.1).

12.1.1 Los datos

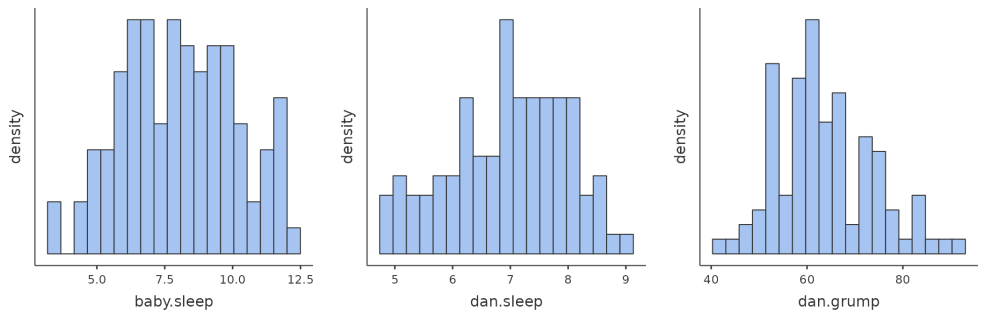
Table 12.1: datos para el análisis de correlación:- estadísticos descriptivos para los datos de paternidad

variable	min	max	mean	median	std. dev	IQR
Dani's grumpiness	41	91	63.71	62	10.05	14
Dani's hours slept	4.84	9.00	6.97	7.03	1.02	1.45
Dani's son's hours slept	3.25	12.07	8.05	7.95	2.07	3.21

Pasemos a un tema cercano al corazón de todos los padres: el sueño. El conjunto de datos que usaremos es ficticio, pero está basado en hechos reales. Supongamos que tengo curiosidad por saber cuánto afectan los hábitos de sueño de mi hijo pequeño a mi estado de ánimo. Digamos que puedo calificar mi mal humor con mucha precisión,

en una escala de 0 (nada malhumorado) a 100 (malhumorado como un anciano o una anciana muy, muy gruñona). Y supongamos también que he estado midiendo mi mal humor, mis patrones de sueño y los patrones de sueño de mi hijo desde hace bastante tiempo. Digamos, durante 100 días. Y, siendo un nerd, guardé los datos en un archivo llamado parenthood.csv. Si cargamos los datos podemos ver que el archivo contiene cuatro variables dani.sleep, baby.sleep, dani.grump y day. Ten en cuenta que cuando cargues este conjunto de datos por primera vez, es posible que jamovi no haya adivinado correctamente el tipo de datos para cada variable, en cuyo caso debes corregirlo: dani.sleep, baby.sleep, dani.grump y day pueden especificarse como variables continuas, e ID es una variable nominal (entera).[1]

A continuación, echaré un vistazo a algunos estadísticos descriptivos básicos y, para dar una descripción gráfica de cómo son cada una de las tres variables interesantes, Figure 12.1 presenta histogramas. Una cosa a tener en cuenta: el hecho de que jamovi pueda calcular docenas de estadísticos diferentes no significa que debas informarlos todos. Si estuviera escribiendo esto para un informe, probablemente elegiría los estadísticos que son de mayor interés para mí (y para mis lectores) y luego los colocaría en una tabla agradable y simple como la de la Tabla 12.1.[2] Ten en cuenta que cuando lo puse en una tabla, le di a todo nombres "legibles por humanos". Esta es siempre una buena práctica. Nota también que no estoy durmiendo lo suficiente. Esta no es una buena práctica, pero otros padres me dicen que es bastante estándar.



(a) (b) (c)

Figure 12.1: Histogramas de jamovi para las tres variables interesantes en el conjunto de datos de paternidad

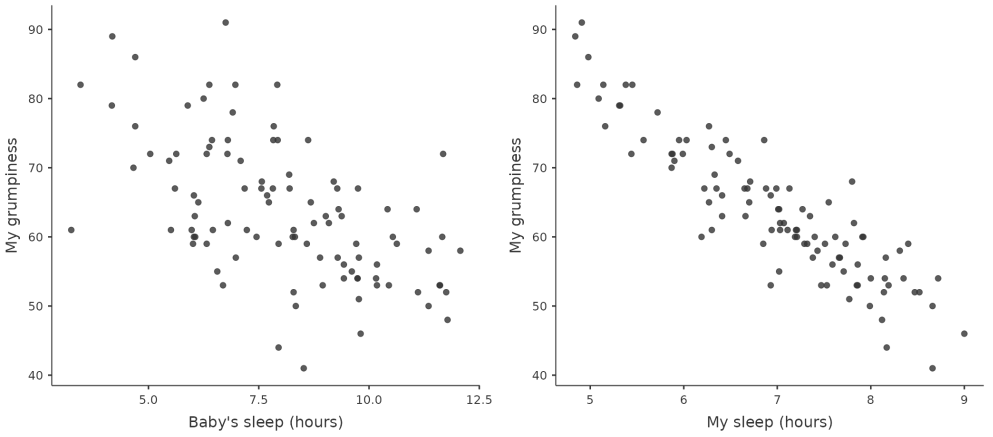
12.1.2 La fuerza y la dirección de una relación

Podemos dibujar diagramas de dispersión para tener una idea general de cuán estrechamente relacionadas están dos variables. Sin embargo, idealmente, podríamos querer decir un poco más al respecto. Por ejemplo, comparemos la relación entre baby.sleep y dani.grump (Figure 12.1a), izquierda, con la de dani.sleep y dani.grump (Figure 12.1b), derecha. Al mirar estos dos gráficos uno al lado del otro, está claro que la relación es cualitativamente la misma en ambos casos: ¡más sueño equivale a menos

[1] he observado que en jamovi también puedes especificar un tipo de variable 'ID', pero para nuestros propósitos no importa cómo especifiquemos la variable ID dado que no lo incluiremos en ningún análisis.

[2] En realidad, incluso esa tabla es más de lo que me molestaría. En la práctica, la mayoría de las personas eligen una medida de tendencia central y una sola medida de variabilidad.

mal humor! Sin embargo, también es bastante obvio que la relación entre dani.sleep y dani.grump es más fuerte que la relación entre baby.sleep y dani.grump. El gráfico de la derecha es "más ordenado" que el de la izquierda. Lo que parece es que si quieres predecir cuál es mi estado de ánimo, te ayudaría un poco saber cuántas horas durmió mi hijo, pero sería más útil saber cuántas horas dormí yo.



(a) (b)

Figure 12.2: Gráficos de dispersión de jamovi que muestran la relación entre baby.sleep y dani.grump (izquierda) y la relación entre dani.sleep y dani.grump (derecha)

Por el contrario, consideremos los dos diagramas de dispersión que se muestran en Figure 12.3. Si comparamos el diagrama de dispersión de "baby.sleep v dani.grump" (izquierda) con el diagrama de dispersión de "'baby.sleep v dani.sleep" (derecha), la fuerza general de la relación es la misma, pero la dirección es diferente. Es decir, si mi hijo duerme más, yo duermo más (relación positiva, lado derecho), pero si él duerme más, yo me pongo menos gruñón (relación negativa, lado izquierdo).

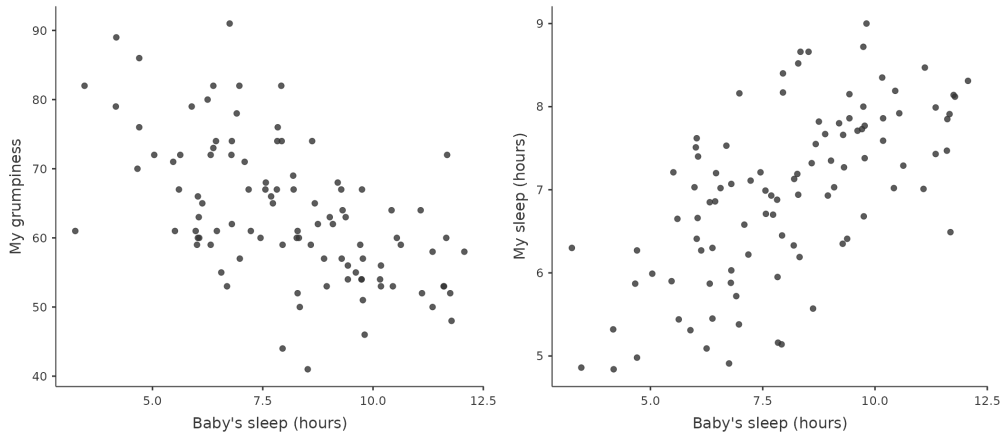
12.1.3 El coeficiente de correlación

Podemos hacer estas ideas un poco más explícitas introduciendo la idea de un **coeficiente de correlación** (o, más específicamente, el coeficiente de correlación de Pearson), que tradicionalmente se denota como r. El coeficiente de correlación entre dos variables X y Y (a veces denominado r_{XY}), que definiremos con más precisión en la siguiente sección, es una medida que varía de -1 a 1. Cuando $r = -1$ significa que tenemos una relación negativa perfecta, y cuando $r = 1$ significa que tenemos una relación positiva perfecta. Cuando $r = 0$, no hay ninguna relación. Si observas Figure 12.4, puedes ver varios gráficos que muestran cómo son las diferentes correlaciones.

[Detalle técnico adicional [3]]

Al estandarizar la covarianza, no solo mantenemos todas las buenas propiedades de la covarianza discutidas anteriormente, sino que los valores reales de r están en una escala

[3] la fórmula para el coeficiente de correlación de Pearson se puede escribir de varias maneras diferentes. Creo que la forma más sencilla de escribir la fórmula es dividirla en dos pasos. En primer lugar, introduzcamos la idea de una **covarianza**. La covarianza entre dos variables X y Y es una generalización de la noción de varianza y es una forma matemáticamente simple de describir la relación entre



(a) (b)

Figure 12.3: Diagramas de dispersión de jamovi que muestran la relación entre baby.sleep y dani.grump (izquierda), en comparación con la relación entre baby.sleep y dani.sleep (derecha)

significativa: r = 1 implica una relación positiva perfecta y $r = -1$ implica una relación negativa perfecta. Me extenderé un poco más sobre este punto más adelante, en la sección sobre [Interpretación de una correlación]. Pero antes de hacerlo, veamos cómo calcular correlaciones en jamovi.

12.1.4 Cálculo de correlaciones en jamovi

El cálculo de correlaciones en jamovi se puede hacer haciendo clic en el botón 'Regresión' - 'Matriz de correlación'. Transfiere las cuatro variables continuas al cuadro de la derecha

---

dos variables que no es muy informativa para los humanos

$$Cov(X, Y) = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

Porque estamos multiplicando (es decir, tomando el "producto" de) una cantidad que depende de X por una cantidad que depende de Y y luego promediando [a], puedes pensar en la fórmula para la covarianza como un "producto cruzado promedio" entre X y Y . La covarianza tiene la buena propiedad de que, si X y Y no están relacionados en absoluto, entonces la covarianza es exactamente cero. Si la relación entre ellos es positiva (en el sentido que se muestra en Figure 12.4, entonces la covarianza también es positiva, y si la relación es negativa, la covarianza también es negativa. En otras palabras, la covarianza captura la idea cualitativa básica de la correlación. Desafortunadamente, la magnitud bruta de la covarianza no es fácil de interpretar, ya que depende de las unidades en las que se expresan X y Y y, peor aún, las unidades reales en las que se expresa la covarianza misma son realmente raras. Por ejemplo, si X se refiere a la variable dani.sleep (unidades: horas) y Y se refiere a la variable dani.grump (unidades: grumps), entonces las unidades para su covarianza son $horas ×gruñones $.Y no tengo ni idea de lo que eso significa. El coeficiente de correlación de Pearson r soluciona este problema de interpretac
En otras palabras, la correlación entre X y Y se puede escribir de la siguiente manera:

$$r_{XY} = \frac{Cov(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

—[a] Tal como vimos con la varianza y la desviación estándar, en la práctica dividimos por $N - 1$ en lugar de N. [b] Esta es una simplificación excesiva, pero servirá para nuestros propósitos.

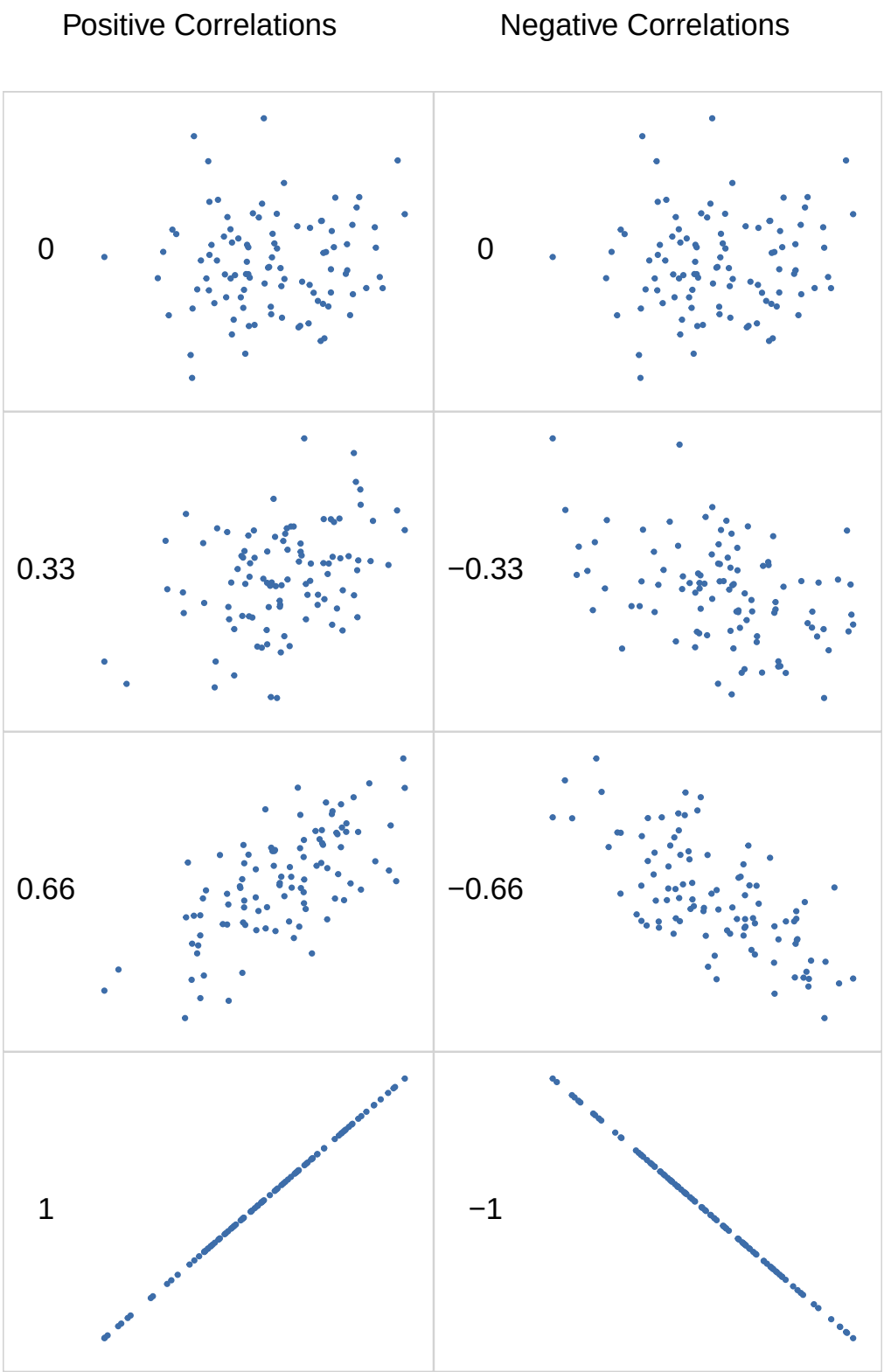


Figure 12.4: Ilustración del efecto de variar la fuerza y la dirección de una correlación. En la columna de la izquierda, las correlaciones son $0, .33, .66$ y 1. En la columna de la derecha, las correlaciones son $0, -.33, -.66$ y -1

para obtener el resultado en Figure 12.5.

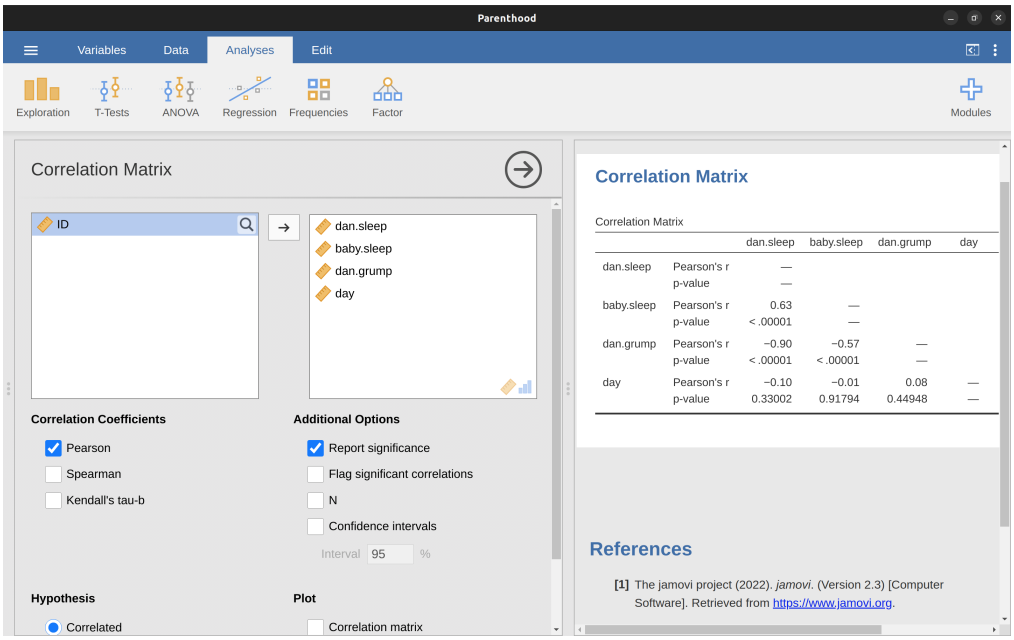


Figure 12.5: Una captura de pantalla jamovi que muestra las correlaciones entre las variables en el archivo parenthood.csv

12.1.5 Interpretar una correlación

Naturalmente, en la vida real no se ven muchas correlaciones de 1. Entonces, ¿cómo deberías interpretar una correlación de, digamos, r = .4? La verdad es que realmente depende de para qué deseas usar los datos y de cómo de fuertes tienden a ser las correlaciones en tu campo. Un amigo mío en ingeniería argumentó una vez que cualquier correlación menor a .95 es completamente inútil (creo que estaba exagerando, incluso para la ingeniería). Por otro lado, hay casos reales, incluso en psicología, en los que realmente deberías esperar correlaciones tan fuertes. Por ejemplo, uno de los conjuntos de datos de referencia utilizados para probar las teorías de cómo las personas juzgan las similitudes es tan limpio que cualquier teoría que no pueda lograr una correlación de al menos .9 realmente no se considera exitosa. Sin embargo, al buscar (digamos) correlatos elementales de inteligencia (por ejemplo, tiempo de inspección, tiempo de respuesta), si obtienes una correlación superior a .3, lo estás haciendo muy bien. En resumen, la interpretación de una correlación depende mucho del contexto. Dicho esto, la guía aproximada que se presenta en Table 12.2 es bastante típica.

Sin embargo, algo que nunca se puede enfatizar lo suficiente es que siempre debes mirar el diagrama de dispersión antes de adjuntar cualquier interpretación a los datos. Una correlación podría no significar lo que crees que significa. La ilustración clásica de esto es el "Cuarteto de Anscombe" (Anscombe, 1973), una colección de cuatro conjuntos de datos. Cada conjunto de datos tiene dos variables, X y Y. Para los cuatro conjuntos de datos, el valor medio para X es 9 y el valor medio para Y es $7,5$. Las desviaciones

Table 12.2: una guía aproximada* para interpretar las correlaciones

Correlation	Strength	Direction
-1.00 to -0.90	Very strong	Negative
-0.90 to -0.70	Strong	Negative
-0.70 to -0.40	Moderate	Negative
-0.40 to -0.20	Weak	Negative
-0.20 to 0.00	Negligible	Negative
0.00 to 0.20	Negligible	Positive
0.20 to 0.40	Weak	Positive
0.40 to 0.70	Moderate	Positive
0.70 to 0.90	Strong	Positive
0.90 to 1.00	Very strong	Positive

**Note that I say a rough guide. There aren't hard and fast rules for what counts as strong or weak relationships. It depends on the context*

estándar de todas las variables X son casi idénticas, al igual que las de las variables Y. Y en cada caso la correlación entre X y Y es $r = 0.816$. Puedes verificar esto tú misma, ya que lo guardé en un archivo llamado anscombe.csv.

Una pensaría que estos cuatro conjuntos de datos serían bastante similares entre sí. Pero no lo son. Si dibujamos diagramas de dispersión de X contra Y para las cuatro variables, como se muestra en Figure 12.6, vemos que los cuatro son espectacularmente diferentes entre sí. La lección aquí, que mucha gente parece olvidar en la vida real, es "siempre representar gráficamente tus datos sin procesar" (ver Chapter 5).

12.1.6 Correlaciones de rango de Spearman

El coeficiente de correlación de Pearson es útil para muchas cosas, pero tiene deficiencias. Destaca una cuestión en particular: lo que realmente mide es la fuerza de la relación lineal entre dos variables. En otras palabras, lo que le da es una medida de la medida en que todos los datos tienden a caer en una sola línea perfectamente recta. A menudo, esta es una aproximación bastante buena a lo que queremos decir cuando decimos "relación", por lo que es bueno calcular la correlación de Pearson. A veces, sin embargo, no lo es.

Una situación muy común en la que la correlación de Pearson no es lo correcto surge cuando un aumento en una variable X realmente se refleja en un aumento en otra variable Y , pero la naturaleza de la relación no es necesariamente lineal. Un ejemplo de esto podría ser la relación entre el esfuerzo y la recompensa al estudiar para un examen. Si no te esfuerzas (X) en aprender una materia, deberías esperar una calificación de 0% (Y). Sin embargo, un poco de esfuerzo causará una mejora masiva. El solo hecho de asistir a las clases significa que aprendes bastante, y si solo llegas a las clases y garabateas algunas cosas, tu calificación podría subir al 35%, todo sin mucho esfuerzo. Sin embargo, no obtienes el mismo efecto en el otro extremo de la escala. Como todo el mundo sabe, se necesita mucho más esfuerzo para obtener una calificación de 90% que

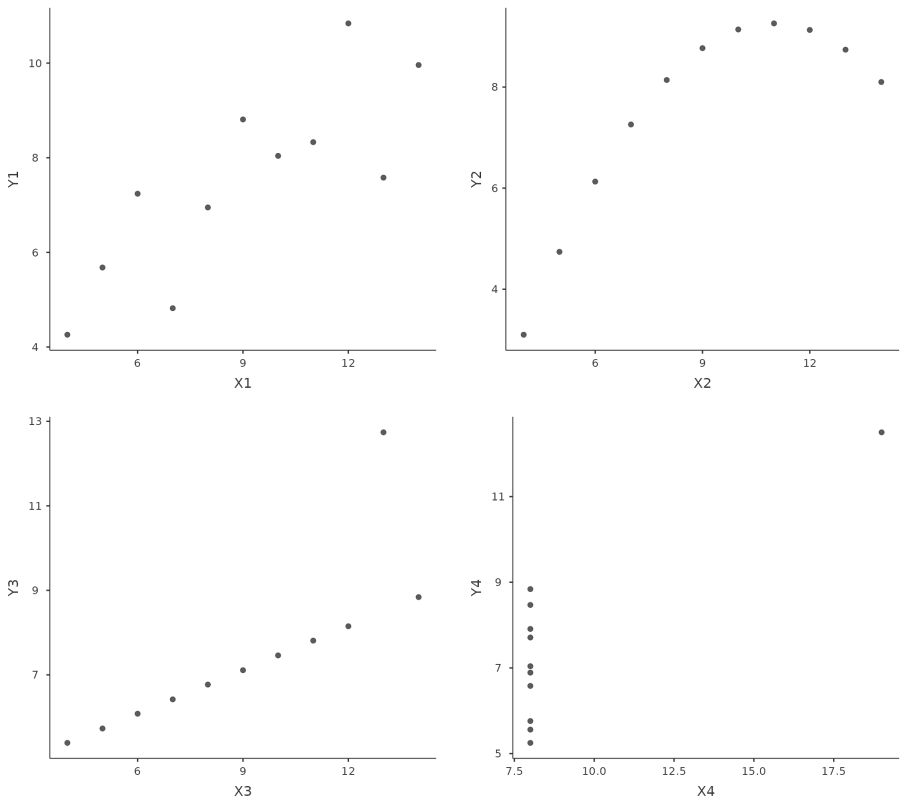


Figure 12.6: Diagramas de dispersión del cuarteto de Anscombe en jamovi. Los cuatro conjuntos de datos tienen una correlación de Pearson de r = .816, pero son cualitativamente diferentes entre sí.

para obtener una calificación de 55%. Lo que esto significa es que, si tengo datos que analizan el esfuerzo de estudio y las calificaciones, hay muchas posibilidades de que las correlaciones de Pearson sean engañosas.

Para ilustrar, considera los datos representados en Figure 12.7, que muestran la relación entre las horas trabajadas y la calificación recibida por 10 estudiantes que reciben alguna clase. Lo curioso de este conjunto de datos (muy ficticios) es que aumentar tu esfuerzo siempre aumenta tu nota. Puede ser por mucho o por poco, pero aumentar el esfuerzo nunca disminuirá tu calificación. Si ejecutamos una correlación estándar de Pearson, muestra una fuerte relación entre las horas trabajadas y la calificación recibida, con un coeficiente de correlación de 0.91. Sin embargo, esto en realidad no refleja el hecho de que aumentar las horas trabajadas siempre aumenta la calificación. Aquí queremos poder decir que la correlación es perfecta pero para una noción algo diferente de lo que es una "relación". Lo que estamos buscando es algo que capte el hecho de que aquí hay una **relación ordinal** perfecta. Es decir, si el estudiante 1 trabaja más horas que el estudiante 2, entonces podemos garantizar que el estudiante 1 obtendrá la mejor calificación. Eso no es lo que dice una correlación de $r = .91$.

¿Cómo debemos abordar esto? En realidad, es muy fácil. Si estamos buscando relaciones ordinales, todo lo que tenemos que hacer es tratar los datos como si fueran una escala ordinal. Así, en lugar de medir el esfuerzo en términos de "horas trabajadas", clasifiquemos nuestros 10 estudiantes en orden de horas trabajadas. Es decir, el estudiante 1 hizo la menor cantidad de trabajo de todos (2 horas), por lo que obtuvo el rango más bajo (rango = 1). El estudiante 4 fue el siguiente más perezoso, dedicando solo 6 horas de trabajo durante todo el semestre, por lo que obtiene el siguiente rango más bajo (rango = 2). Ten en cuenta que estoy usando "rango = 1" para hacer referencia a "rango bajo". A veces, en el lenguaje cotidiano, hablamos de "rango = $ 1 $" para hacer referencia a "rango superior" en lugar de "rango inferior". Así que ten cuidado, puedes clasificar "del valor más pequeño al valor más grande" (es decir, pequeño equivale a rango 1) o puedes clasificar "del valor más grande al valor más pequeño" (es decir, grande equivale a rango 1). En este caso, estoy clasificando de menor a mayor, pero como es muy fácil olvidar de qué manera configuraste las cosas, ¡tienes que esforzarte un poco para recordar!

Bien, echemos un vistazo a nuestros estudiantes cuando los clasificamos de peor a mejor en términos de esfuerzo y recompensa Table 12.3.

Mmm. Estos son idénticos. El estudiante que se esforzó más obtuvo la mejor calificación, el estudiante que se esforzó menos obtuvo la peor calificación, etc. Como muestra la tabla anterior, estas dos clasificaciones son idénticas, por lo que si ahora las correlacionamos obtenemos una relación perfecta, con una correlación de 1.0.

Lo que acabamos de reinventar es la **correlación de orden de rango de Spearman**, generalmente denominada ρ para distinguirla de la correlación r de Pearson. Podemos calcular el ρ de Spearman usando jamovi simplemente haciendo clic en la casilla de verificación 'Spearman' en la pantalla 'Correlation Matrix'.

12.2 Gráfico de dispersión

Los diagramas de dispersión son una herramienta simple pero efectiva para visualizar la relación entre dos variables, como vimos con las figuras en la sección sobre

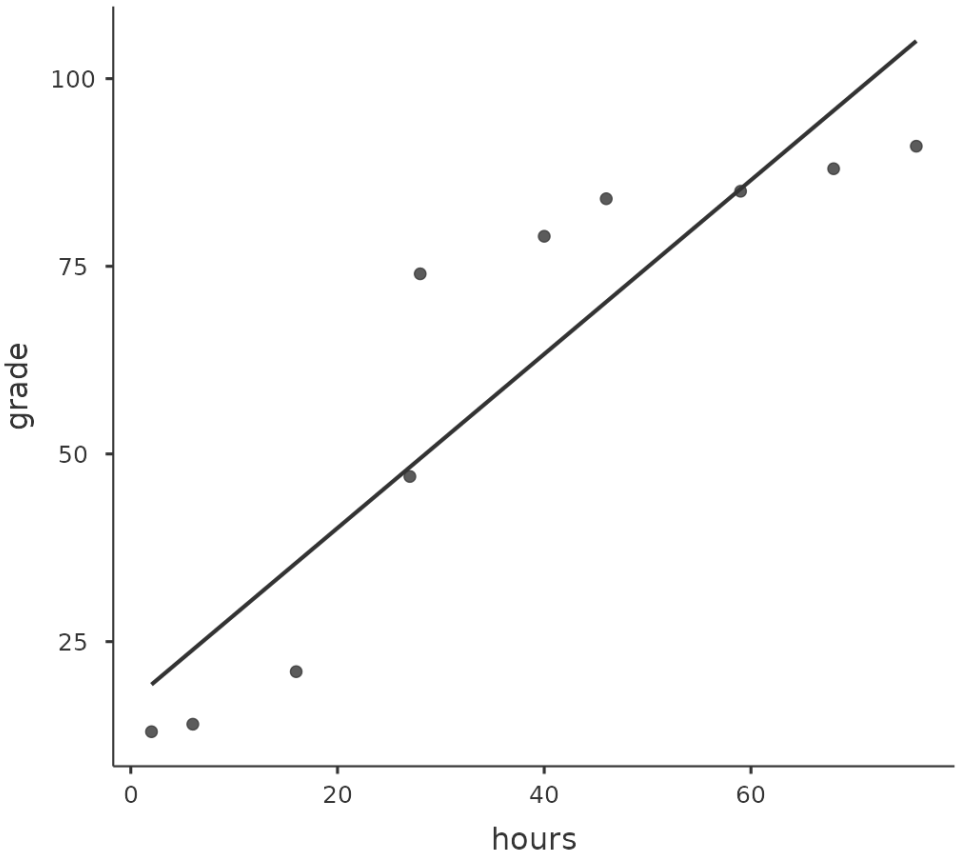


Figure 12.7: La relación entre las horas trabajadas y la calificación recibida para un juego de datos que consta de solo 10 estudiantes (cada punto corresponde a un estudiante). La línea que pasa por el medio muestra la relación lineal entre las dos variables. Esto produce una fuerte correlación de Pearson de $r = .91$. Sin embargo, lo interesante a tener en cuenta aquí es que en realidad existe una relación monótona perfecta entre las dos variables. En este juego de datos, aumentar las horas trabajadas siempre aumenta la calificación recibida, como lo ilustra la línea continua. Esto se refleja en una correlación de Spearman de $\rho = 1$. Sin embargo, con un conjunto de datos tan pequeño, la pregunta es qué versión describe mejor la relación real involucrada.

Table 12.3: Estudiantes clasificados en términos de esfuerzo y recompensa

	rank (hours worked)	**rank (grade received)**
student 1	1	1
student 2	10	10
student 3	6	6
student 4	2	2
student 5	3	3
student 6	5	5
student 7	4	4
student 8	8	8
student 9	7	7
student 10	9	9

Correlaciones. Es esta última aplicación la que generalmente tenemos en mente cuando usamos el término "diagrama de dispersión". En este tipo de gráfico, cada observación corresponde a un punto. La ubicación horizontal del punto traza el valor de la observación en una variable y la ubicación vertical muestra su valor en la otra variable. En muchas situaciones, realmente no tienes una opinión clara sobre cuál es la relación causal (por ejemplo, A causa B, o B causa A, o alguna otra variable C controla tanto A como B). Si ese es el caso, realmente no importa qué variable representas gráficamente en el eje x y cuál representas en el eje y. Sin embargo, en muchas situaciones tienes una idea bastante clara de qué variable crees que es más probable que sea causal, o al menos tienes algunas sospechas que van en esa dirección. Si es así, entonces es usual representar la variable de causa en el eje x y la variable de efecto en el eje y. Con eso en mente, veamos cómo dibujar diagramas de dispersión en jamovi, usando el mismo conjunto de datos de paternidad (es decir, parenthood.csv) que usé cuando introduje las correlaciones.

Supongamos que mi objetivo es dibujar un diagrama de dispersión que muestre la relación entre la cantidad de sueño que duermo (dani.sleep) y lo malhumorada que estoy al día siguiente (dani.grump). Podemos obtener el gráfico que buscamos de dos maneras diferentes en jamovi. La primera forma es usar la opción 'Gráfica' debajo del botón 'Regresión' - 'Matriz de correlación', dándonos el resultado que se muestra en Figure 12.8. Ten en cuenta que jamovi dibuja una línea a través de los puntos, hablaremos de esto un poco más adelante en la sección sobre ¿Qué es un modelo de regresión lineal?. Trazar un diagrama de dispersión de esta manera también te permite especificar 'Densidades para variables' y esta opción agrega una curva de densidad que muestra cómo se distribuyen los datos en cada variable.

La segunda forma de hacerlo es usar uno de los módulos complementarios de jamovi. Este módulo se llama 'scatr' y puedes instalarlo haciendo clic en el icono grande '+' en la parte superior derecha de la pantalla de jamovi, abriendo la librería de jamovi, desplazándote hacia abajo hasta encontrar 'scatr' y haciendo clic en 'instalar'. Cuando hayas hecho esto, encontrarás un nuevo comando 'Gráfico de dispersión' disponible en el botón 'Exploración'. Este gráfico es un poco diferente al primero, ver Figure 12.9,

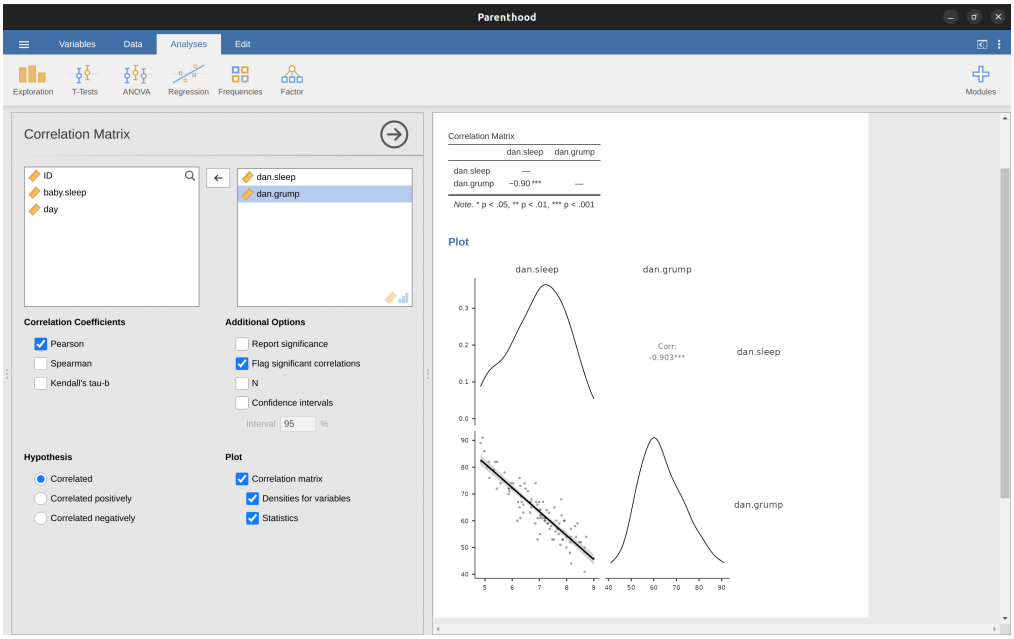


Figure 12.8: Diagrama de dispersión a través del comando 'Matriz de correlación' en jamovi

pero la información relevante es la misma.

12.2.1 Opciones más elaboradas

A menudo querrás ver las relaciones entre varias variables a la vez, usando una **matriz de diagrama de dispersión** (en jamovi a través del comando 'Matriz de correlación' - 'Gráfica'). Simplemente agrega otra variable, por ejemplo, baby.sleep a la lista de variables a correlacionar, y jamovi creará una matriz de diagrama de dispersión para ti, como la de Figure 12.10.

12.3 ¿Qué es un modelo de regresión lineal?

Reducidos a lo esencial, los modelos de regresión lineal son básicamente una versión un poco más elegante de la correlación de Pearson (consulta Correlaciones), aunque, como veremos, los modelos de regresión son herramientas mucho más poderosas.

Dado que las ideas básicas de la regresión están estrechamente relacionadas con la correlación, volveremos al archivo parenthood.csv que estábamos usando para ilustrar cómo funcionan las correlaciones. Recuerda que, en este conjunto de datos, estábamos tratando de averiguar por qué Dani está tan malhumorada todo el tiempo y nuestra hipótesis de trabajo era que no estoy durmiendo lo suficiente. Dibujamos algunos diagramas de dispersión para ayudarnos a examinar la relación entre la cantidad de sueño que duermo y mi mal humor al día siguiente, como en Figure 12.9, y como vimos anteriormente, esto corresponde a una correlación de $r = -.90 $, pero nos encontramos

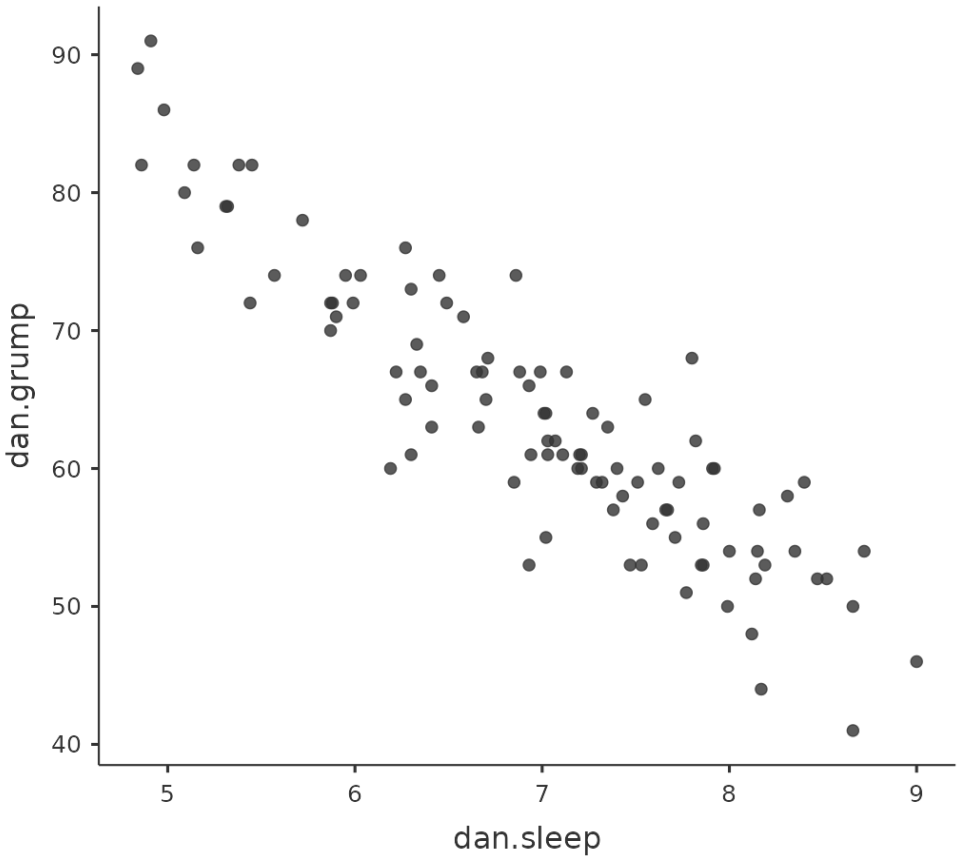


Figure 12.9: Diagrama de dispersión a través del módulo adicional 'scatr' en - jamovi

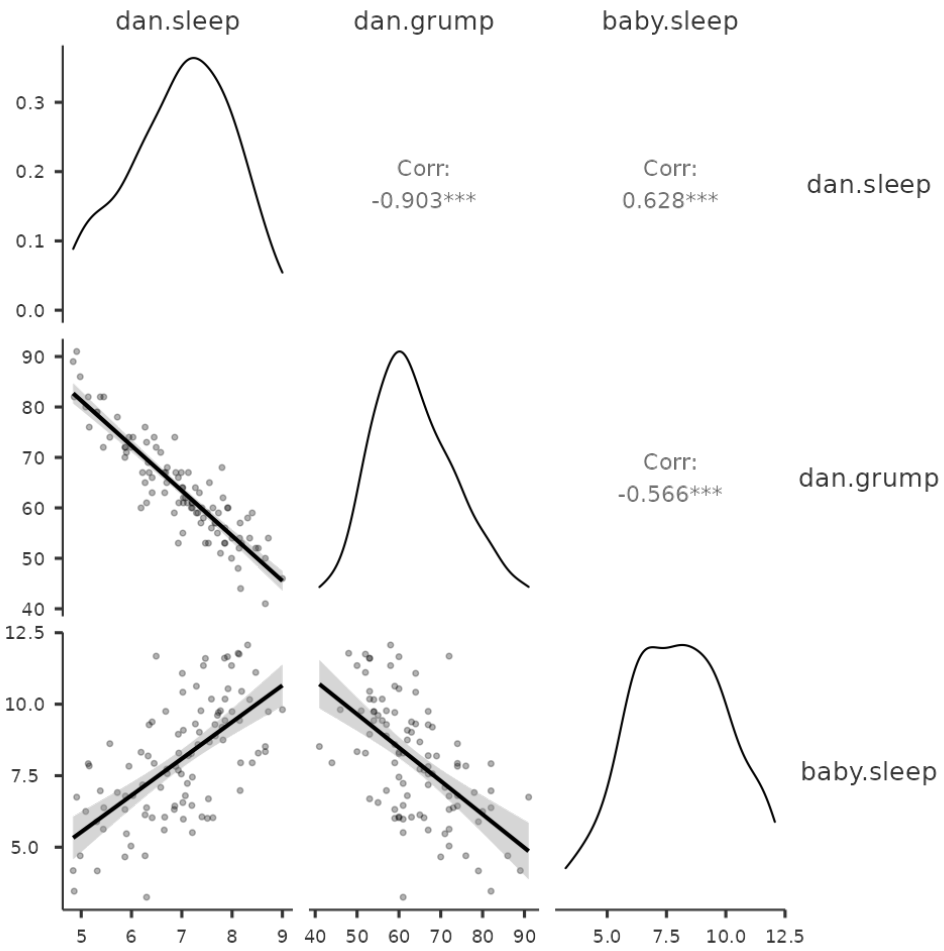


Figure 12.10: una matriz de diagramas de dispersión producidos usando jamovi

imaginando en secreto algo que se parece más a Figure 12.11 (a). Es decir, dibujamos mentalmente una línea recta a través de la mitad de los datos. En estadística, esta línea que estamos dibujando se llama **línea de regresión**. Ten en cuenta que la línea de regresión pasa por la mitad de los datos. No nos imaginamos nada parecido al gráfico que se muestra en Figure 12.11 (b).

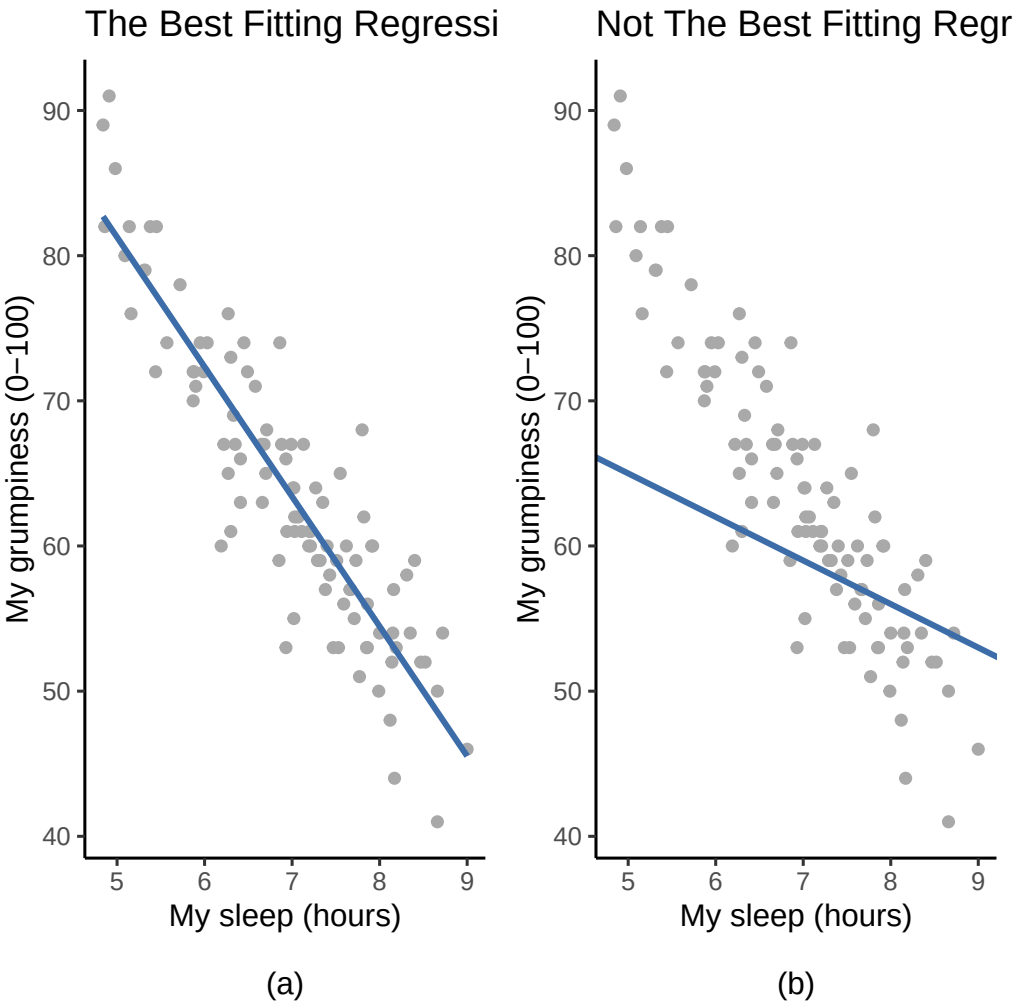


Figure 12.11: El panel (a) muestra el diagrama de dispersión de sueño-mal humor de Figure 12.9 con la línea de regresión de mejor ajuste dibujada en la parte superior. No es sorprendente que la línea pase por la mitad de los datos. Por el contrario, el panel (b) muestra los mismos datos, pero con una elección muy pobre de la línea de regresión dibujada en la parte superior.

Esto no es muy sorprendente. La línea que he dibujado en Figure 12.11 (b) no "encaja" muy bien con los datos, por lo que no tiene mucho sentido proponerla como una forma de resumir los datos, ¿verdad? Esta es una observación muy simple, pero resulta ser muy poderosa cuando empezamos a tratar de envolverla con un poco de matemática. Para hacerlo, comencemos con un repaso de algunas matemáticas de la escuela secundaria.

La fórmula de una línea recta generalmente se escribe así

$$y = a + bx$$

O, al menos, así era cuando fui a la escuela secundaria hace tantos años. Las dos variables son x y y, y tenemos dos coeficientes, a y b.[4] El coeficiente a representa la intersección de y de la recta, y el coeficiente b representa la pendiente de la recta. Profundizando más en nuestros recuerdos decadentes de la escuela secundaria (lo siento, para algunas de nosotras la escuela secundaria fue hace mucho tiempo), recordamos que la intersección se interpreta como "el valor de y que obtienes cuando $x = 0$". De manera similar, una pendiente de b significa que si aumentas el valor de x en 1 unidad, entonces el valor de y sube b unidades, y una pendiente negativa significa que el valor de y bajaría en lugar de subir. Ah, sí, ahora me acuerdo de todo. Ahora que lo hemos recordado no debería sorprendernos descubrir que usamos exactamente la misma fórmula para una recta de regresión. Si Y es la variable de resultado (la VD) y X es la variable predictora (la VI), entonces la fórmula que describe nuestra regresión se escribe así

$$\hat{Y}_i = b_0 + b_1 X_i$$

Mmm. Parece la misma fórmula, pero hay algunas partes extra en esta versión. Asegurémonos de entenderlos. En primer lugar, fíjate que he escrito X_i y Y_i en lugar de simplemente X y Y. Esto se debe a que queremos recordar que estamos tratando con datos reales. En esta ecuación, X_i es el valor de la variable predictora para la i-ésima observación (es decir, la cantidad de horas de sueño que dormí el día i de mi pequeño estudio), y Y_i es el valor correspondiente de la variable de resultado (es decir, mi mal humor ese día). Y aunque no lo he dicho explícitamente en la ecuación, lo que estamos asumiendo es que esta fórmula funciona para todas las observaciones en el conjunto de datos (es decir, para todo i). En segundo lugar, observa que escribí $\hat{Y}_i$ y no Y_i . Esto se debe a que queremos hacer la distinción entre los datos reales Y_i y la estimación $\hat{Y}_i$ (es decir, la predicción que hace nuestra recta de regresión). En tercer lugar, cambié las letras utilizadas para describir los coeficientes de a y b a b_0 y b_1. Así es como a los estadísticos les gusta referirse a los coeficientes en un modelo de regresión. No tengo ni idea de por qué eligieron b, pero eso es lo que hicieron. En cualquier caso, b_0 siempre se refiere al término de intersección y b_1 se refiere a la pendiente.

Excelente, excelente. A continuación, no puedo dejar de notar que, independientemente de si estamos hablando de la recta de regresión buena o mala, los datos no caen perfectamente en la recta. O, dicho de otra forma, los datos Y_i no son idénticos a las predicciones del modelo de regresión $\hat{Y}_i$. Dado que a los estadísticos les encanta adjuntar letras, nombres y números a todo, nos referiremos a la diferencia entre la predicción del modelo y ese punto de datos real como un valor residual, y lo llamaremos ϵ_i.[5] En términos matemáticos, los residuales se definen como

$$\epsilon_i = Y_i - \hat{Y}_i$$

[4] también se escribe a veces como $y = mx + c$ donde m es el coeficiente de pendiente y c es el coeficiente de intersección (constante).

[5] El símbolo ϵ es la letra griega epsilon. Es tradicional usar ϵ_i o e_i para indicar un residual.

lo que a su vez significa que podemos escribir el modelo de regresión lineal completo como

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

12.4 Estimación de un modelo de regresión lineal

Bien, ahora volvamos a dibujar nuestras imágenes, pero esta vez agregaré algunas líneas para mostrar el tamaño del residual para todas las observaciones. Cuando la recta de regresión es buena, nuestros residuales (las longitudes de las líneas negras continuas) son bastante pequeñas, como se muestra en Figure 12.12 (a), pero cuando la recta de regresión es mala, los residuales son mucho más grandes, como puedes ver en Figure 12.12 (b). Mmm. Tal vez lo que "queremos" en un modelo de regresión son residuales *pequeños*. Sí, eso tiene sentido. De hecho, creo que me atreveré a decir que la recta de regresión de "mejor ajuste" es la que tiene los residuales más pequeños. O, mejor aún, dado que a los estadísticos parece gustarles sacar cuadrados de todo, ¿por qué no decir eso?

> Los coeficientes de regresión estimados, $\hat{b}_0$ y $\hat{b}_1$, son aquellos que minimizan la suma de los residuales al cuadrado, que podemos escribir como $\sum_i (Y_i - \hat{Y}_i)^2$ o como $\sum_i \epsilon_i^2$.

Sí, sí, eso suena aún mejor. Y como lo he marcado así, probablemente signifique que esta es la respuesta correcta. Y dado que esta es la respuesta correcta, probablemente valga la pena tomar nota del hecho de que nuestros coeficientes de regresión son estimaciones (¡estamos tratando de adivinar los parámetros que describen una población!), razón por la cual agregué los sombreritos, para que obtengamos $\hat{b}_0$ y $\hat{b}_1$ en lugar de b_0 y b_1. Finalmente, también debo señalar que, dado que en realidad hay más de una forma de estimar un modelo de regresión, el nombre más técnico para este proceso de estimación es **regresión de mínimos cuadrados ordinarios (OLS, por sus siglas en inglés)**.

En este punto, ahora tenemos una definición concreta de lo que cuenta como nuestra "mejor" elección de coeficientes de regresión, $\hat{b}_0$ y $\hat{b}_1$. La pregunta natural a hacer a continuación es, si nuestros coeficientes de regresión óptimos son aquellos que minimizan la suma de los residuales al cuadrado, ¿cómo encontramos estos maravillosos números? La respuesta a esta pregunta es complicada y no te ayuda a entender la lógica de la regresión.[6] Esta vez te voy a dejar libre. En lugar de mostrarte primero el camino largo y tedioso y luego "revelarte" el maravilloso atajo que ofrece jamovi, vayamos directo al grano y usemos jamovi para hacer todo el trabajo pesado.

[6] O al menos, asumo que no ayuda a la mayoría de las personas. Pero en el caso de que alguien que lea esto sea un verdadero maestro de kung fu de álgebra lineal (y para ser justos, siempre tengo algunas de estas personas en mi clase de introducción a la estadística), te ayudará saber que la solución al problema de estimación resulta ser $\hat{b} = (X'X)^{-1}X'y$, donde $\hat{b}$ es un vector que contiene los coeficientes de regresión estimados, X es la "matriz de diseño" que contiene las variables predictoras (más una columna adicional que contiene todos unos; estrictamente X es una matriz de los regresores, pero aún no he discutido la distinción), e y es una vector que contiene la variable de resultado. Para todos los demás, esto no es exactamente útil y puede ser francamente aterrador. Sin embargo, dado que bastantes cosas en la regresión lineal se pueden escribir en términos de álgebra lineal, verás un montón de notas al pie como esta en este capítulo. Si puedes seguir las matemáticas en ellas, genial. Si no, ignóralas.

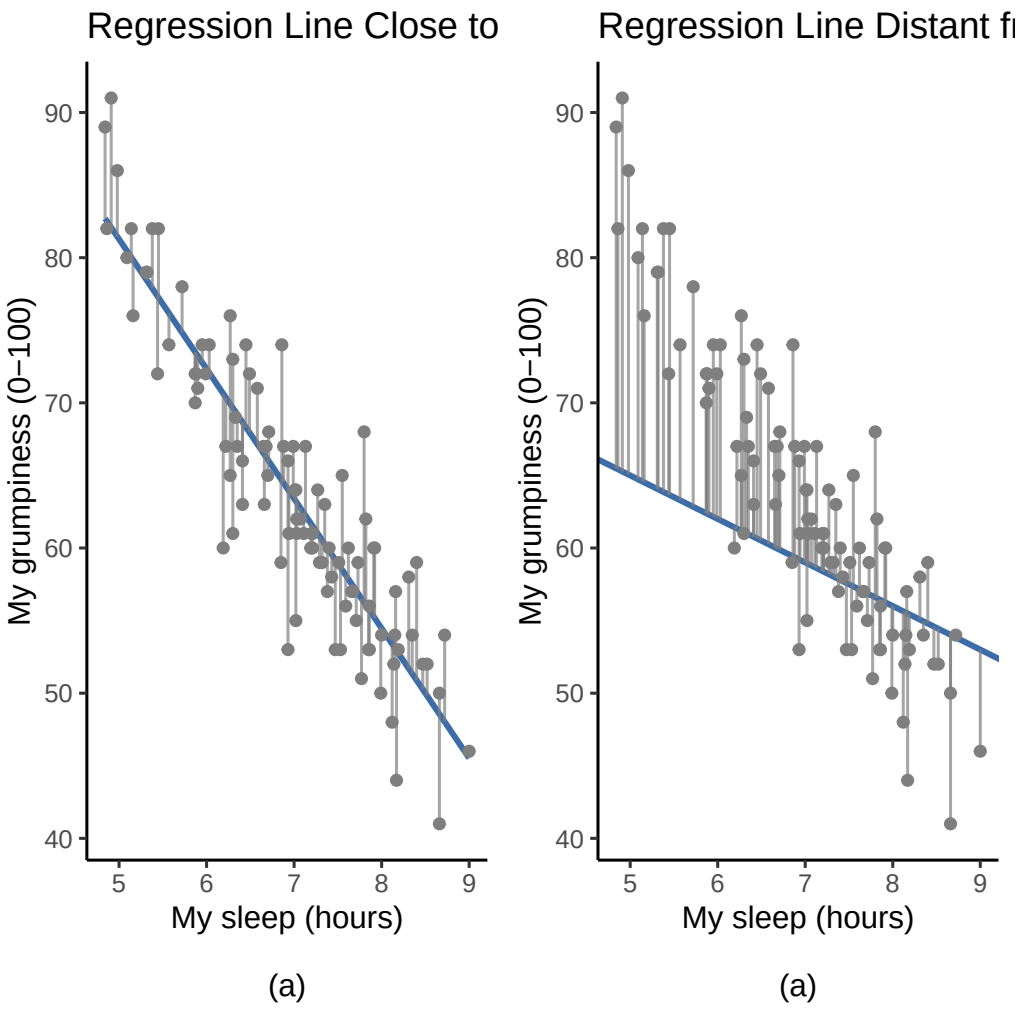


Figure 12.12: Una representación de los residuales asociados con la recta de regresión de mejor ajuste (panel a) y los residuales asociados con una línea de regresión pobre (panel b). Los residuales son mucho más pequeños para la línea de regresión buena. Una vez más, esto no es una sorpresa dado que la línea buena es la que pasa por la mitad de los datos.

12.4.1 Regresión lineal en jamovi

Para ejecutar mi regresión lineal, abre el análisis 'Regresión' - 'Regresión lineal' en jamovi, utilizando el archivo de datos parenthood.csv. Luego especifica dani.grump como la 'Variable dependiente' y dani.sleep como la variable ingresada en el cuadro 'Covariables'. Esto da los resultados que se muestran en Figure 12.13, mostrando una intersección $\hat{b}_0 = 125,96$ y la pendiente $\hat{b}_1 = -8,94$. En otras palabras, la recta de regresión de mejor ajuste que tracé en Figure 12.11 tiene esta fórmula:

$$\hat{Y}_i = 125,96 + (-8,94 X_i)$$

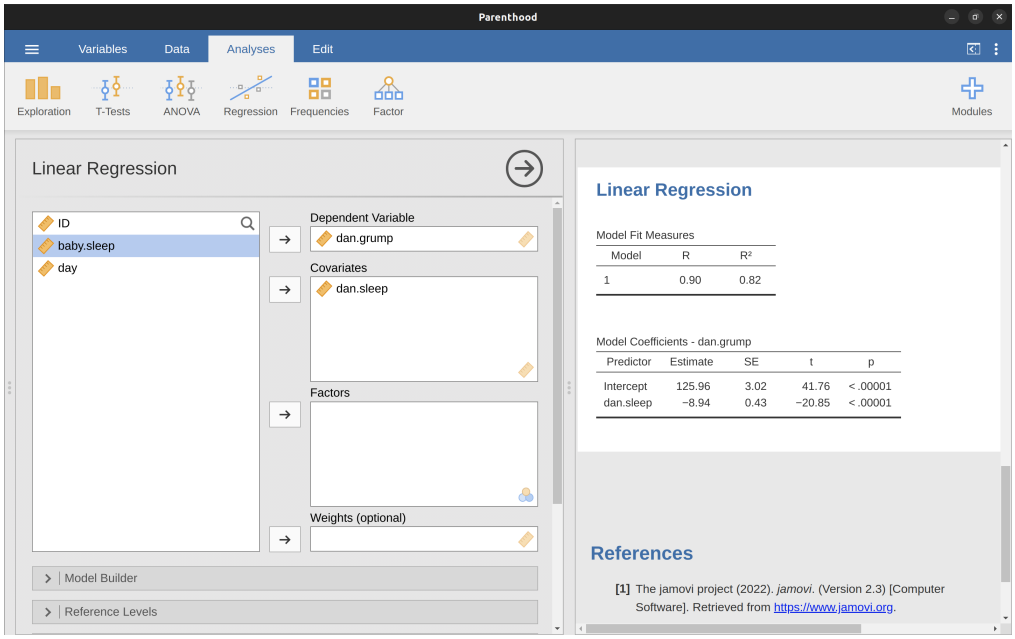


Figure 12.13: una captura de pantalla de jamovi que muestra un análisis de regresión lineal simple

12.4.2 Interpretando el modelo estimado

Lo más importante es entender cómo interpretar estos coeficientes. Comencemos con $\hat{b}_1$, la pendiente. Si recordamos la definición de la pendiente, un coeficiente de regresión de $\hat{b}_1 = -8.94$ significa que si aumento Xi en 1, entonces estoy disminuyendo Yi en 8.94. Es decir, cada hora adicional de sueño que gane mejorará mi estado de ánimo, reduciendo mi mal humor en 8,94 puntos de mal humor. ¿Qué pasa con la intersección? Bueno, dado que $\hat{b}_0$ corresponde al "valor esperado de Y_i cuando X_i es igual a 0", es bastante sencillo. Implica que si duermo cero horas ($X_i = 0$), entonces mi mal humor se saldrá de la escala, a un valor insano de ($Y_i = 125.96$). Creo que es mejor evitarlo.

12.5 Regresión lineal múltiple

El modelo de regresión lineal simple que hemos discutido hasta este punto asume que hay una sola variable predictora que te interesa, en este caso dani.sleep. De hecho, hasta este punto, todas las herramientas estadísticas de las que hemos hablado han asumido que tu análisis utiliza una variable predictora y una variable de resultado. Sin embargo, en muchos (quizás la mayoría) de los proyectos de investigación, en realidad tienes múltiples predictores que deseas examinar. Si es así, sería bueno poder extender el marco de regresión lineal para poder incluir múltiples predictores. ¿Quizás sería necesario algún tipo de modelo de **regresión múltiple**?

La regresión múltiple es conceptualmente muy sencilla. Todo lo que hacemos es agregar más términos a nuestra ecuación de regresión. Supongamos que tenemos dos variables que nos interesan; quizás queramos usar tanto dani.sleep como baby.sleep para predecir la variable dani.grump. Como antes, hacemos que Y_i se refiera a mi mal humor en el i-ésimo día. Pero ahora tenemos dos variables $ X $: la primera corresponde a la cantidad de sueño que dormí y la segunda corresponde a la cantidad de sueño que durmió mi hijo. Así que dejaremos que X_{i1} se refiera a las horas que dormí el i-ésimo día y X_{i2} se refiera a las horas que durmió el bebé ese día. Si es así, entonces podemos escribir nuestro modelo de regresión así:

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \epsilon_i$$

Como antes, ϵ_i es el residual asociado con la i-ésima observación, $\epsilon_i = Y_i - \hat{Y}_i$. En este modelo, ahora tenemos tres coeficientes que deben estimarse: b0 es la intersección, b1 es el coeficiente asociado con mi sueño y b2 es el coeficiente asociado con el sueño de mi hijo. Sin embargo, aunque el número de coeficientes que deben estimarse ha cambiado, la idea básica de cómo funciona la estimación no ha cambiado: nuestros coeficientes estimados $\hat{b}_0$, $\hat{b}_1$ y $\hat{b}_2$ son los que minimizan la seuma de los residuales al cuadrado.

12.5.1 Haciéndolo en jamovi

La regresión múltiple en jamovi no es diferente a la regresión simple. Todo lo que tenemos que hacer es agregar variables adicionales al cuadro 'Covariables' en jamovi. Por ejemplo, si queremos usar dani.sleep y baby.sleep como predictores en nuestro intento de explicar por qué estoy tan malhumorado, entonces mueve baby.sleep al cuadro 'Covariables' junto a dani.sleep. Por defecto, jamovi asume que el modelo debe incluir una intersección. Los coeficientes que obtenemos esta vez se muestran en Table 12.4.

Table 12.4: agregar múltiples variables como predictores en una regresión

(Intercept)	dani.sleep	baby.sleep
125.97	-8.95	0.01

El coeficiente asociado con dani.sleep es bastante grande, lo que sugiere que cada hora de sueño que pierdo me vuelve mucho más gruñona. Sin embargo, el coeficiente de sueño del bebé es muy pequeño, lo que sugiere que en realidad no importa cuánto duerma mi hijo. Lo que importa en cuanto a mi mal humor es cuánto duermo. Para tener una idea de cómo es este modelo de regresión múltiple, Figure 12.14 muestra un gráfico 3D que representa las tres variables, junto con el propio modelo de regresión.

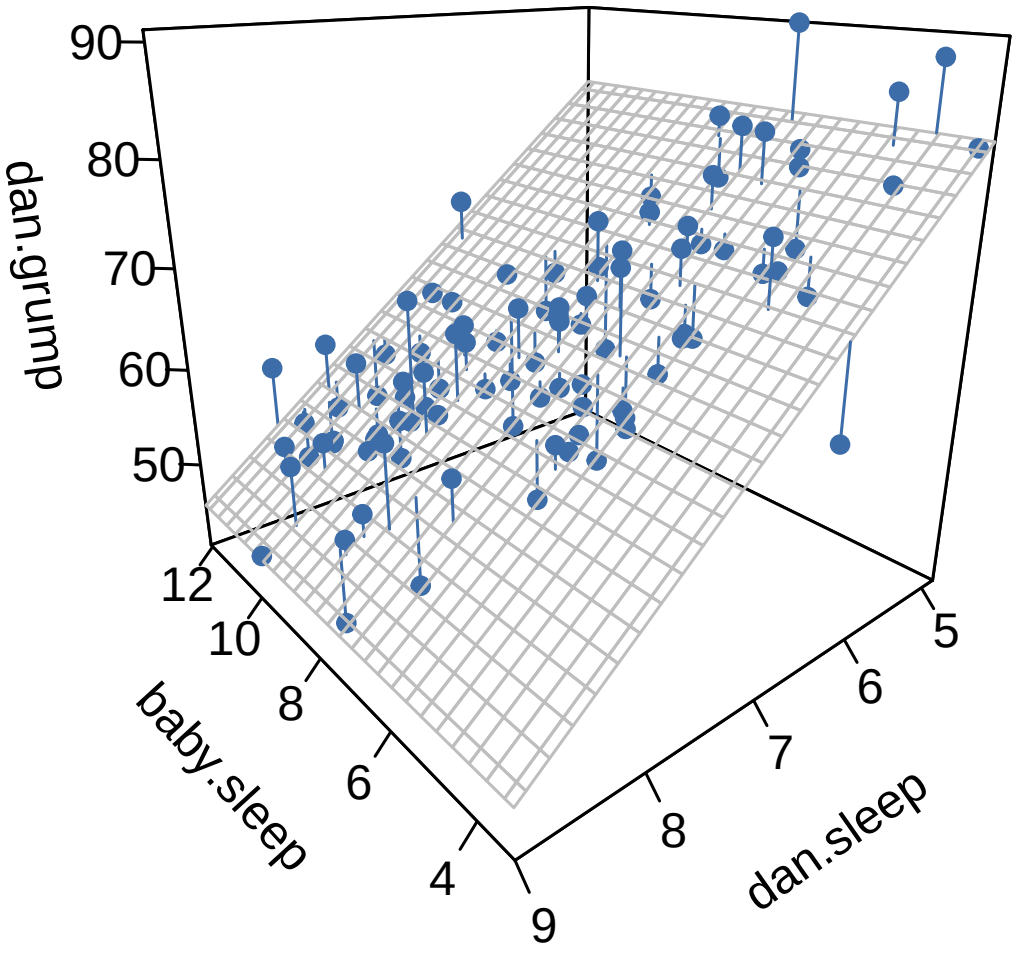


Figure 12.14: una visualización en 3D de un modelo de regresión múltiple. Hay dos predictores en el modelo, dani.sleep y baby.sleep y la variable de resultado es dani.grump. Juntas, estas tres variables forman un espacio 3D. Cada observación (punto) es un punto en este espacio. De la misma manera que un modelo de regresión lineal simple forma una línea en el espacio 2D, este modelo de regresión múltiple forma un plano en el espacio 3D. Cuando estimamos los coeficientes de regresión, lo que intentamos hacer es encontrar un plano que esté lo más cerca posible de todos los puntos azules.

[Detalle técnico adicional[7]]

12.6 Cuantificando el ajuste del modelo de regresión

Ahora sabemos cómo estimar los coeficientes de un modelo de regresión lineal. El problema es que aún no sabemos si este modelo de regresión es bueno. Por ejemplo, el modelo de regresión.1 afirma que cada hora de sueño mejorará bastante mi estado de ánimo, pero podría ser una tontería. Recuerda, el modelo de regresión solo produce una predicción $\hat{Y}_i$ sobre cómo es mi estado de ánimo, pero mi estado de ánimo real es Y_i. Si estos dos están muy cerca, entonces el modelo de regresión ha hecho un buen trabajo. Si son muy diferentes, entonces ha hecho un mal trabajo.

12.6.1 El valor de R^2

Una vez más, pongamos un poco de matemática alrededor de esto. En primer lugar, tenemos la suma de los residuales al cuadrado

$$SS_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

que esperamos que sea bastante pequeña. Específicamente, lo que nos gustaría es que sea muy pequeña en comparación con la variabilidad total en la variable de resultado.

$$SS_{tot} = \sum_i (Y_i - \bar{Y})^2$$

Ya que estamos aquí, calculemos estos valores nosotras mismas, aunque no a mano. Usemos algo como Excel u otro programa de hoja de cálculo estándar. Hice esto abriendo el archivo parenthood.csv en Excel y guardándolo como parenthood rsquared.xls para poder trabajar en él. Lo primero que debes hacer es calcular los valores de $\hat{Y}$, y para el modelo simple que usa solo un único predictor, haríamos lo siguiente:

1. crea una nueva columna llamada 'Y.pred' usando la fórmula '= 125.97 + (-8.94 × dani.sleep)'
2. calcula el SS(resid) creando una nueva columna llamada '(YY.pred)^2' utilizando la fórmula ' = (dani.grump - Y.pred)^2 '.
3. Luego, en la parte inferior de esta columna, calcula la suma de estos valores, es decir, ' sum((YY.pred)^2) .
4. En la parte inferior de la columna dani.grump, calcula el valor medio para dani.grump (NB Excel usa la palabra 'PROMEDIO' en lugar de 'promedio' en su función).

[7] la fórmula general: la ecuación que di en el texto principal muestra cómo es un modelo de regresión múltiple cuando incluye dos predictores. Entonces, no es sorprendente que si deseas más de dos predictores, todo lo que tienes que hacer es agregar más términos X y más coeficientes b. En otras palabras, si tienes K variables predictoras en el modelo, la ecuación de regresión se verá así

$$Y_i = b_0 + (\sum_{k=1}^{K} b_k X_{ik}) + \epsilon_i$$

5. Luego crea una nueva columna, llamada ' (Y - mean(Y))^2)' usando la fórmula ' = (dani.grump - AVERAGE(dani.grump))^2 '.
6. Luego, en la parte inferior de esta columna, calcula la suma de estos valores, es decir, 'sum((Y - mean(Y))^2)'.
7. Calcula R.squared escribiendo en una celda en blanco lo siguiente: '= 1 - (SS(resid) / SS(tot))'.

Esto da un valor para R^2 de '0.8161018'. El valor R^2, a veces llamado **coeficiente de determinación**[8] tiene una interpretación simple: es la proporción de la varianza en la variable de resultado que puede ser explicada por el predictor. Entonces, en este caso, el hecho de que hayamos obtenido $R^2 = .816$ significa que el predictor (my.sleep) explica 81.6% de la varianza del resultado (my.grump).

Naturalmente, no necesitas escribir todos estos comandos en Excel tú misma si deseas obtener el valor de R^2 para tu modelo de regresión. Como veremos más adelante en la sección sobre [Ejecutar las pruebas de hipótesis en jamovi], todo lo que necesitas hacer es especificar esto como una opción en jamovi. Sin embargo, dejemos eso a un lado por el momento. Hay otra propiedad de R^2 que quiero señalar.

12.6.2 La relación entre regresión y correlación

En este punto podemos revisar mi afirmación anterior de que la regresión, en esta forma tan simple que he discutido hasta ahora, es básicamente lo mismo que una correlación. Anteriormente, usamos el símbolo r para indicar una correlación de Pearson. ¿Podría haber alguna relación entre el valor del coeficiente de correlación r y el valor de R^2 de la regresión lineal? Por supuesto que la hay: la correlación al cuadrado r^2 es idéntica al valor de R^2 para una regresión lineal con un solo predictor. En otras palabras, ejecutar una correlación de Pearson es más o menos equivalente a ejecutar un modelo de regresión lineal que usa solo una variable predictora.

12.6.3 El valor R^2 ajustado

Una última cosa a señalar antes de continuar. Es bastante común informar de una medida ligeramente diferente del rendimiento del modelo, conocida como "R^2 ajustado". La razón que subyace al cálculo del valor de R^2 ajustado es la observación de que agregar más predictores al modelo siempre hará que el valor de R^2 aumente (o al menos no disminuya).

[Detalle técnico adicional[9]]

Este ajuste es un intento de tener en cuenta los grados de libertad. La gran ventaja del valor de R^2 ajustado es que cuando agregas más predictores al modelo, el valor de R^2 ajustado solo aumentará si las nuevas variables mejoran el rendimiento del modelo más de lo esperado por casualidad. La gran desventaja es que el valor ajustado de R^2

[8] Y por "a veces" me refiero a "casi nunca". En la práctica, todo el mundo lo llama simplemente "R-cuadrado".

[9] el valor R^2 ajustado introduce un ligero cambio en el cálculo, como se muestra a continuación. Para un modelo de regresión con K predictores, ajustado a un conjunto de datos que contiene N observaciones, el R^2 ajustado es:

$$\text{adj.}R^2 = 1 - (\frac{SS_{res}}{SS_{tot}} \times \frac{N-1}{NK-1})$$

no se puede interpretar de la forma elegante en que se puede interpretar R^2. R^2 tiene una interpretación simple como la proporción de varianza en la variable de resultado que se explica por el modelo de regresión. Que yo sepa, no existe una interpretación equivalente para R^2 ajustado.

Entonces, una pregunta obvia es si debes informar R^2 o ajustar R^2. Esto es probablemente una cuestión de preferencia personal. Si te importa más la interpretabilidad, entonces R^2 es mejor. Si te importa más corregir el sesgo, probablemente sea mejor ajustar R^2. Personalmente, prefiero R^2. Mi sensación es que es más importante poder interpretar la medida del rendimiento del modelo. Además, como veremos en Pruebas de hipótesis para modelos de regresión, si te preocupa que la mejora en R^2 que obtienes al agregar un predictor se deba solo al azar y no porque sea un mejor modelo, bueno, tenemos pruebas de hipótesis para eso.

12.7 Pruebas de hipótesis para modelos de regresión

Hasta ahora hemos hablado sobre qué es un modelo de regresión, cómo se estiman los coeficientes de un modelo de regresión y cómo cuantificamos el rendimiento del modelo (el último de estos, por cierto, es básicamente nuestra medida del tamaño del efecto). Lo siguiente de lo que tenemos que hablar es de las pruebas de hipótesis. Hay dos tipos diferentes (pero relacionados) de pruebas de hipótesis de las que debemos hablar: aquellas en las que probamos si el modelo de regresión como un todo está funcionando significativamente mejor que un modelo nulo, y aquellas en las que probamos si un coeficiente de regresión particular es significativamente diferente de cero.

12.7.1 Probando el modelo como un todo

Bien, supongamos que has estimado tu modelo de regresión. La primera prueba de hipótesis que puedes probar es la hipótesis nula de que no existe una relación entre los predictores y el resultado, y la hipótesis alternativa de que los datos se distribuyen exactamente de la manera que predice el modelo de regresión.

[Detalle técnico adicional[10]]

[10] formalmente, nuestro "modelo nulo" corresponde al modelo de "regresión" bastante trivial en el que incluimos 0 predictores y solo incluimos el término de intersección b_0: $H_0 : Y_0 = b_0 + \epsilon_i$ Si nuestro modelo de regresión tiene K predictores, el "modelo alternativo" se describe utilizando la fórmula habitual para un modelo de regresión múltiple:

$$H_1 : Y_i = b_0 + (\sum_{k=1}^{K} b_k X_{ik}) + \epsilon_i$$

¿Cómo podemos contrastar estas dos hipótesis? El truco es entender que es posible dividir la varianza total $SStot$ en la suma de la varianza residual SCres y la varianza del modelo de regresión SCmod. Me saltaré los tecnicismos, ya que llegaremos a eso más adelante cuando veamos ANOVA en Chapter 13. Pero ten en cuenta que $SS_{mod} = SS_{tot} - SS_{res}$ Y podemos convertir las sumas de cuadrados en medias cuadráticas dividiendo por los grados de libertad.

$$MS_{mod} = \frac{SS_{mod}}{df_{mod}}$$

$$MS_{res} = \frac{SS_{res}}{df_{res}}$$

Entonces, ¿cuántos grados de libertad tenemos? Como es de esperar, el gl asociado con el modelo está estrechamente relacionado con la cantidad de predictores que hemos incluido. De hecho, resulta que

Veremos mucho más del estadístico F en Chapter 13, pero por ahora solo ten en cuenta que podemos interpretar valores grandes de F como una indicación de que la hipótesis nula está funcionando mal en comparación con la hipótesis alternativa. En un momento te mostraré cómo hacer la prueba en jamovi de la manera más fácil, pero primero echemos un vistazo a las pruebas para los coeficientes de regresión individuales.

12.7.2 Pruebas para coeficientes individuales

La prueba F que acabamos de presentar es útil para comprobar que el modelo en su conjunto funciona mejor que el azar. Si tu modelo de regresión no produce un resultado significativo para la prueba F, probablemente no tengas un modelo de regresión muy bueno (o, muy posiblemente, no tengas muy buenos datos). Sin embargo, mientras que fallar esta prueba es un indicador bastante fuerte de que el modelo tiene problemas, pasar la prueba (es decir, rechazar el valor nulo) no implica que el modelo sea bueno. ¿Por qué ocurre eso, te estarás preguntando? La respuesta se puede encontrar mirando los coeficientes del modelo Regresión lineal múltiple que ya hemos visto (Table 12.4)

No puedo dejar de notar que el coeficiente de regresión estimado para la variable baby.sleep es pequeño (0.01), en relación al valor que obtenemos para dani.sleep (−8.95). Dado que estas dos variables están en la misma escala (ambas se miden en "horas dormidas"), me parece que esto es esclarecedor. De hecho, estoy empezando a sospechar que en realidad solo la cantidad de sueño que duermo es lo que importa para predecir mi mal humor. Podemos reutilizar una prueba de hipótesis que discutimos anteriormente, la prueba t. La prueba que nos interesa tiene una hipótesis nula de que el verdadero coeficiente de regresión es cero ($b = 0$), que debe probarse contra la hipótesis alternativa de que no lo es ($b \neq 0$). Eso es:

$$H_0 : b = 0$$
$$H_1 : b \neq 0$$

¿Cómo podemos probar esto? Bueno, si el teorema central del límite es bueno con nosotros, podríamos suponer que la distribución muestral de $\hat{b}$, el coeficiente de regresión estimado, es una distribución normal con la media centrada en b. Lo que eso significaría es que si la hipótesis nula fuera cierta, entonces la distribución muestral de $\hat{b}$ tiene una media cero y una desviación estándar desconocida. Suponiendo que podemos llegar a una buena estimación del error estándar del coeficiente de regresión, $se(\hat{b})$, entonces tenemos suerte. Esa es exactamente la situación para la que introdujimos la prueba t de una muestra en Chapter 11. Así que definamos un estadístico t como este

$$t = \frac{\hat{b}}{SE(\hat{b})}$$

Pasaré por alto las razones, pero nuestros grados de libertad en este caso son $df = N - K - 1$. De manera irritante, la estimación del error estándar del coeficiente de

$df_mod = K$. Para los residuales, los grados de libertad totales son $df_res = N - K - 1$. Ahora que tenemos nuestras medias cuadráticas, podemos calcular un estadístico F como este

$$F = \frac{MS_{mod}}{MS_{res}}$$

y los grados de libertad asociados con esto son $K $ y $N - K - 1$.

regresión, $se(\hat{b})$, no es tan fácil de calcular como el error estándar de la media que usamos para las pruebas t más sencillas en Chapter 11. De hecho, la fórmula es algo fea y no muy útil de ver.[11] Para nuestros propósitos, es suficiente señalar que el error estándar del coeficiente de regresión estimado depende de las variables predictoras y de resultado, y es algo sensible a las violaciones del supuesto de homogeneidad de varianzas (discutido en breve).

En cualquier caso, este estadístico t se puede interpretar de la misma manera que los estadísticos t que analizamos en Chapter 11. Suponiendo que tienes una alternativa de dos colas (es decir, no te importa si b $>$ 0 o b $<$ 0), entonces son los valores extremos de t (es decir, mucho menos que cero o mucho mayor que cero) que sugieren que debes rechazar la hipótesis nula.

12.7.3 Ejecutando las pruebas de hipótesis en jamovi

Para calcular todos los estadísticos de los que hemos hablado hasta ahora, todo lo que necesitas hacer es asegurarte de que las opciones relevantes estén marcadas en jamovi y luego ejecutar la regresión. Si hacemos eso, como en Figure 12.15, obtenemos una gran cantidad de resultados útiles.

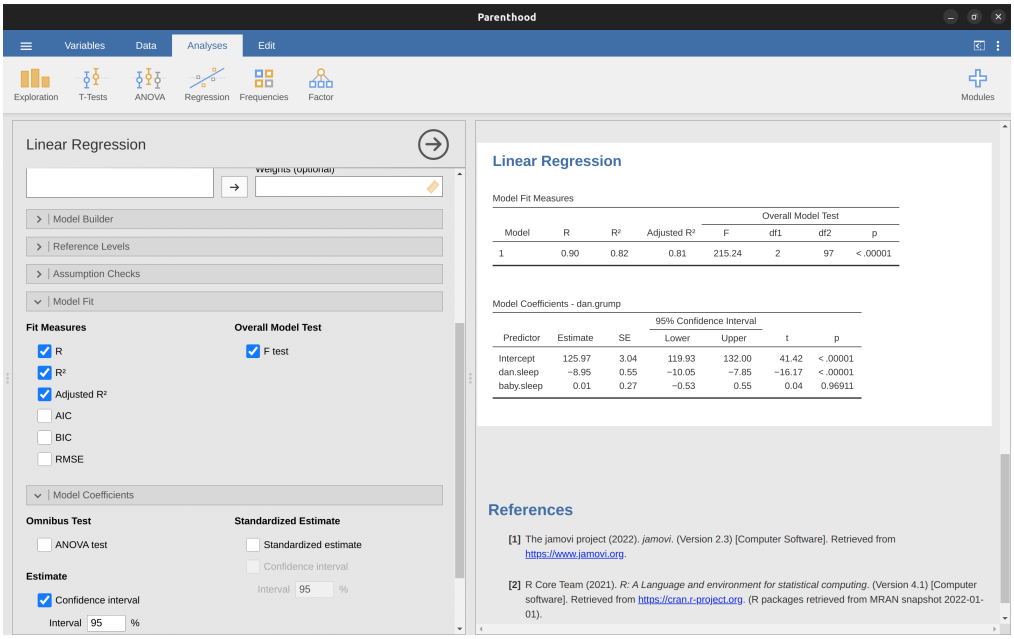


Figure 12.15: una captura de pantalla de jamovi que muestra un análisis de regresión lineal múltiple, con algunas opciones útiles marcadas

Los 'Coeficientes del modelo' en la parte inferior de los resultados del análisis jamovi que se muestran en Figure 12.15 proporcionan los coeficientes del modelo de regresión. Cada

[11] solo para lectores avanzados. El vector de residuales es $\epsilon = y - X\hat{b}$. Para K predictores más la intersección, la varianza residual estimada es $\hat{\sigma}^2 = \frac{\epsilon' \epsilon}{(N-K-1)}$. La matriz de covarianza estimada de los coeficientes es $\hat{\sigma}^2 (X' X)^{-1}$, cuya diagonal principal es $se(\hat{b})$, nuestros errores estándar estimados.

fila de esta tabla se refiere a uno de los coeficientes del modelo de regresión. La primera fila es el término de intersección, y las últimas miran cada uno de los predictores. Las columnas te ofrecen toda la información relevante. La primera columna es la estimación real de b (por ejemplo, 125, 97 para la intersección y -8,95 para el predictor dani.sleep). La segunda columna es la estimación del error estándar $\hat{\sigma}_b$. Las columnas tercera y cuarta proporcionan los valores inferior y superior para el intervalo de confianza del 95% alrededor de la estimación b (más sobre esto más adelante). La quinta columna te da el estadístico t, y vale la pena notar que en esta tabla $t = \frac{\hat{b}}{se(\hat{b})}$. Finalmente, la última columna te muestra el valor p real para cada una de estas pruebas.[12]

Lo único que la tabla de coeficientes en sí no incluye son los grados de libertad utilizados en la prueba t, que siempre es $N - K - 1$ y se enumeran en la parte superior de la tabla, etiquetada como 'Medidas de ajuste del modelo'. Podemos ver en esta tabla que el modelo funciona significativamente mejor de lo que cabría esperar por casualidad ($F(2, 97) = 215.24, p < .001$), lo cual no es tan sorprendente: el valor $R^2 = .81$ indica que el modelo de regresión representa 81% de la variabilidad en la medida de resultado (y 82% para el R^2 ajustado). Sin embargo, cuando volvemos a mirar las pruebas t para cada uno de los coeficientes individuales, tenemos pruebas bastante sólidas de que la variable sueño del bebé no tiene un efecto significativo. Todo el trabajo en este modelo lo realiza la variable dani.sleep. En conjunto, estos resultados sugieren que este modelo de regresión es en realidad el modelo incorrecto para los datos. Probablemente sea mejor que deje por completo el predictor del sueño del bebé. En otras palabras, el modelo de regresión simple con el que comenzamos es el mejor modelo.

12.8 Sobre los coeficientes de regresión

Antes de pasar a discutir los supuestos subyacentes a la regresión lineal y lo que puedes hacer para verificar si se cumplen, hay dos temas más que quiero discutir brevemente, los cuales se relacionan con los coeficientes de regresión. Lo primero de lo que hablar es de calcular los intervalos de confianza para los coeficientes. Después de eso, discutiré la cuestión un tanto turbia de cómo determinar qué predictor es el más importante.

12.8.1 Intervalos de confianza para los coeficientes

Como cualquier parámetro poblacional, los coeficientes de regresión b no pueden estimarse con total precisión a partir de una muestra de datos; eso es parte de por qué necesitamos pruebas de hipótesis. Dado esto, es muy útil poder informar intervalos de confianza que capturen nuestra incertidumbre sobre el verdadero valor de b. Esto es especialmente útil cuando la pregunta de investigación se centra en gran medida en un intento de averiguar la relación entre la variable X y la variable Y, ya que en esas situaciones el interés está principalmente en el peso de regresión b.

[Detalle técnico adicional[13]]

[12] ten en cuenta que, aunque jamovi ha realizado varias pruebas aquí, no ha realizado una corrección de Bonferroni ni nada (consulta Chapter 13). Estas son pruebas t estándar de una muestra con una alternativa bilateral. Si quieres realizar correcciones para varias pruebas, debes hacerlo tú misma.

[13] Afortunadamente, los intervalos de confianza para los pesos de regresión se pueden construir de la forma habitual $CI(b) = \hat{b} \pm (t_{crit} \times SE(\hat{b}))$ donde $se(\hat{b})$ es el error estándar del coeficiente de regresión y t_crit es el valor crítico relevante de la distribución t apropiada. Por ejemplo, si lo que queremos es un intervalo de confianza del 95 %, entonces el valor crítico es el cuantil 97, 5 de la distribución at con

En jamovi ya habíamos especificado el 'intervalo de confianza del 95%' como se muestra en Figure 12.15, aunque podríamos haber elegido fácilmente otro valor, digamos un 'intervalo de confianza del 99%' si eso es lo que decidimos.

12.8.2 Cálculo de coeficientes de regresión estandarizados

Una cosa más que quizás quieras hacer es calcular los coeficientes de regresión "estandarizados", a menudo denominados β. La lógica de los coeficientes estandarizados es la siguiente. En muchas situaciones, tus variables están en escalas diferentes. Supongamos, por ejemplo, que mi modelo de regresión pretende predecir las puntuaciones de IQ de las personas utilizando su nivel educativo (número de años de educación) y sus ingresos como predictores. Obviamente, el nivel educativo y los ingresos no están en la misma escala. La cantidad de años de escolaridad solo puede variar en decenas de años, mientras que los ingresos pueden variar en $ 10,000 dólares (o más). Las unidades de medida tienen una gran influencia en los coeficientes de regresión. Los coeficientes b solo tienen sentido cuando se interpretan a la luz de las unidades, tanto de las variables predictoras como de la variable resultado. Esto hace que sea muy difícil comparar los coeficientes de diferentes predictores. Sin embargo, hay situaciones en las que realmente deseas hacer comparaciones entre diferentes coeficientes. Específicamente, es posible que quieras algún tipo de medida estándar de qué predictores tienen la relación más fuerte con el resultado. Esto es lo que pretenden hacer los **coeficientes estandarizados**.

La idea básica es bastante simple; los coeficientes estandarizados son los coeficientes que habrías obtenido si hubieras convertido todas las variables a puntuaciones z antes de ejecutar la regresión.[14] La idea aquí es que, al convertir todos los predictores en puntuaciones z, todos entran en la regresión en la misma escala, eliminando así el problema de tener variables en diferentes escalas. Independientemente de cuáles fueran las variables originales, un valor β de 1 significa que un aumento en el predictor de 1 desviación estándar producirá un aumento correspondiente de 1 desviación estándar en la variable de resultado. Por lo tanto, si la variable A tiene un valor absoluto de β mayor que la variable B, se considera que tiene una relación más fuerte con el resultado. O al menos esa es la idea. Vale la pena ser un poco cauteloso aquí, ya que esto depende en gran medida del supuesto de que "un cambio de 1 desviación estándar" es fundamentalmente lo mismo para todas las variables. No siempre es obvio que esto sea cierto.

[Detalle técnico adicional[15]]

Para simplificar aún más las cosas, jamovi tiene una opción que calcula los coeficientes β por ti usando la casilla de verificación 'Estimación estandarizada' en las opciones 'Coeficientes del modelo', ver los resultados en Figure 12.16.

$N - K - 1$ grados de libertad. En otras palabras, este es básicamente el mismo enfoque para calcular los intervalos de confianza que hemos usado en todo momento.

[14] Estrictamente, estandarizas todos los *regresores*. Es decir, cada "cosa" que tiene asociado un coeficiente de regresión en el modelo. Para los modelos de regresión de los que he hablado hasta ahora, cada variable predictora se asigna exactamente a un regresor y viceversa. Sin embargo, eso no es cierto en general y veremos algunos ejemplos de esto más adelante en Chapter 14. Pero, por ahora, no necesitamos preocuparnos demasiado por esta distinción.

[15] Dejando de lado los problemas de interpretación, veamos cómo se calcula. Lo que podrías hacer es estandarizar todas las variables tú misma y luego ejecutar una regresión, pero hay una forma mucho más sencilla de hacerlo. Resulta que el coeficiente β para un predictor X y un resultado Y tiene una fórmula muy simple, a saber, $\beta_X = b_X \times \frac{\sigma_X}{\sigma_Y}$ donde $sigma_X$ es la desviación estándar del predictor, y Y es la desviación estándar de la variable de resultado Y. Esto simplifica mucho las cosas.

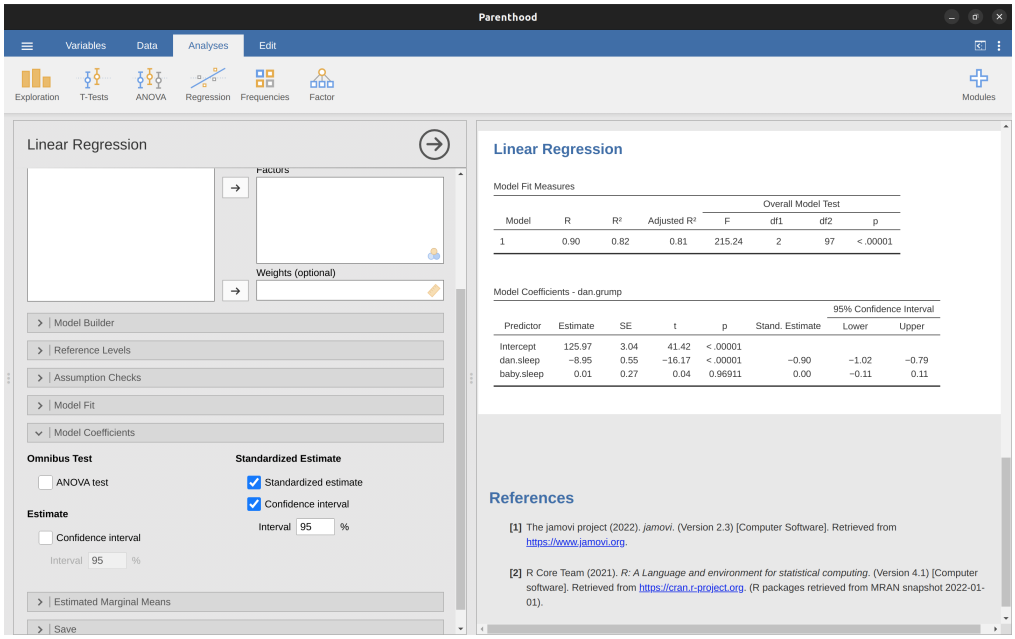


Figure 12.16: Coeficientes estandarizados, con intervalos de confianza del 95%, para regresión lineal múltiple

Estos resultados muestran claramente que la variable dani.sleep tiene un efecto mucho más fuerte que la variable baby.sleep. Sin embargo, este es un ejemplo perfecto de una situación en la que probablemente tendría sentido utilizar los coeficientes b originales en lugar de los coeficientes estandarizados β. Después de todo, mi sueño y el sueño del bebé ya están en la misma escala: número de horas dormidas. ¿Por qué complicar las cosas al convertirlos en puntuaciones z?

12.9 Supuestos de regresión

El modelo de regresión lineal que he estado discutiendo se basa en varios supuestos. En Comprobación de modelos hablaremos mucho más sobre cómo comprobar que se cumplen estos supuestos, pero primero echemos un vistazo a cada uno de ellos.

- **L**inealidad. Un supuesto bastante fundamental del modelo de regresión lineal es que la relación entre X y Y en realidad es lineal. Independientemente de si se trata de una regresión simple o una regresión múltiple, asumimos que las relaciones involucradas son lineales.
- **I**ndependencia: los residuales son independientes entre sí. En realidad, se trata de un supuesto general, en el sentido de que "no hay nada raro en los residuales". Si ocurre algo extraño (p. ej., todos los residuales dependen en gran medida de alguna otra variable no medida), podría estropear las cosas. La independencia no es algo que puedas verificar directa y específicamente con herramientas de diagnóstico, pero si tus diagnósticos de regresión están en mal estado, piensa detenidamente en la independencia de tus observaciones y residuales.

- **N**ormalidad. Como muchos de los modelos en estadística, la regresión lineal simple o múltiple básica se basa en un supuesto de normalidad. Específicamente, asume que los residuales se distribuyen normalmente. En realidad, está bien si los predictores X y las variables de resultado Y no son normales, siempre que los residuales ϵ sean normales. Consulta la sección [Comprobación de la normalidad de los residuos].
- **E**calidad (u 'homogeneidad') de la varianza. Estrictamente hablando, el modelo de regresión asume que cada residual ϵ_i se genera a partir de una distribución normal con media 0 y (más importante para los propósitos actuales) con una desviación estándar σ que es la misma para cada residual. En la práctica, es imposible probar el supuesto de que todos los residuales se distribuyen de manera idéntica. En cambio, lo que nos interesa es que la desviación estándar del residual sea la misma para todos los valores de $\hat{Y}$ y (si somos especialmente diligentes) para todos los valores de cada predictor X en el modelo.

Entonces, tenemos cuatro supuestos principales para la regresión lineal (que forman claramente el acrónimo '**LINE**'). Y además hay un par de cosas que también debemos verificar:

- Predictores no correlacionados. La idea aquí es que, en un modelo de regresión múltiple, no deseas que tus predictores estén demasiado correlacionados entre sí. Esto no es "técnicamente" un supuesto del modelo de regresión, pero en la práctica es necesario. Los predictores que están demasiado correlacionados entre sí (lo que se conoce como "colinealidad") pueden causar problemas al evaluar el modelo. Consulta la sección Comprobación de la colinealidad.
- No hay valores atípicos "malos". Nuevamente, en realidad no es un supuesto técnico del modelo (o más bien, está implícito en todos los demás), pero hay un supuesto implícito de que tu modelo de regresión no está muy influenciado por uno o dos puntos de datos anómalos porque esto plantea dudas sobre la idoneidad del modelo y la fiabilidad de los datos en algunos casos. Consulta la sección sobre Datos atípicos y anómalos.

12.10 Comprobación del modelo

Esta sección se centra en el **diagnóstico de regresión**, un término que se refiere al arte de verificar que se hayan cumplido los supuestos de tu modelo de regresión, descubrir cómo arreglar el modelo si se violan los supuestos y, en general, comprobar que no pasa nada "raro". Me refiero a esto como el "arte" de la verificación de modelos por una buena razón. No es fácil, y aunque hay muchas herramientas fácilmente disponibles que puedes usar para diagnosticar y tal vez incluso arreglar los problemas que afectan a tu modelo (si es que hay alguno), realmente hay que tener cierto criterio al hacerlo.

En esta sección, describo varias cosas diferentes que puedes hacer para comprobar que tu modelo de regresión está haciendo lo que se supone que debe hacer. No cubre todas las cosas que podrías hacer, pero aún así es mucho más detallado de lo que se hace a menudo en la práctica, desafortunadamente. Pero es importante que tengas una idea de las herramientas que tienes a tu disposición, así que trataré de presentar algunas de ellas aquí. Finalmente, debo señalar que esta sección se basa en gran medida en @ Fox2011, el libro asociado con el paquete 'car' que se usa para realizar análisis de regresión en R. El paquete 'car' se destaca por proporcionar algunas herramientas

excelentes para el diagnóstico de regresión, y el libro mismo habla de ellos de una manera admirablemente clara. No quiero sonar demasiado efusiva al respecto, pero creo que vale la pena leer @ Fox2011, incluso si algunas de las técnicas de diagnóstico avanzadas solo están disponibles en R y no en jamovi.

12.10.1 Tres tipos de residuales

La mayoría de los diagnósticos de regresión giran en torno a la observación de los residuales, y existen varios tipos diferentes de residuales que podríamos considerar. En particular, en esta sección se hace referencia a los siguientes tres tipos de residuales: "residuales ordinarios", "residuales estandarizados" y "residuales estudentizados". Hay un cuarto tipo al que verás que se hace referencia en algunas de las Figuras, y ese es el "residual de Pearson". Sin embargo, para los modelos de los que estamos hablando en este capítulo, el residual de Pearson es idéntico al residual ordinario.

El primer y más simple tipo de residuales que nos interesan son los **residuales ordinarios**. Estos son los residuales brutos reales de los que he estado hablando a lo largo de este capítulo hasta ahora. El residual ordinario es simplemente la diferencia entre el valor predicho $\hat{Y}_i$ y el valor observado Y_i. He estado usando la notación ϵ_i para referirme al residual ordinario i-ésimo y así, con esto en mente, tenemos la ecuación muy simple

$$\epsilon_i = Y_i - \hat{Y}_i$$

Por supuesto, esto es lo que vimos antes y, a menos que me refiera específicamente a algún otro tipo de residual, este es del que estoy hablando. Así que no hay nada nuevo aquí. Solo quería repetirme. Una desventaja de usar residuales ordinarios es que siempre están en una escala diferente, dependiendo de cuál sea la variable de resultado y cómo de bueno sea el modelo de regresión. Es decir, a menos que hayas decidido ejecutar un modelo de regresión sin un término de intersección, los residuales ordinarios tendrán media 0 pero la varianza es diferente para cada regresión. En muchos contextos, especialmente donde solo estás interesada en el patrón de los residuales y no en sus valores reales, es conveniente estimar los **residuales estandarizados**, que se normalizan de tal manera que tienen una desviación estándar de 1.

[Detalle técnico adicional[16]]

El tercer tipo de residuales son los **residuales estudentizados** (también llamados "residuales jackknifed") y son incluso más sofisticados que los residuales estandarizados. Nuevamente, la idea es coger el residuo ordinario y dividirlo por alguna cantidad para estimar alguna noción estandarizada del residual. [17]

[16] la forma en que los calculamos es dividir el residal ordinario por una estimación de la desviación estándar (poblacional) de estos residuales. Por razones técnicas, la fórmula para esto es

$$\epsilon_i' = \frac{\epsilon_i}{\hat{\sigma}\sqrt{1-h_i}}$$

donde $\hat{\sigma}$ en este contexto es la desviación estándar de la población estimada de los residuales ordinarios, y h_i es el "valor sombrero" de la $i$ésima observación. Todavía no te he explicado los valores sombrero, así que esto no tendrá mucho sentido. Por ahora, basta con interpretar los residuales estandarizados como si hubiéramos convertido los residuales ordinarios en puntuaciones z.

[17] La fórmula para hacer los cálculos esta vez es sutilmente diferente $\epsilon_i^* = \frac{\epsilon_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}}$ Fíjate que nuestra estimación de la desviación estándar aquí se escribe $\hat{\sigma}_{(-i)}$. Esto corresponde a la estimación de la desviación estándar residual que habría obtenido si hubiera eliminado la i-ésima observación del

Antes de continuar, debo señalar que a menudo no es necesario obtener estos residuales por ti misma, a pesar de que son la base de casi todos los diagnósticos de regresión. La mayoría de las veces, las diversas opciones que proporcionan los diagnósticos, o las comprobaciones de supuestos, se encargarán de estos cálculos por ti. Aun así, siempre es bueno saber cómo obtener estas cosas tú misma en caso de que alguna vez necesites hacer algo no estándar.

12.10.2 Verificando la linealidad de la relación

Deberíamos verificar la linealidad de las relaciones entre los predictores y los resultados. Hay diferentes cosas que podrías hacer para verificar esto. En primer lugar, nunca está de más trazar la relación entre los valores predichos $\hat{Y}_i$ y los valores observados Y_i para la variable de resultado, como se ilustra en Figure 12.17. Para dibujar esto en jamovi, guardamos los valores predichos en el conjunto de datos y luego dibujamos un diagrama de dispersión de los valores observados contra los predichos (ajustados). Esto te da una especie de "vista general": si este gráfico se ve aproximadamente lineal, entonces probablemente no lo estemos haciendo tan mal (aunque eso no quiere decir que no haya problemas). Sin embargo, si puedes ver grandes desviaciones de la linealidad aquí, entonces sugiere que necesitas hacer algunos cambios.

En cualquier caso, para obtener una imagen más detallada, a menudo es más informativo observar la relación entre los valores pronosticados y los residuales mismos. Nuevamente, en jamovi puedes guardar los residuales en el conjunto de datos y luego dibujar un diagrama de dispersión de los valores pronosticados contra los valores residuales, como en Figure 12.18. Como puedes ver, no solo dibuja el diagrama de dispersión que muestra el valor pronosticado contra los residuales, sino que también puede trazar una línea a través de los datos que muestra la relación entre los dos. Idealmente, debería ser una línea recta y perfectamente horizontal. En la práctica, buscamos una línea razonablemente recta o plana. Es una cuestión de criterio.

Se producen versiones algo más avanzadas del mismo gráfico al marcar 'Gráficos de residuales' en las opciones de análisis de regresión 'Comprobaciones de supuestos' en jamovi. Estos son útiles no solo para verificar la linealidad, sino también para verificar la normalidad y el supuesto de homogeneidad de varianzas, y los analizamos con más detalle en Section 12.10.3. Esta opción no solo dibuja gráficos que comparan los valores pronosticados con los residuales, sino que también lo hace para cada predictor individual.

12.10.3 Comprobación de la normalidad de los residuales

Como muchas de las herramientas estadísticas que hemos discutido en este libro, los modelos de regresión se basan en un supuesto de normalidad. En este caso, asumimos que los residuales se distribuyen normalmente. Lo primero que podemos hacer es dibujar un gráfico QQ a través de la opción 'Comprobaciones de supuestos' - 'Comprobaciones de supuestos' - 'Gráfico QQ de residuales'. El resultado se muestra en Figure 12.19, que muestra los residuales estandarizados representados en función de sus cuantiles teóricos según el modelo de regresión.

conjunto de datos. Esto parece una pesadilla de calcular, ya que parece estar diciendo que tienes que ejecutar N nuevos modelos de regresión (incluso un ordenador moderno podría quejarse un poco de eso, especialmente si tienes un gran conjunto de datos). Afortunadamente, esta desviación estándar estimada en realidad viene dada por la siguiente ecuación: $\hat{\sigma}_{(-i)} = \hat{\sigma}\sqrt{\frac{NK-1-\epsilon_i'^2}{NK-2}}$

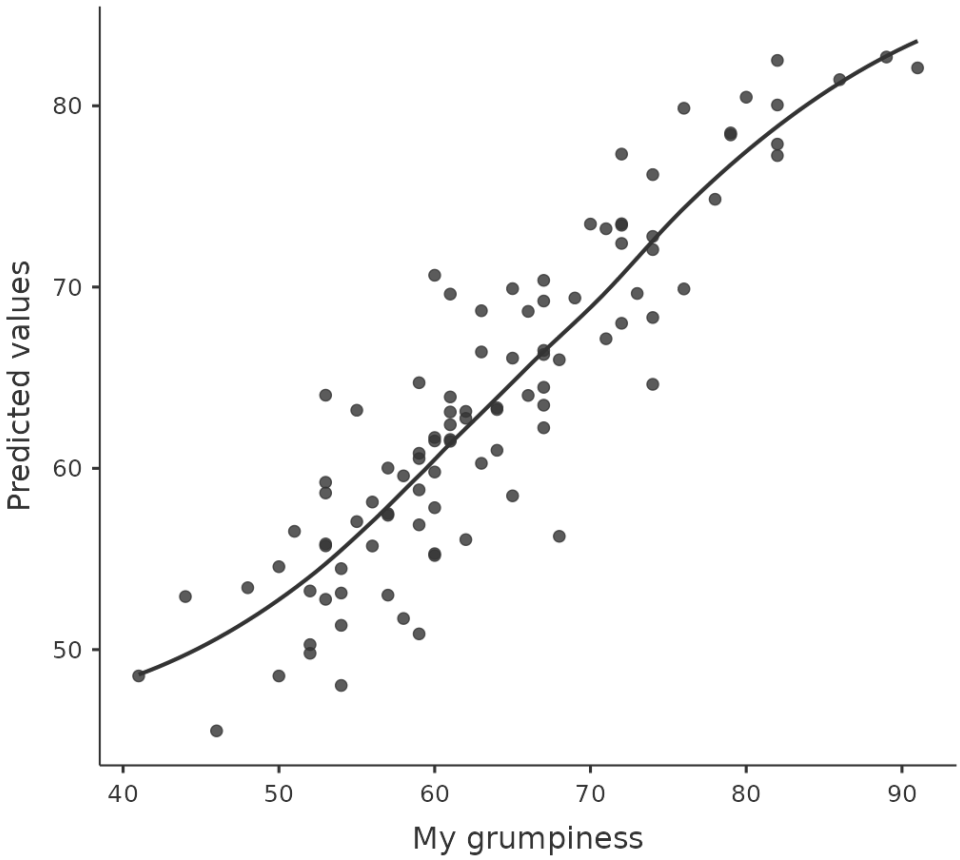


Figure 12.17: gráfico de jamovi de los valores predichos contra los valores observados de la variable de resultado. Lo que esperamos ver aquí es una línea más o menos recta. Esto se ve bastante bien, lo que sugiere que no hay nada muy mal.

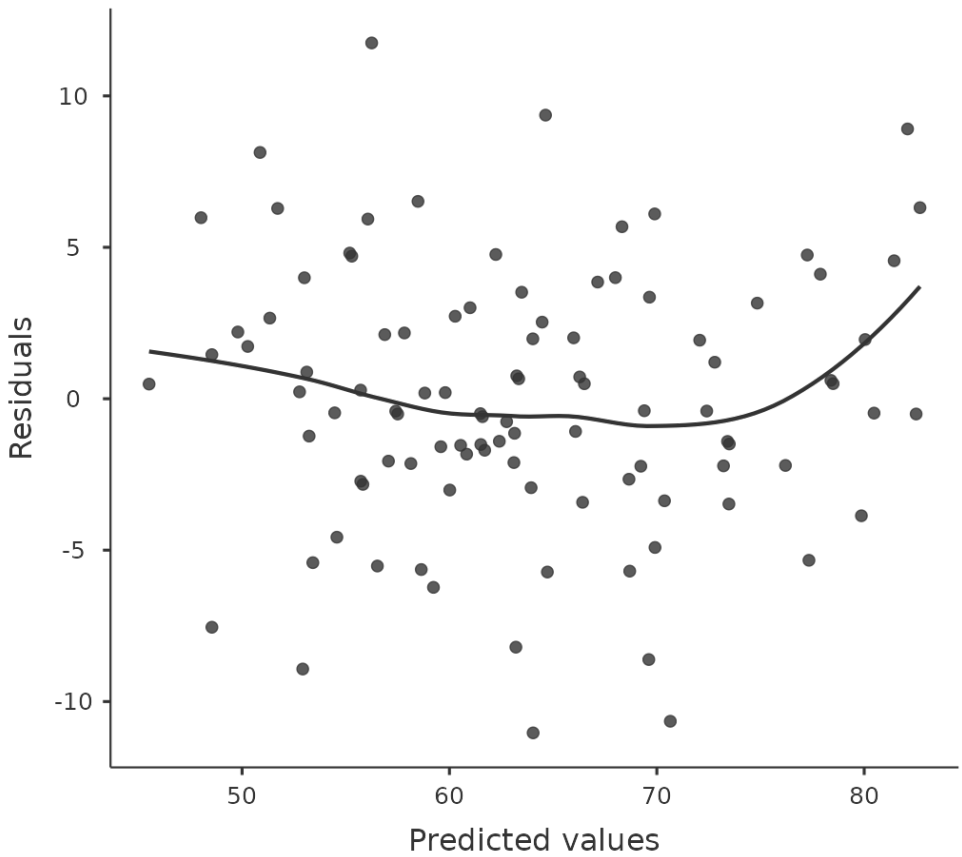


Figure 12.18: gráfico de jamovi de los valores predichos frente a los residuales, con una línea que muestra la relación entre ambos. Si la línea es horizontal y mas o menos recta, podemos sentirnos razonablemente seguros de que el "residual promedio" para todos los "valores predichos" es más o menos el mismo.

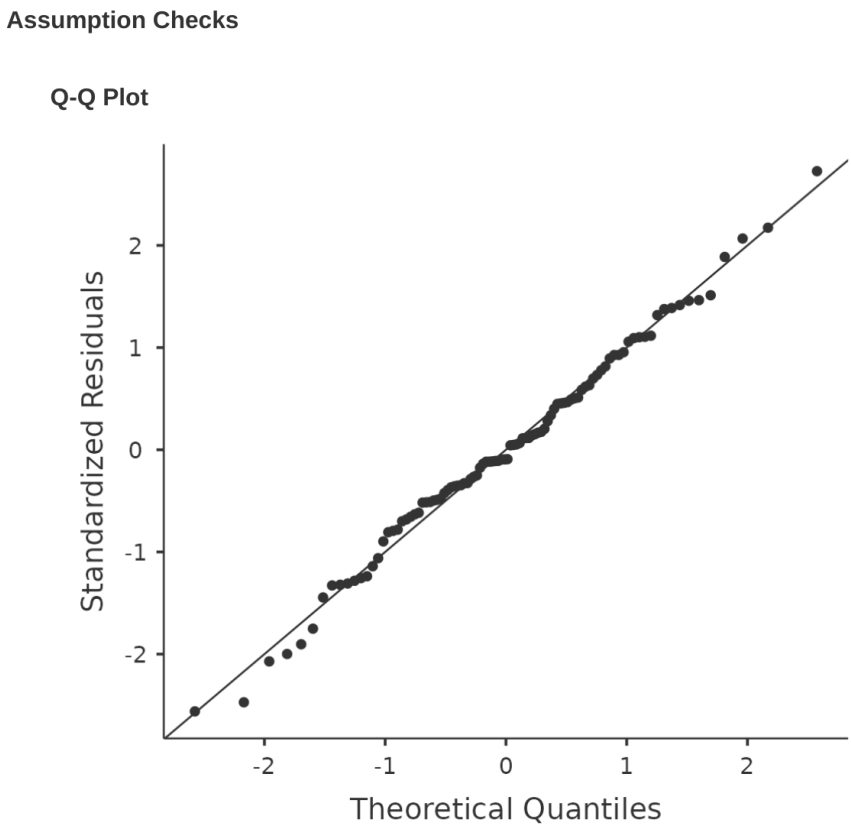


Figure 12.19: Gráfico de los cuantiles teóricos según el modelo, contra los cuantiles de los residuos estandarizados, producidos en jamovi

Otra cosa que debemos comprobar es la relación entre los valores predichos (ajustados) y los propios residuales. Podemos hacer que jamovi lo haga usando la opción 'Gráficos de residuales', que proporciona un gráfico de dispersión para cada variable predictora, la variable de resultado y los valores pronosticados frente a los residuales, ver Figure 12.20. En estos gráficos buscamos una distribución bastante uniforme de los 'puntos', sin agrupamientos ni patrones claros de los 'puntos'. Observando estos gráficos, no hay nada especialmente preocupante, ya que los puntos están distribuidos de manera bastante uniforme por todo el gráfico. Puede haber un poco de falta de uniformidad en el gráfico (b), pero no es una desviación importante y probablemente no valga la pena preocuparse.

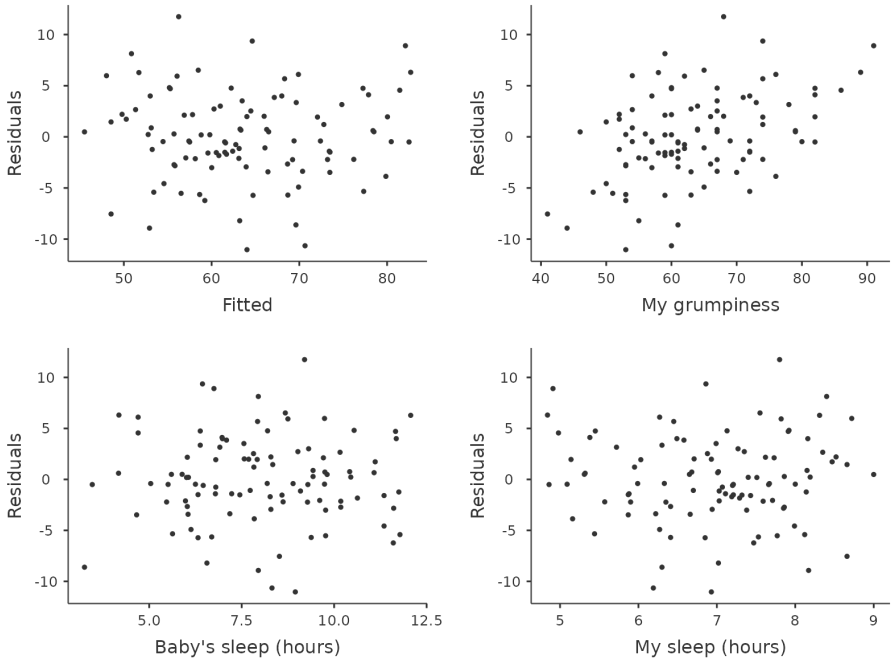


Figure 12.20: Gráficos de residuales producidos en jamovi

Si estábamos preocupadas, en muchos casos la solución a este problema (y muchos otros) es transformar una o más de las variables. Discutimos los conceptos básicos de la transformación de variables en Section 6.3, pero quiero hacer una nota especial de una posibilidad adicional que no expliqué completamente antes: la transformación de Box-Cox. La función Box-Cox es bastante simple y se usa mucho. [18]

Puedes calcularlo usando la función BOXCOX en la pantalla de variables 'Calcular' en jamovi.

12.10.4 Comprobación de la igualdad de varianzas

Todos los modelos de regresión de los que hemos hablado hacen un supuesto de igualdad (es decir, homogeneidad) de la varianza: se supone que la varianza de los residuales es constante. Para representar esto gráficamente en jamovi primero necesitamos calcular

[18] $f(x, \lambda) = \frac{x^\lambda - 1}{\lambda}$ para todos los valores de λ excepto $\lambda = 0$. Cuando $\lambda = 0$ simplemente cogemos el logaritmo natural (es decir, $\ln(x)$).

la raíz cuadrada del tamaño (absoluto) del residual[19] y luego representar esto contra los valores predichos, como en Figure 12.21. Ten en cuenta que este gráfico en realidad usa los residuales estandarizados en lugar de los brutos, pero es irrelevante desde nuestro punto de vista. Lo que queremos ver aquí es una línea recta horizontal que atraviesa el centro de la gráfica.[20]

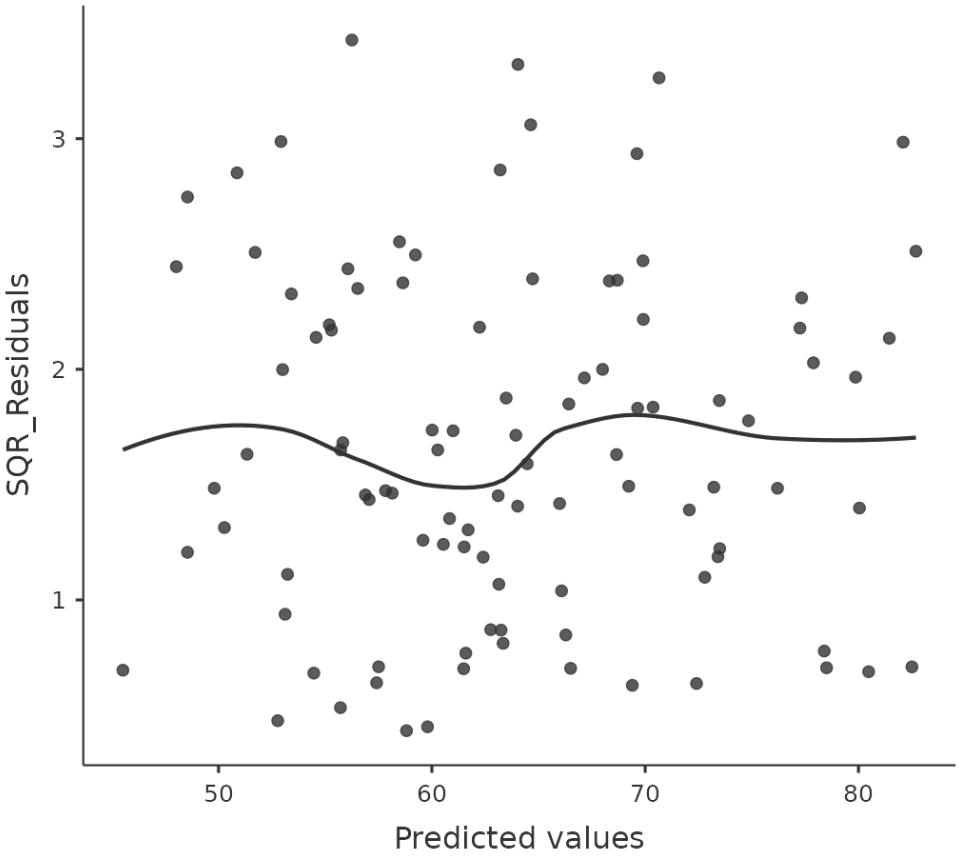


Figure 12.21: diagrama jamovi de los valores predichos (predicciones del modelo) frente a la raíz cuadrada de los residuales estandarizados absolutos. Este gráfico se utiliza para diagnosticar violaciones de la homogeneidad de varianzas. Si la varianza es realmente constante, entonces la línea que pasa por el medio debe ser horizontal y plana (más o menos).

[19] En jamovi, puedes calcular esta nueva variable usando la fórmula 'SQRT(ABS(Residuals))'

[20] Está un poco más allá del alcance de este capítulo hablar sobre cómo tratar las violaciones de la homogeneidad de varianzas, pero te daré una idea rápida de lo que debes tener en cuenta. Lo **principal** de lo que preocuparse, si se viola la homogeneidad de varianzas, es que las estimaciones del error estándar asociadas con los coeficientes de regresión ya no son completamente fiables, por lo que tus pruebas de t para los coeficientes no son del todo correctas. Una solución simple al problema es hacer uso de una "matriz de covarianza corregida por heteroscedasticidad" al estimar los errores estándar. Estos a menudo se denominan ***estimadores sándwich***, y se pueden estimar en R (pero no directamente en jamovi).

12.10.5 Comprobación de la colinealidad

Otro diagnóstico de regresión lo proporcionan los **factores de inflación de varianza** (FIV), que son útiles para determinar si los predictores en tu modelo de regresión están demasiado correlacionados entre sí o no. Hay un factor de inflación de varianza asociado con cada predictor X_k en el modelo. [21]

Si solo tienes dos predictores, los valores de FIV siempre serán los mismos, como podemos ver si hacemos clic en la casilla de verificación 'Colinealidad' en las opciones 'Regresión' - 'Supuestos' en jamovi. Tanto para dani.sleep como para baby.sleep el FIV es de 1.65. Y dado que la raíz cuadrada de 1.65 es 1.28, vemos que la correlación entre nuestros dos predictores no está causando mucho problema.

Para dar una idea de cómo podríamos terminar con un modelo que tiene mayores problemas de colinealidad, supongamos que tuvieras que ejecutar un modelo de regresión mucho menos interesante, en el que intentaras predecir el día en que se recogieron los datos, en función de todas las demás variables en el conjunto de datos. Para ver por qué esto sería un pequeño problema, echemos un vistazo a la matriz de correlación para las cuatro variables (Figure 12.22).

Correlation Matrix

Correlation Matrix

	My grumpiness	Baby's sleep (hours)	My sleep (hours)	day
My grumpiness	—			
Baby's sleep (hours)	−0.57	—		
My sleep (hours)	−0.90	0.63	—	
day	0.08	−0.01	−0.10	—

Figure 12.22: matriz de correlación en jamovi para las cuatro variables

¡Tenemos algunas correlaciones bastante grandes entre algunas de nuestras variables predictoras! Cuando ejecutamos el modelo de regresión y observamos los valores FIV, vemos que la colinealidad está causando mucha incertidumbre sobre los coeficientes. Primero, ejecuta la regresión, como en Figure 12.23 y puedes ver a partir de los valores FIV que, sí, ahí hay una colinealidad muy fina.

[21] La fórmula para el k-ésimo FIV es: $VIF_k = \frac{1}{1-R^2_{-k}}$ donde $R^2_(-k)$ se refiere al valor R-cuadrado que obtendrías si ejecutaras una regresión utilizando X_k como variable de resultado y todas las demás variables X como predictores. La idea aquí es que $R^2_(-k)$ es una muy buena medida de hasta qué punto X_k se correlaciona con todas las demás variables del modelo. Mejor aún, la raíz cuadrada del FIV es bastante interpretable: te dice cuánto más amplio es el intervalo de confianza para el coeficiente correspondiente b_k, en relación con lo que cabría esperar si todos los predictores fueran buenos y no estuvieran correlacionados entre sí.

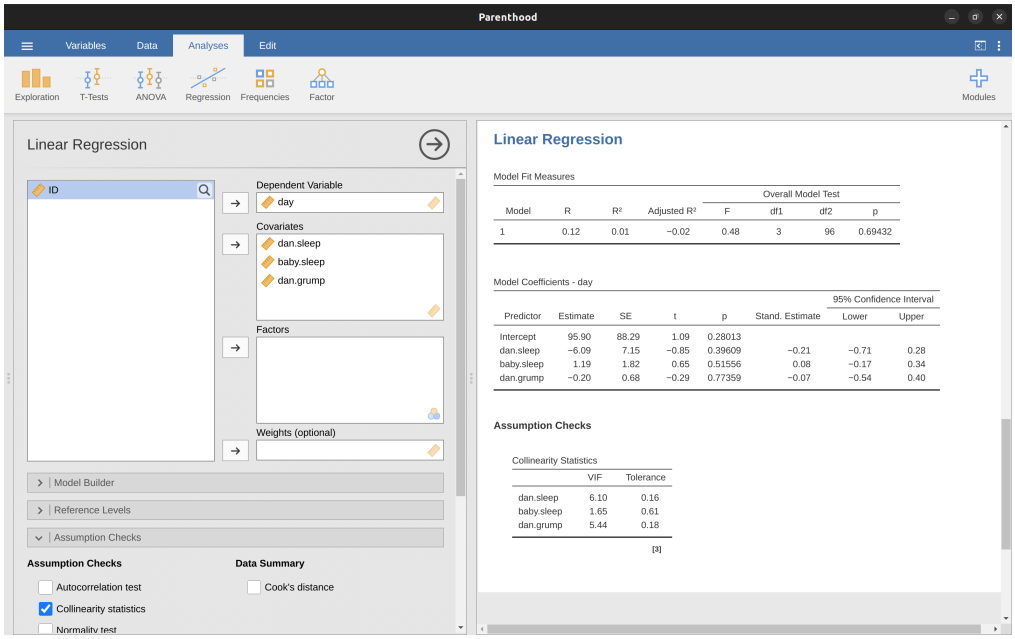


Figure 12.23: estadísticos de colinealidad para regresión múltiple, producidas en jamovi

12.10.6 Datos atípicos y anómalos

Un peligro con el que puedes encontrarte con los modelos de regresión lineal es que tu análisis puede ser desproporcionadamente sensible a un pequeño número de observaciones "inusuales" o "anómalas". Discutí esta idea anteriormente en Section 5.2.3 cuando comenté los valores atípicos que se identifican automáticamente mediante la opción de diagrama de caja en 'Exploración' - 'Descriptivos', pero esta vez necesitamos ser mucho más precisas. En el contexto de la regresión lineal, hay tres formas conceptualmente distintas en las que una observación puede llamarse "anómala". Las tres son interesantes, pero tienen implicaciones bastante diferentes para tu análisis.

El primer tipo de observación inusual es un **valor atípico**. La definición de un valor atípico (en este contexto) es una observación que es muy diferente de lo que predice el modelo de regresión. Se muestra un ejemplo en Figure 12.24. En la práctica, operacionalizamos este concepto diciendo que un valor atípico es una observación que tiene un residual muy grande, ϵ_i^*. Los valores atípicos son interesantes: un valor atípico grande puede corresponder a datos basura, por ejemplo, las variables pueden haberse registrado incorrectamente en el conjunto de datos, o puede detectarse algún otro defecto. Ten en cuenta que no debes descartar una observación solo porque es un valor atípico. Pero el hecho de que sea un caso atípico es a menudo una señal para mirar con más de talle ese caso y tratar de descubrir por qué es tan diferente.

La segunda forma en que una observación puede ser inusual es si tiene un alto **apalancamiento**, lo que sucede cuando la observación es muy diferente de todas las demás observaciones. Esto no necesariamente tiene que corresponder a un residual grande. Si la observación resulta ser inusual en todas las variables precisamente de la misma manera, en realidad puede estar muy cerca de la línea de regresión. Un ejemplo de esto

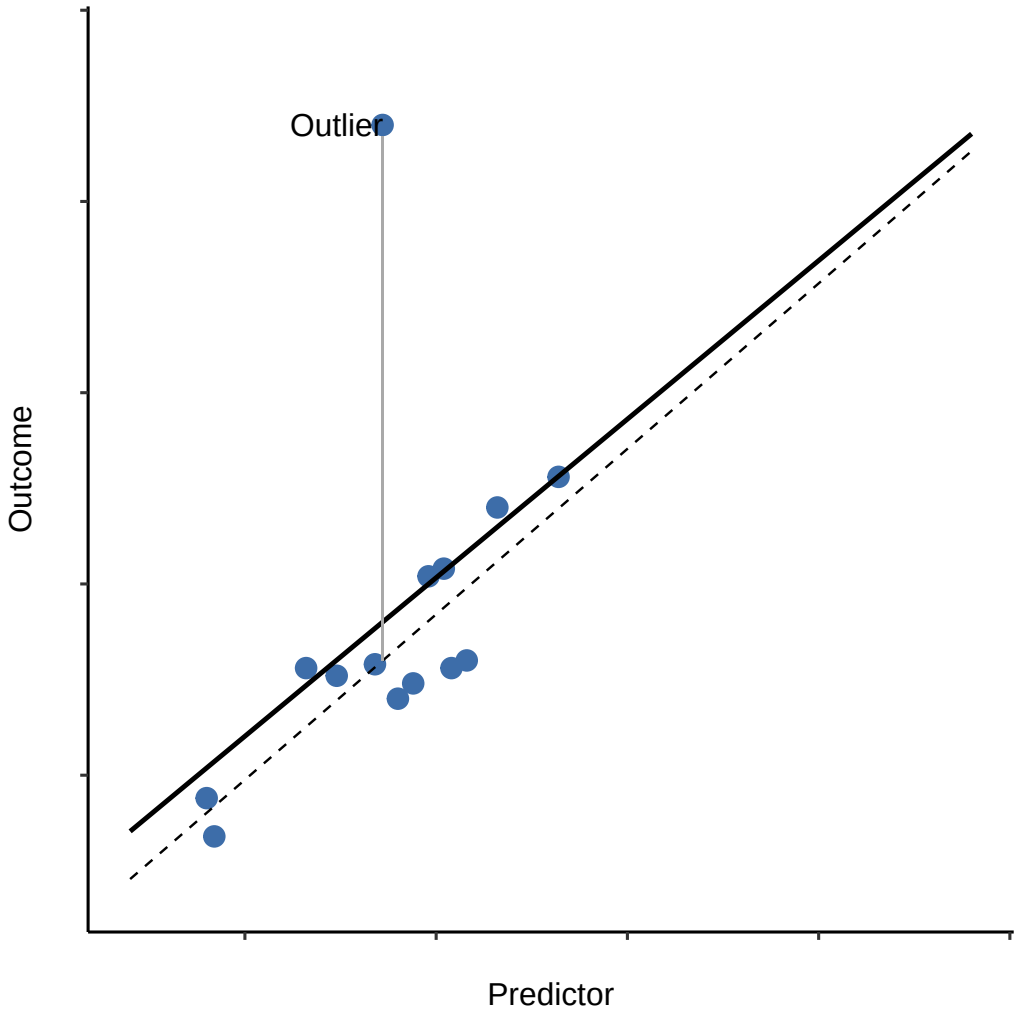


Figure 12.24: Una ilustración de valores atípicos. La línea continua muestra la línea de regresión con la observación de valores atípicos anómalos incluida. La línea discontinua traza la línea de regresión estimada sin incluir la observación anómala de valores atípicos. La línea vertical desde el punto del valor atípico hasta la línea de regresión discontinua ilustra el gran error residual del valor atípico. El valor atípico tiene un valor inusual en el resultado (ubicación del eje y) pero no en el predictor (ubicación del eje x), y se encuentra muy lejos de la línea de regresión

se muestra en Figure 12.24. El apalancamiento de una observación se operacionaliza en términos de su valor sombrero, generalmente escrito h_i. La fórmula para el valor sombrero es bastante complicada[22] pero su interpretación no lo es: h_i es una medida de hasta qué punto la i-ésima observación "controla" hacia dónde se dirige la línea de regresión.

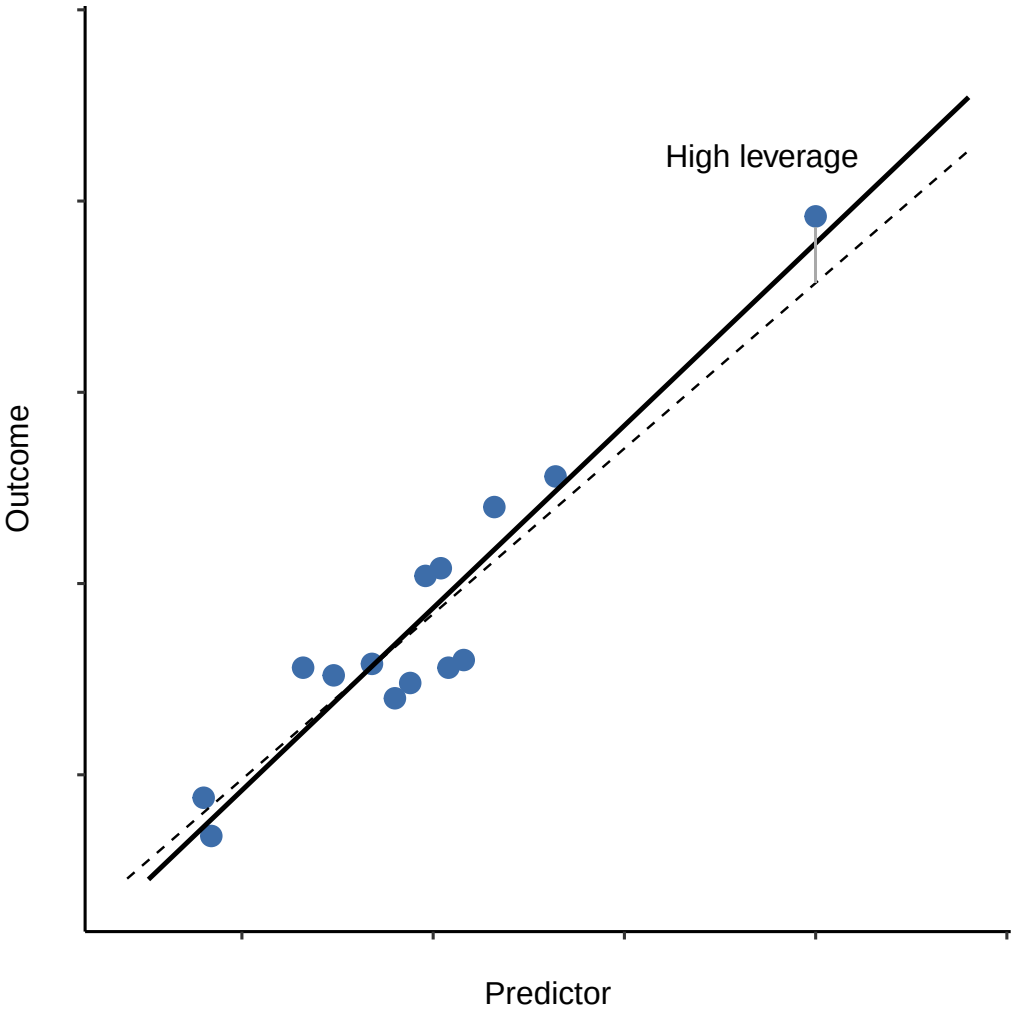


Figure 12.25: Una ilustración de puntos de alto apalancamiento. La observación anómala en este caso es inusual tanto en términos del predictor (eje x) como del resultado (eje y), pero esta inusualidad es muy consistente con el patrón de correlaciones que existe entre las otras observaciones. La observación cae muy cerca de la línea de regresión y no la distorsiona mucho.

[22] Nuevamente, para los fanáticos del álgebra lineal: la "matriz sombrero" se define como la matriz H que convierte el vector de valores observados y en un vector de valores predichos $\hat{y}$, tal que $\hat{y} = Hy$. El nombre proviene del hecho de que esta es la matriz que "le pone un sombrero a y". El valor sombrero de la i-ésima observación es el i-ésimo elemento diagonal de esta matriz (así que técnicamente deberías escribirlo como h_{ii} en lugar de h_i). Y así es como se calcula: $H = X(X^{'}X)^{1}X^{'}$.

En general, si una observación se encuentra muy alejada de las demás en términos de las variables predictoras, tendrá un valor de sombrero grande (como guía aproximada, la influencia alta es cuando el valor de sombrero es más de 2 o 3 veces el promedio; y ten en cuenta que la suma de los valores sombrero está limitada a ser igual a $K + 1$). También vale la pena analizar con más detalle los puntos de alto apalancamiento, pero es mucho menos probable que sean motivo de preocupación a menos que también sean valores atípicos.

Esto nos lleva a nuestra tercera medida de inusualidad, la **influencia** de una observación. Una observación de alta influencia es un valor atípico que tiene una alta influencia. Es decir, es una observación que es muy diferente a todas las demás en algún aspecto, y también se encuentra muy lejos de la línea de regresión. Esto se ilustra en Figure 12.26. Nota el contraste con las dos figuras anteriores. Los valores atípicos no mueven mucho la línea de regresión y tampoco los puntos de aalta influencia. Pero algo que es un valor atípico y tiene una alta influencia, bueno, eso tiene un gran efecto en la línea de regresión. Por eso llamamos a estos puntos de alta influencia, y es por eso que son la mayor preocupación. Operacionalizamos la influencia en términos de una medida conocida como **distancia de Cook**. [23]

Para tener una distancia de Cook grande, una observación debe ser un valor atípico bastante sustancial y tener una alta influencia. Como guía aproximada, la distancia de Cook superior a 1 a menudo se considera grande (eso es lo que normalmente uso como una regla rápida).

En jamovi, la información sobre la distancia de Cook se puede calcular haciendo clic en la casilla de verificación 'Distancia de Cook' en las opciones 'Comprobaciones de supuestos' - 'Resumen de datos'. Cuando haces esto, para el modelo de regresión múltiple que hemos estado usando como ejemplo en este capítulo, obtienes los resultados que se muestran en Figure 12.27.

Puedes ver que, en este ejemplo, el valor medio de la distancia de Cook es $ 0.01 $, y el rango es de $ 0.00 $ a $ 0.11 $, por lo que esto se aleja de la regla general mencionada anteriormente de que una distancia de Cook mayor que 1 se considera grande.

Una pregunta obvia para hacer a continuación es, si tienes valores grandes de distancia de Cook, ¿qué debes hacer? Como siempre, no hay una regla estricta y rápida. Probablemente, lo primero que debes hacer es intentar ejecutar la regresión con el valor atípico con la mayor distancia de Cook[24] excluido y ver qué sucede con el rendimiento del modelo y con los coeficientes de regresión. Si realmente son sustancialmente diferentes, es hora de comenzar a profundizar en tu conjunto de datos y las notas que sin duda escribías mientras realizabas tu estudio. Trata de averiguar por qué el dato es tan diferente. Si estás convencida de que este punto de datos está distorsionando gravemente sus resultados, entonces podrías considerar excluirlo, pero eso no es ideal a menos que tengas una explicación sólida de por qué este caso en particular es cualitativamente diferente de los demás y, por lo tanto, merece ser manejado por separado.

[23] $D_i = \frac{{\epsilon_i^*}^2}{K+1} \times \frac{h_i}{1-h_i}$ Observa que esto es una multiplicación de algo que mide el valor atípico de la observación (la parte de la izquierda) y algo que mide la influencia de la observación (la parte de la derecha).

[24] en jamovi, puedes guardar los valores de distancia de Cook en el conjunto de datos y luego dibujar un diagrama de caja de los valores de distancia de Cook para identificar los valores atípicos específicos. O podrías usar un programa de regresión más poderoso, como el paquete 'car' en R, que tiene más opciones para el análisis de diagnóstico de regresión avanzado.

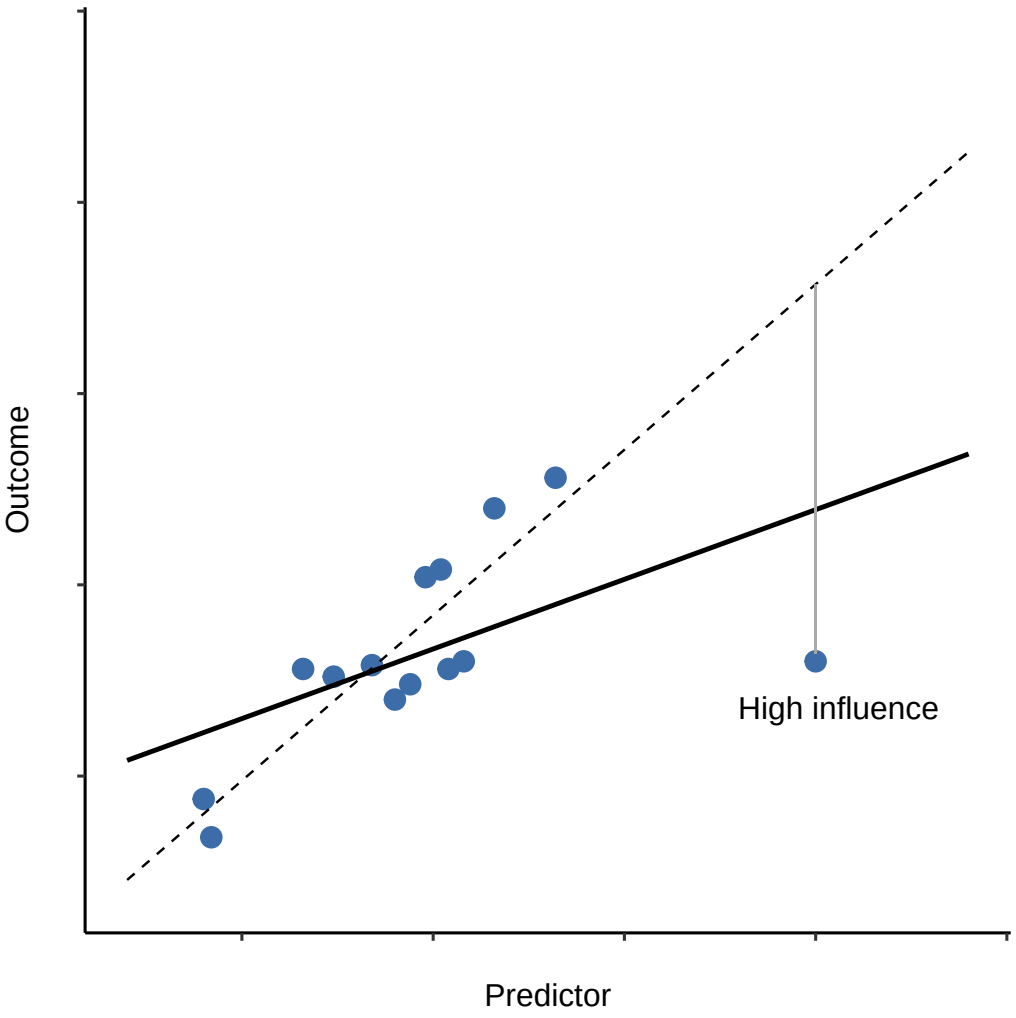


Figure 12.26: Una ilustración de puntos de alta influencia. En este caso, la observación anómala es muy inusual en la variable predictora (eje x) y se aleja mucho de la línea de regresión. Como consecuencia, la línea de regresión está muy distorsionada, aunque (en este caso) la observación anómala es completamente típica en términos de la variable de resultado (eje y)

Data Summary

Cook's Distance

			Range	
Mean	Median	SD	Min	Max
0.01	0.00	0.02	0.00	0.11

Figure 12.27: salida de jamovi que muestra la tabla de estadísticos de distancia de Cook

12.11 Selección del modelo

Un problema bastante importante que persiste es el problema de la "selección del modelo". Es decir, si tenemos un conjunto de datos que contiene varias variables, ¿cuáles debemos incluir como predictores y cuáles no? En otras palabras, tenemos un problema de **selección de variables**. En general, la selección de modelos es un asunto complejo, pero se simplifica un poco si nos restringimos al problema de elegir un subconjunto de las variables que deberían incluirse en el modelo. Sin embargo, no voy a tratar de abarcar este tema con mucho detalle. En su lugar, hablaré sobre dos principios generales en los que debes pensar y luego analizaré una herramienta concreta que proporciona jamovi para ayudarte a seleccionar un subconjunto de variables para incluir en tu modelo. En primer lugar, los dos principios:

- Es bueno tener una base sustantiva real para tus elecciones. Es decir, en muchas situaciones, tú, la investigadora, tienes buenas razones para seleccionar un pequeño número de posibles modelos de regresión que son de interés teórico. Estos modelos tendrán una interpretación sensata en el contexto de tu campo. Nunca restes importancia a esto. La estadística sirve al proceso científico, no al revés.
- En la medida en que tus elecciones se basen en la inferencia estadística, existe un equilibrio entre la simplicidad y la bondad de ajuste. A medida que agregas más predictores al modelo, lo haces más complejo. Cada predictor agrega un nuevo parámetro libre (es decir, un nuevo coeficiente de regresión), y cada nuevo parámetro aumenta la capacidad del modelo para "absorber" variaciones aleatorias. Por lo tanto, la bondad del ajuste (por ejemplo, R^2) continúa aumentando, a veces de manera trivial o por casualidad, a medida que agregas más predictores sin importar qué. Si deseas que tu modelo pueda generalizarse bien a nuevas observaciones, debes evitar incluir demasiadas variables.

Este último principio a menudo se conoce como **la navaja de Ockham** y a menudo se resume en la siguiente frase: no multipliques las entidades más allá de lo necesario. En este contexto, significa no introducir un montón de predictores en gran medida

irrelevantes solo para aumentar tu R2. Mmm. Sí, el original era mejor.

En cualquier caso, lo que necesitamos es un criterio matemático real que implemente el principio cualitativo detrás de la navaja de Ockham en el contexto de la selección de un modelo de regresión. Pues resulta que hay varias posibilidades. Del que hablaré es del **criterio de información de Akaike** (Akaike, 1974) simplemente porque está disponible como una opción en jamovi. [25]

Cuanto menor sea el valor de AIC, mejor será el rendimiento del modelo. Si ignoramos los detalles de bajo nivel, es bastante obvio lo que hace el AIC. A la izquierda tenemos un término que aumenta a medida que empeoran las predicciones del modelo; a la derecha tenemos un término que aumenta a medida que aumenta la complejidad del modelo. El mejor modelo es el que se ajusta bien a los datos (residuales bajos, lado izquierdo) usando la menor cantidad de predictores posible (K bajo, lado derecho). En resumen, esta es una implementación simple de la navaja de Ockham.

AIC se puede agregar a la tabla de resultados 'Model Fit Measures' cuando se hace clic en la casilla de verificación 'AIC', y una forma bastante torpe de evaluar diferentes modelos es ver si el valor 'AIC' es más bajo si eliminas uno o más de los predictores en el modelo de regresión. Esta es la única forma implementada actualmente en jamovi, pero existen alternativas en otros programas más potentes, como R. Estos métodos alternativos pueden automatizar el proceso de eliminar (o agregar) variables predictoras de forma selectiva para encontrar el mejor AIC. Aunque estos métodos no están implementados en jamovi, los mencionaré brevemente a continuación para que los conozcas.

12.11.1 Eliminación hacia atrás

En la eliminación hacia atrás, comienzas con el modelo de regresión completo, incluidos todos los predictores posibles. Luego, en cada "paso" probamos todas las formas posibles de eliminar una de las variables, y se acepta la que sea mejor (en términos del valor AIC más bajo). Este se convierte en nuestro nuevo modelo de regresión, y luego intentamos todas las eliminaciones posibles del nuevo modelo, eligiendo nuevamente la opción con el AIC más bajo. Este proceso continúa hasta que terminamos con un modelo que tiene un valor de AIC más bajo que cualquiera de los otros modelos posibles que podrías producir eliminando uno de tus predictores.

12.11.2 Selección hacia adelante

Como alternativa, también puedes probar la **selección hacia adelante**. Esta vez comenzamos con el modelo más pequeño posible como punto de partida y solo consideramos las posibles adiciones al modelo. Sin embargo, hay una complicación. También debes especificar cuál es el modelo más grande posible que estás dispuesta a aceptar.

Aunque la selección hacia atrás y hacia adelante pueden llevar a la misma conclusión, no siempre es así.

[25] en el contexto de un modelo de regresión lineal (¡e ignorando los términos que no dependen del modelo de ninguna manera!), el AIC para un modelo que tiene un predictor con K variables más una intersección es $AIC = \frac{SS_{res}}{\hat{\sigma}^2} + 2K$

12.11.3 Una advertencia

Los métodos automatizados de selección de variables son seductores, especialmente cuando están agrupados en funciones (bastante) simples en poderosos programas estadísticos. Brindan un elemento de objetividad a la selección de tu modelo, y eso es bueno. Desafortunadamente, a veces se usan como excusa para la desconsideración. Ya no tienes que pensar detenidamente qué predictores agregar al modelo y cuál podría ser la base teórica para su inclusión. Todo se soluciona con la magia de AIC. Y si empezamos a lanzar frases como la navaja de Ockham, parece que todo está envuelto en un pequeño y bonito paquete con el que nadie puede discutir.

O, quizás no. En primer lugar, hay muy poco acuerdo sobre lo que cuenta como un criterio de selección de modelo adecuado. Cuando me enseñaron la eliminación hacia atrás como estudiante universitaria, usamos pruebas F para hacerlo, porque ese era el método predeterminado que usaba el software. He descrito el uso de AIC, y dado que este es un texto introductorio, ese es el único método que he descrito, pero el AIC no es la Palabra de los Dioses de la Estadística. Es una aproximación, derivada bajo ciertos supuestos, y se garantiza que funcionará solo para muestras grandes cuando se cumplan esos supuestos. Modifica esos supuestos y obtendrás un criterio diferente, como el BIC, por ejemplo (también disponible en jamovi). Vuelve a adoptar un enfoque diferente y obtendrás el criterio NML. Decide que eres un bayesiano y obtienes una selección de modelo basada en razones de probabilidades posteriores. Luego hay un montón de herramientas específicas de regresión que no he mencionado. Y así sucesivamente. Todos estos métodos diferentes tienen fortalezas y debilidades, y algunos son más fáciles de calcular que otros (AIC es probablemente el más fácil de todos, lo que podría explicar su popularidad). Casi todos producen las mismas respuestas cuando la respuesta es "obvia", pero hay bastante desacuerdo cuando el problema de selección del modelo se vuelve difícil.

¿Qué significa esto en la práctica? Bueno, podrías pasar varios años aprendiendo por ti misma la teoría de la selección de modelos, aprendiendo todos los entresijos de ella para que finalmente puedas decidir qué es lo que personalmente crees que es lo correcto. Hablando como alguien que realmente hizo eso, no lo recomendaría. Probablemente saldrás aún más confundida que cuando empezaste. Una mejor estrategia es mostrar un poco de sentido común. Si estás mirando los resultados de un procedimiento de selección automatizado hacia atrás o hacia adelante, y el modelo que tiene sentido está cerca de tener el AIC más pequeño pero es derrotado por poco por un modelo que no tiene ningún sentido, entonces confía en tu instinto. La selección de modelos estadísticos es una herramienta inexacta y, como dije al principio, la interpretabilidad es importante.

12.11.4 Comparación de dos modelos de regresión

Una alternativa al uso de procedimientos automatizados de selección de modelos es que el investigador seleccione explícitamente dos o más modelos de regresión para compararlos entre sí. Puedes hacer esto de diferentes maneras, según la pregunta de investigación que estés tratando de responder. Supongamos que queremos saber si la cantidad de sueño que mi hijo durmió o no tiene alguna relación con mi mal humor, más allá de lo que podríamos esperar de la cantidad de sueño que dormí. También queremos asegurarnos de que el día en que tomamos la medida no influya en la relación. Es decir, estamos interesadas en la relación entre baby.sleep y dani.grump, y desde esa perspectiva dani.sleep y day son variables molestas o **covariables** que queremos controlar. En

esta situación, lo que nos gustaría saber es si dani.grump ~ dani.sleep + day + baby .sleep (que llamaré Modelo 2 o M2) es un mejor modelo de regresión para estos datos que dani.grump ~ dani.sleep + day (que llamaré Modelo 1 o M1). Hay dos formas diferentes en que podemos comparar estos dos modelos, una basada en un criterio de selección de modelo como AIC, y la otra basada en una prueba de hipótesis explícita. Primero te mostraré el enfoque basado en AIC porque es más simple y se deriva naturalmente de la discusión en la última sección. Lo primero que debo hacer es ejecutar las dos regresiones, anotar el AIC para cada una y luego seleccionar el modelo con el valor de AIC más pequeño, ya que se considera que es el mejor modelo para estos datos. De hecho, no lo hagas todavía. Sigue leyendo porque en jamovi hay una manera fácil de obtener los valores de AIC para diferentes modelos incluidos en una tabla.[26]

Un enfoque algo diferente del problema surge del marco de prueba de hipótesis. Supón que tienes dos modelos de regresión, donde uno de ellos (Modelo 1) contiene un subconjunto de los predictores del otro (Modelo 2). Es decir, el Modelo 2 contiene todos los predictores incluidos en el Modelo 1, además de uno o más predictores adicionales. Cuando esto sucede, decimos que el Modelo 1 está anidado dentro del Modelo 2, o posiblemente que el Modelo 1 es un submodelo del Modelo 2. Independientemente de la terminología, lo que esto significa es que podemos pensar en el Modelo 1 como una hipótesis nula y el Modelo 2 como una hipótesis alternativa. Y, de hecho, podemos construir una prueba F para esto de una manera bastante sencilla. [27]

Bien, esa es la prueba de hipótesis que usamos para comparar dos modelos de regresión entre sí. Ahora bien, ¿cómo lo hacemos en jamovi? La respuesta es usar la opción

[26] Mientras estoy en este tema, debo señalar que la evidencia empírica sugiere que BIC es un mejor criterio que AIC. En la mayoría de los estudios de simulación que he visto, BIC trabaja mucho mejor al seleccionar el modelo correcto.

[27] podemos ajustar ambos modelos a los datos y obtener una suma de cuadrados residual para ambos modelos. Los denotaré como $SS_{res}^{(1)}$ y $SS_{res}^{(2)}$ respectivamente. El superíndice aquí solo indica de qué modelo estamos hablando. Entonces nuestro estadístico F es

$$F = \frac{\frac{SS_{res}^{(1)} - SS_{res}^{(2)}}{k}}{\frac{SS_{res}^{2}}{N p - 1}}$$

donde N es el número de observaciones, p es el número de predictores en el modelo completo (sin incluir la intersección) y k es la diferencia en el número de parámetros entre los dos modelos. [d] Los grados de libertad aquí son k y $N - p - 1$. Ten en cuenta que a menudo es más conveniente pensar en la diferencia entre esos dos valores de SC como una suma de cuadrados en sí. Eso es

$$SS_{\Delta} = SS_{res}^{(1)} - SS_{res}^{(2)}$$

La razón por la que esto es útil es que podemos expresar SS_{Δ} como una medida de hasta qué punto los dos modelos hacen diferentes predicciones sobre la variable de resultado. Específicamente,

$$SS_{\Delta} = \sum_{i} (\hat{y}_i^{(2)} - \hat{y}_i^{(1)})^2$$

donde $\hat{y}_{i^{(1)}}$ es el valor previsto para y_i según el modelo M_1 y $\hat{y}_{i^{(2)}}$ es el valor previsto para $y_i $ según modelo M_2. — [d] Vale la pena señalar de paso que este mismo estadístico F se puede usar para probar una gama mucho más amplia de hipótesis que las que estoy mencionando aquí. Muy brevemente, observa que el modelo anidado M1 corresponde al modelo completo M2 cuando restringimos algunos de los coeficientes de regresión a cero. A veces es útil construir submodelos colocando otros tipos de restricciones en los coeficientes de regresión. Por ejemplo, quizás dos coeficientes diferentes tengan que sumar cero. También puedes construir pruebas de hipótesis para ese tipo de restricciones, pero es un poco más complicado y la distribución muestral de F puede terminar siendo algo conocido como distribución F no central, que está mucho más allá del alcance de este libro. Todo lo que quiero hacer es alertarte de esta posibilidad.

'Model Builder' y especificar los predictores del Modelo 1 dani.sleep y day en el 'Bloque 1' y luego agregar el predictor adicional del Modelo 2 (baby.sleep) en el 'Bloque 2', como en Figure 12.27. Esto muestra, en la tabla de 'Comparaciones de modelos', que para las comparaciones entre el Modelo 1 y el Modelo 2, $F(1, 96) = 0.00$, $p = 0.954$. Como tenemos p > .05 mantenemos la hipótesis nula (M1). Este enfoque de regresión, en el que agregamos todas nuestras covariables en un modelo nulo, luego agregamos las variables de interés en un modelo alternativo y luego comparamos los dos modelos en un marco de prueba de hipótesis, a menudo se denomina **regresión jerárquica** .

También podemos usar esta opción de 'Comparación de modelos' para construir una tabla que muestra el AIC y BIC para cada modelo, lo que facilita la comparación e identificación de qué modelo tiene el valor más bajo, como en Figure 12.28.

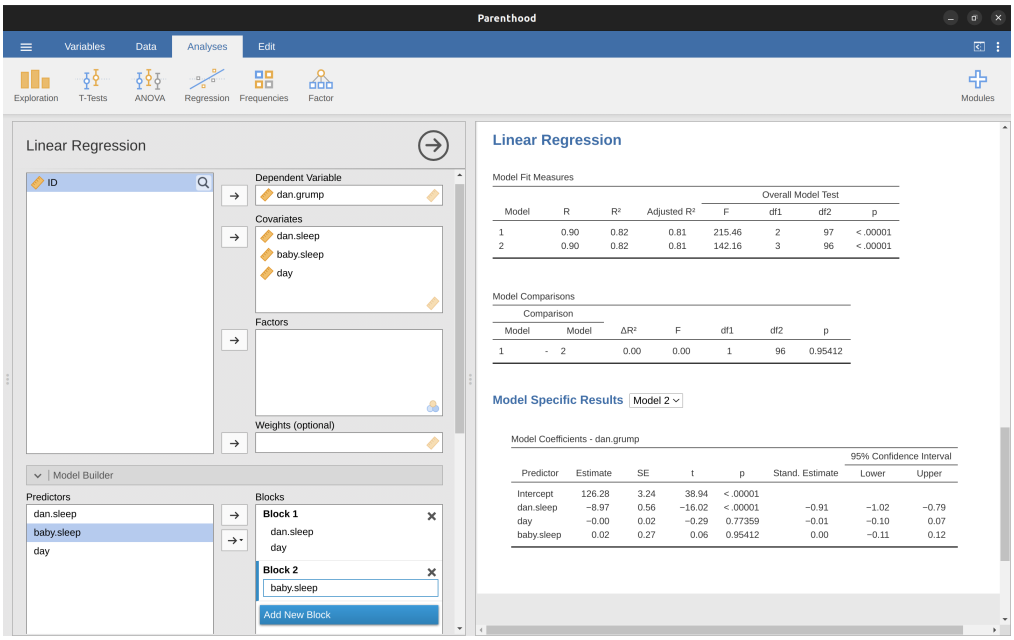


Figure 12.28: Comparación de modelos en jamovi usando la opción 'Creador de modelos'

12.12 Resumen

- ¿Quieres saber cómo de fuerte es la relación entre dos variables? Calcular correlaciones
- Dibujo [diagramas de dispersión]
- Ideas básicas sobre ¿Qué es un modelo de regresión lineal? y Estimación de un modelo de regresión lineal
- Regresión lineal múltiple
- Cuantificando el ajuste del modelo de regresión usando R^2.
- Pruebas de hipótesis para modelos de regresión
- En Sobre los coeficientes de regresión hablamos sobre calcular Intervalos de confianza para los coeficientes y Cálculo de coeficientes de regresión estandarizados
- Los Supuestos de regresión y Comprobación del modelo (#sec-Model-checking)

- Regresión [Selección de modelo]

Chapter 13

Comparación de varias medias (ANOVA unidireccional)

Este capítulo presenta una de las herramientas más utilizadas en estadística psicológica, conocida como "análisis de la varianza", pero generalmente denominada ANOVA. La técnica básica fue desarrollada por Sir Ronald Fisher a principios del siglo XX y es a él a quien le debemos la terminología bastante desafortunada. El término ANOVA es un poco engañoso, en dos aspectos. En primer lugar, aunque el nombre de la técnica se refiere a las varianzas, ANOVA se ocupa de investigar las diferencias en las medias. En segundo lugar, hay diferentes cosas que se conocen como ANOVA, algunas de las cuales tienen poca relación. Más adelante en el libro, encontraremos diferentes métodos ANOVA que se aplican en situaciones bastante diferentes, pero para los propósitos de este capítulo solo consideraremos la forma más simple de ANOVA, en la que tenemos varios grupos diferentes de observaciones, y nos interesa averiguar si esos grupos difieren en términos de alguna variable de resultado de interés. Esta es la pregunta que se aborda mediante un ANOVA unifactorial.

La estructura de este capítulo es la siguiente: primero presentaré un conjunto de datos ficticios que usaremos como ejemplo a lo largo del capítulo. Después de presentar los datos, describiré la mecánica de cómo funciona realmente un ANOVA unifactorial Cómo funciona ANOVA y luego me centraré en cómo puedes ejecutar uno en jamovi [Ejecutar un ANOVA en jamovi]. Estas dos secciones son el núcleo del capítulo.

El resto del capítulo analiza algunos temas importantes que inevitablemente surgen cuando se ejecuta un ANOVA, a saber, cómo calcular los tamaños del efecto, las pruebas post hoc y las correcciones para comparaciones múltiples y los supuestos en las que se basa el ANOVA. También hablaremos sobre cómo verificar esos supuestos y algunas de las cosas que puedes hacer si se violan los supuestos. Luego hablaremos de ANOVA de medidas repetidas.

13.1 Un conjunto de datos ilustrativos

Imagina que llevas a cabo un ensayo clínico en el que estás probando un nuevo fármaco antidepresivo llamado *Joyzepam*. Con el fin de construir una prueba justa de la eficacia

del fármaco, el estudio implica la administración de tres fármacos separados. Uno es un placebo y el otro es un medicamento antidepresivo/ansiolítico existente llamado *Anxifree*. Se reclutan 18 participantes con depresión moderada a severa para la prueba inicial. Debido a que los fármacos a veces se administran junto con la terapia psicológica, tu estudio incluye a 9 personas que se someten a terapia cognitiva conductual (TCC) y 9 que no la reciben. A los participantes se les asigna aleatoriamente (doble ciego, por supuesto) un tratamiento, de modo que haya 3 personas con TCC y 3 personas sin terapia asignadas a cada uno de los 3 medicamentos. Un psicólogo evalúa el estado de ánimo de cada persona después de 3 meses de tratamiento con cada medicamento, y la mejora general en el estado de ánimo de cada persona se evalúa en una escala que va de -5 a $+5$. Con ese diseño del estudio, ahora carguemos el archivo de datos en Clinicaltrial.csv. Podemos ver que este conjunto de datos contiene las tres variables fármaco, terapia y humor.ganancia.

Para los objetivos de este capítulo, lo que realmente nos interesa es el efecto de los fármacos sobre el estado de ánimo. Lo primero que debes hacer es calcular algunos estadísticos descriptivos y dibujar algunos gráficos. En el capítulo Chapter 4 te mostramos cómo hacer esto, y algunos de los estadísticos descriptivos que podemos calcular en jamovi se muestran en Figure 13.1

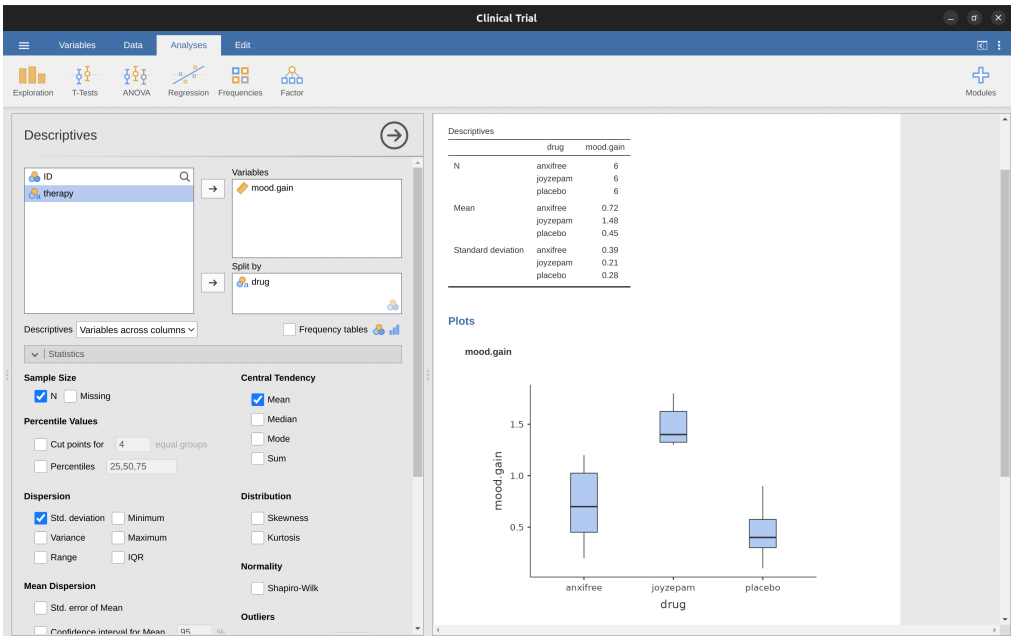


Figure 13.1: **?(caption)**

Como el gráfico muestra, hay una mayor mejora en el estado de ánimo de los participantes en el grupo de Joyzepam que en el grupo de Anxifree o en el grupo de placebo. El grupo Anxifree muestra una mayor mejora del estado de ánimo que el grupo de control, pero la diferencia no es tan grande. La pregunta que queremos responder es si estas diferencias son "reales" o solo se deben al azar.

13.2 Cómo funciona ANOVA

Para responder a la pregunta planteada por los datos de nuestro ensayo clínico, vamos a ejecutar un ANOVA unifactorial. Comenzaré mostrándote cómo hacerlo de la manera difícil, construyendo la herramienta estadística desde cero y mostrándote cómo podrías hacerlo si no tuvieras acceso a ninguna de las geniales funciones de ANOVA integradas en jamovi. Y espero que lo leas atentamente, intenta hacerlo de la manera larga una o dos veces para asegurarte de que realmente comprendes cómo funciona ANOVA, y luego, una vez que hayas comprendido el concepto, nunca vuelvas a hacerlo de esta manera.

El diseño experimental que describí en la sección anterior sugiere que nos interesa comparar el cambio de estado de ánimo promedio para los tres fármacos diferentes. En ese sentido, estamos hablando de un análisis similar a la prueba t (ver Chapter 11) pero involucrando a más de dos grupos. Si hacemos que μ_P denote la media de la población para el cambio de estado de ánimo inducido por el placebo, y que μ_A y μ_J denote las medias correspondientes para nuestros dos fármacos, Anxifree y Joyzepam, entonces la (algo pesimista) hipótesis nula que queremos probar es que las medias de las tres poblaciones son idénticas. Es decir, ninguno de los dos fármacos es más efectivo que un placebo. Podemos escribir esta hipótesis nula como:

$$H_0 : \text{ es cierto que } \mu_P = \mu_A = \mu_J$$

Como consecuencia, nuestra hipótesis alternativa es que al menos uno de los tres tratamientos es diferente de los demás. Es un poco complicado escribir esto matemáticamente, porque (como veremos) hay bastantes maneras diferentes en las que la hipótesis nula puede ser falsa. Así que por ahora escribiremos la hipótesis alternativa así:

$$H_1 : \text{ eso } \underline{\textit{es no}} \text{ es cierto que } \mu_P = \mu_A = \mu_J$$

Esta hipótesis nula es mucho más difícil de probar que cualquiera de las que hemos visto anteriormente. ¿Cómo lo haremos? Una forma sensata sería "hacer un ANOVA", ya que ese es el título del capítulo, pero no está particularmente claro por qué un "análisis de varianzas" nos ayudará a aprender algo útil sobre las medias. De hecho, esta es una de las mayores dificultades conceptuales que tienen las personas cuando se encuentran por primera vez con ANOVA. Para ver cómo funciona, me parece más útil comenzar hablando de variancias, específicamente variabilidad entregrupo y variabilidad intragrupo (Figure 13.2).

13.2.1 Dos fórmulas para la varianza de Y

En primer lugar, comencemos introduciendo algo de notación. Usaremos G para referirnos al número total de grupos. Para nuestro conjunto de datos hay tres fármacos, por lo que hay $G = 3$ grupos. A continuación, usaremos N para referirnos al tamaño total de la muestra; hay un total de $N = 18$ personas en nuestro conjunto de datos. De manera similar, usemos N_k para indicar el número de personas en el k-ésimo grupo. En nuestro ensayo clínico falso, el tamaño de la muestra es $N_k = 6$ para los tres grupos.[1]

[1] cuando todos los grupos tienen el mismo número de observaciones, se dice que el diseño experimental está "equilibrado". El equilibrio no es un gran problema para ANOVA unifactorial, que es el tema de este capítulo. Es más importante cuando haces ANOVA más complicados.

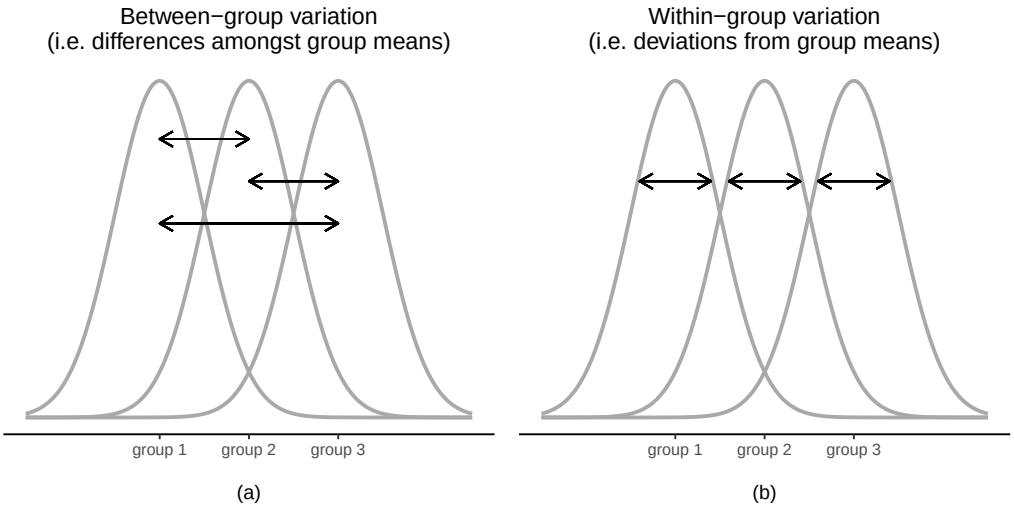


Figure 13.2: Ilustración gráfica de la variación 'entre grupos' (panel (a)) y la variación 'intra grupos' (panel (b)). A la izquierda, las flechas muestran las diferencias en las medias de los grupos. A la derecha, las flechas resaltan la variabilidad dentro de cada grupo.

Finalmente, usaremos Y para indicar la variable de resultado. En nuestro caso, Y se refiere al cambio de estado de ánimo. Específicamente, usaremos Yik para referirnos al cambio de estado de ánimo experimentado por el i-ésimo miembro del k-ésimo grupo. De manera similar, usaremos $\bar{Y}$ para el cambio de estado de ánimo promedio, recogido entre las 18 personas en el experimento, y $\bar{Y}_k$ para referirnos al cambio de estado de ánimo promedio experimentado por las 6 personas en el grupo k.

Ahora que hemos resuelto nuestra notación, podemos comenzar a escribir fórmulas. Para empezar, recordemos la fórmula para la varianza que usamos en Section 4.2, en aquellos días más amables cuando solo hacíamos estadística descriptiva. La varianza muestral de Y se define de la siguiente manera

$$Var(Y) = \frac{1}{N} \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Esta fórmula parece bastante idéntica a la fórmula para la varianza en Section 4.2. La única diferencia es que ahora tengo dos sumas aquí: estoy sumando entre grupos (es decir, valores para k) y las personas dentro de los grupos (es decir, valores para i). Esto es puramente un detalle cosmético. Si, en cambio, hubiera usado la notación Y_p para referirme al valor de la variable de resultado para la persona p en la muestra, tendría una sola suma. La única razón por la que tenemos una suma doble aquí es porque clasifiqué a las personas en grupos y luego asigné números a las personas dentro de los grupos.

Un ejemplo concreto podría sernos útil. Consideremos Table 13.1, en el que tenemos un total de $N = 5$ personas clasificadas en $G = 2$ grupos. Arbitrariamente, digamos que las personas "geniales" son el grupo 1 y las personas "no geniales" son el grupo 2. Resulta que tenemos tres personas geniales ($N_1 = 3$) y dos personas no geniales ($N_2 = 2$)

Table 13.1: mal humor en personas en grupos geniales y no geniales

name	person P	group	group num. k	index in group	grumpiness Y_{ik} or Y_p
Ann	1	cool	1	1	20
Ben	2	cool	1	2	55
Cat	3	cool	1	3	21
Tim	4	uncool	2	1	91
Egg	5	uncool	2	2	22

Ten en cuenta que he construido dos esquemas de etiquetado diferentes aquí. Tenemos una variable de "persona" p, por lo que sería perfectamente sensato referirse a Yp como el mal humor de la p-ésima persona en la muestra. Por ejemplo, la tabla muestra que Tim es el cuarto, entonces diríamos $p = 4$. Así, cuando hablamos del mal humor Y de esta persona "Tim", quienquiera que sea, podríamos referirnos a su mal humor diciendo que $Y_p = 91$, para la persona $p = 4$. Sin embargo, esa no es la única forma en que podemos referirnos a Tim. Como alternativa, podemos señalar que Tim pertenece al grupo "no geniales" ($k = 2$) y, de hecho, es la primera persona que figura en el grupo no geniales ($i = 1$). Así que es igualmente válido referirse al mal humor de Tim diciendo que $Y_{ik} = 91$, donde $k = 2$ y $i = 1$.

En otras palabras, cada persona p corresponde a una única combinación ik, por lo que la fórmula que di arriba es en realidad idéntica a nuestra fórmula original para la varianza, que sería

$$Var(Y) = \frac{1}{N}\sum_{p=1}^{N}(Y_p - \bar{Y})^2$$

En ambas fórmulas, lo único que hacemos es sumar todas las observaciones de la muestra. La mayoría de las veces solo usaríamos la notación Yp más simple; la ecuación que usa Y_p es claramente la más simple de las dos. Sin embargo, al hacer un ANOVA es importante hacer un seguimiento de qué participantes pertenecen a qué grupos, y necesitamos usar la notación Yik para hacer esto.

13.2.2 De varianzas a sumas de cuadrados

Bien, ahora que sabemos cómo se calcula la varianza, definamos algo llamado **suma de cuadrados total** , que se denota como SCtot. Es muy simple. En lugar de promediar las desviaciones al cuadrado, que es lo que hacemos cuando calculamos la varianza, simplemente las sumamos.[2]

Cuando hablamos de analizar las varianzas en el contexto de ANOVA, lo que realmente estamos haciendo es trabajar con las sumas de cuadrados totales en lugar de la varianza

[2]Por lo tanto, la fórmula para la suma de cuadrados total es casi idéntica a la fórmula para la varianza

$$SS_{tot} = \sum_{k=1}^{G}\sum_{i=1}^{N_k}(Y_{ik} - \bar{Y})^2$$

real. [3]

A continuación, podemos definir una tercera noción de variación que recoge solo las diferencias entre grupos. Hacemos esto observando las diferencias entre las medias de grupo $\bar{Y}_k$ y la media total $\bar{Y}$. [4]

No es difícil mostrar que la variación total entre las personas en el experimento (SS_{tot} es en realidad la suma de las diferencias entre los grupos SS_b y la variación dentro de los grupos SS_w. Es decir,

$$SS_w + SS_b = SS_{tot}$$

Sí.

Bien, entonces, ¿qué hemos descubierto? Hemos descubierto que la variabilidad total asociada con la variable de resultado (SS_{tot}) se puede dividir matemáticamente en la suma de "la variación debida a las diferencias en las medias de la muestra para los diferentes grupos" (SS_b) más "todo el resto de la variación" (SS_w) [5].

¿Cómo me ayuda eso a averiguar si los grupos tienen diferentes medias poblacionales? Um. Espera. Espera un segundo. Ahora que lo pienso, esto es exactamente lo que estábamos buscando. Si la hipótesis nula es verdadera, esperaría que todas las medias de la muestra fueran bastante similares entre sí, ¿verdad? Y eso implicaría que esperaría que el valor de SS_b fuera realmente pequeño, o al menos esperaría que fuera mucho más pequeño que "la variación asociada con todo lo demás", SS_w. Mmm. Detecto que se acerca una prueba de hipótesis.

13.2.3 De sumas de cuadrados a la prueba F

Como vimos en la última sección, la idea detrás de ANOVA es comparar los valores de dos sumas de cuadrados SS_b y SS_w entre sí. Si la variación entre grupos SS_b es grande en relación con la variación dentro del grupo SS_w, entonces tenemos razones para sospechar que las medias poblacionales para los diferentes grupos no son idénticas entre sí. Para convertir esto en una prueba de hipótesis viable, se necesita "jugar" un

[3] Una cosa muy positiva acerca de la suma de cuadrados total es que podemos dividirla en dos tipos diferentes de variación. Primero, podemos hablar de la suma de cuadrados intragrupo, en la que buscamos ver qué tan diferente es cada persona individual de su propia media de grupo

$$SS_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

donde $\bar{Y}_k$ es la media del grupo. En nuestro ejemplo, $\bar{Y}_k$ sería el cambio de estado de ánimo promedio experimentado por aquellas personas que recibieron el k-ésimo fármaco. Así, en lugar de comparar individuos con el promedio de todas las personas en el experimento, solo los estamos comparando con aquellas personas del mismo grupo. Como consecuencia, esperaríamos que el valor de SS_w fuera menor que la suma de cuadrados total, porque ignora por completo las diferencias de grupo, es decir, si los fármacos tendrán efectos diferentes en el estado de ánimo de las personas.

[4] para cuantificar el alcance de esta variación, lo que hacemos es calcular la suma de cuadrados entre grupos

$$\begin{aligned} SS_b &= \sum_{k=1}^{G} \sum_{i=1}^{N_k} (\bar{Y}_k - \bar{Y})^2 \\ &= sum_{k=1}^{G} N_k (\bar{Y}_k - \bar{Y})^2 \end{aligned}$$

[5] SS_w también se conoce en un ANOVA independiente como la varianza del error, o SS_{error}

poco. Lo que haré será mostrarte primero lo que hacemos para calcular nuestra prueba estadística, la **razón de F**, y luego trataré de darte una idea de por qué lo hacemos de esta manera.

Para convertir nuestros valores SC en una razón de F, lo primero que debemos calcular son los **grados de libertad** asociados con los valores SS_b y SS_w. Como es habitual, los grados de libertad corresponden al número de "datos" únicos que contribuyen a un cálculo particular, menos el número de "restricciones" que deben satisfacer. Para la variabilidad dentro de los grupos, lo que estamos calculando es la variación de las observaciones individuales (N datos) alrededor de las medias del grupo (G restricciones). Por el contrario, para la variabilidad entre grupos, nos interesa la variación de las medias de los grupos (datos G) alrededor de la media total (restricción 1). Por lo tanto, los grados de libertad aquí son:

$$df_b = G - 1$$
$$df_w = NG$$

Bueno, eso parece bastante simple. Lo que hacemos a continuación es convertir nuestro valor de sumas de cuadrados en un valor de "medias cuadráticas", lo que hacemos dividiendo por los grados de libertad:

$$MS_b = \frac{SS_b}{df_b}$$
$$MS_w = \frac{SS_w}{df_w}$$

Finalmente, calculamos la razón F dividiendo la MC entre grupos por la MC intra grupos:

$$F = \frac{MS_b}{MS_w}$$

A un nivel muy general, la explicación del estadístico F es sencilla. Los valores más grandes de F significan que la variación entre grupos es grande en relación con la variación dentro de los grupos. Como consecuencia, cuanto mayor sea el valor de F, más evidencia tendremos en contra de la hipótesis nula. Pero, ¿qué tamaño tiene que tener F para rechazar realmente H_0? Para comprender esto, necesitas una comprensión un poco más profunda de qué es ANOVA y cuáles son realmente los valores de las medias cuadráticas.

La siguiente sección trata eso con un poco de detalle, pero para quien no tenga interés en los detalles de lo que realmente mide la prueba, iré al grano. Para completar nuestra prueba de hipótesis, necesitamos conocer la distribución muestral de F si la hipótesis nula es verdadera. No es sorprendente que la distribución muestral para el estadístico F bajo la hipótesis nula sea una distribución F. Si recuerdas nuestra discusión sobre la distribución F en Chapter 7, la distribución F tiene dos parámetros, correspondientes a los dos grados de libertad involucrados. El primero df_1 son los grados de libertad entre grupos df_b, y el segundo df_2 son los grados de libertad intra grupos df_w.

Table 13.2: todas las cantidades clave involucradas en un ANOVA organizadas en una tabla ANOVA 'estándar'. Se muestran las fórmulas para todas las cantidades (excepto el valor p que tiene una fórmula muy fea y sería terriblemente difícil de calcular sin una computadora)

	between groups	**within groups**
df	$df_b = G - 1$	$df_w = N - G$
sum of squares	$SS_b = \sum_{k=1}^{G} N_k(\bar{Y}_k - \bar{Y})^2$	$SS_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k}(Y_{ik} - \bar{Y}_k)^2$
mean squares	$MS_b = \frac{SS_b}{df_b}$	$MS_w = \frac{SS_w}{df_w}$
F-statistic	$F = \frac{MS_b}{df_b}$	-
p-value	[complicated]	-

En Table 13.2 se muestra un resumen de todas las cantidades clave involucradas en un ANOVA unifactorial, incluidas las fórmulas que muestran cómo se calculan.

[Detalle técnico adicional [6]]

[6]En un nivel básico, ANOVA es una competición entre dos modelos estadísticos diferentes, H_0 y H_1. Cuando describí las hipótesis nula y alternativa al comienzo de la sección, fui un poco imprecisa acerca de cuáles son realmente estos modelos. Arreglaré eso ahora, aunque probablemente no te agradaré por hacerlo. Si recuerdas, nuestra hipótesis nula era que todas las medias de los grupos son idénticas entre sí. Si es así, entonces una forma natural de pensar en la variable de resultado Y_{ik} es describir las puntuaciones individuales en términos de una sola media poblacional μ, más la desviación de esa media poblacional. Esta desviación generalmente se denota ϵ_{ik} y tradicionalmente se le llama el error o residual asociado con esa observación. PEro ten cuidado. Tal como vimos con la palabra "significativo", la palabra "error" tiene un significado técnico en estadística que no es exactamente igual a su definición cotidiana en español. En el lenguaje cotidiano, "error" implica un error de algún tipo, pero en estadística no (o al menos, no necesariamente). Con eso en mente, la palabra "residual" es un término mejor que la palabra "error". En estadística, ambas palabras significan "variabilidad sobrante", es decir, "cosas" que el modelo no puede explicar. En cualquier caso, así es como se ve la hipótesis nula cuando la escribimos como un modelo estadístico

$$Y_{ik} = \mu + \epsilon_{ik}$$

donde asumimos (discutido más adelante) que los valores residuales ϵ_{ik} se distribuyen normalmente, con media 0 y una desviación estándar σ que es igual para todos los grupos. Para usar la notación que presentamos en Chapter 7, escribiríamos esta suposición así

$$\epsilon_{ik} \sum Normal(0, \sigma^2)$$

¿Qué pasa con la hipótesis alternativa, H_1? ? La única diferencia entre la hipótesis nula y la hipótesis alternativa es que permitimos que cada grupo tenga una media poblacional diferente. Así, si dejamos que μ_k denote la media de la población para el k-ésimo grupo en nuestro experimento, entonces el modelo estadístico correspondiente a H_1 es

$$Y_{ik} = \mu_k + \epsilon_{ik}$$

donde, una vez más, asumimos que los términos de error se distribuyen normalmente con media 0 y desviación estándar σ. Es decir, la hipótesis alternativa también asume que $\epsilon \sim Normal(0, \sigma^2)$ Bueno, una vez que hemos descrito los modelos estadísticos que sustentan H_0 y H_1 con más detalle, ahora es bastante sencillo decir qué miden los valores de las medias cuadráticas y qué significa esto para la interpretación de F. No te aburriré con la justificación de esto, pero resulta que la media cuadrática intra grupos, MS_w, puede verse como un estimador de la varianza del error σ^2 . La media cuadrática entre grupos MS_b también es un estimador, pero lo que estima es la varianza del error más una cantidad que depende de las verdaderas diferencias entre las medias de los grupos. Si llamamos a esta cantidad Q, podemos ver que el estadístico F es básicamente [a]

$$F = \frac{\hat{Q} + \hat{\sigma}^2}{\hat{\sigma}^2}$$

donde el valor verdadero $Q = 0$ si la hipótesis nula es verdadera, y Q < 0 si la hipótesis alternativa es verdadera (p. ej., Hays (1994), cap. 10). Por lo tanto, como mínimo, el *valor F debe ser mayor que* 1 para tener alguna posibilidad de rechazar la hipótesis nula. Ten en cuenta que esto no significa que sea imposible obtener un valor F menor que 1. Lo que significa es que si la hipótesis nula es verdadera, la distribución muestral de la razón F tiene una media de 1,[ˆb] por lo que necesitamos ver valores F mayores que 1 para rechazar con seguridad el valor nulo. Para ser un poco más precisas sobre la distribución muestral, observa que si la hipótesis nula es verdadera, tanto la media cuadrática entre grupos como la media cuadrática intra grupos son estimadores de la varianza de los residuales ϵ_{ik}. Si esos residuales se distribuyen normalmente, entonces podrías sospechar que la estimación de la varianza de ϵ_{ik} tiene una distribución de ji cuadrado, porque (como se discutió en Section 7.6) eso es lo que la distribución ji cuadrado es: lo que obtienes cuando elevas al cuadrado un montón de cosas normalmente distribuidas y las sumas. Y dado que la distribución F es (nuevamente, por definición) lo que obtienes cuando calculas la relación entre dos cosas que están distribuidas en χ^2, tenemos nuestra distribución muestral. Obviamente, estoy pasando por alto un montón de cosas cuando digo esto, pero en términos generales, de aquí es de donde proviene nuestra distribución muestral. — [a] Si sigues leyendo Chapter 14 y observas cómo se define el "efecto de tratamiento" en el nivel k de un factor en términos de $ _k $ (ver sección sobre ANOVA factorial 2: diseños balanceados, interacciones permitidas]), resulta que Q se refiere a una media ponderada de los efectos del tratamiento al cuadrado, $Q = \frac{(\sum_{k=1}^{G} N_k \alpha_k^2)}{(G-1)}$ [b] O, si queremos ser rigurosos con la precisión, $1 + \frac{2}{df_2 - 2}$

13.2.4 Un ejemplo resuelto

La discusión anterior fue bastante abstracta y un poco técnica, por lo que creo que en este punto podría ser útil ver un ejemplo resuelto. Para ello, volvamos a los datos de los ensayos clínicos que introduje al principio del capítulo. Los estadísticos descriptivos que calculamos al principio nos dicen las medias de nuestro grupo: una mejora promedio en el estado de ánimo de $ 0.45 $ para el placebo, $ 0.72 $ para Anxifree y $ 1.48 $ para Joyzepam. Con eso en mente, imaginemos que estamos en 1899 [7] y comencemos haciendo algunos cálculos con lápiz y papel. Solo haré esto para las primeras observaciones de 5 porque no estamos en 1899 y soy muy vaga. Comencemos por calcular SS_w, las sumas de cuadrados intra grupo. Primero, elaboremos una tabla para ayudarnos con nuestros cálculos (Table 13.3)

Table 13.3: Un ejemplo resuelto...1

group k	**outcome Y_{ik}**
placebo	0.5
placebo	0.3
placebo	0.1
anxifree	0.6
anxifree	0.4

En este punto, lo único que he incluido en la tabla son los datos sin procesar. Es decir, la variable de agrupación (el fármaco) y la variable de resultado (el estado de ánimo.ganancia) para cada persona. Ten en cuenta que la variable de resultado aquí corresponde al valor $\bar{Y}_{ik}$ en nuestra ecuación anterior. El próximo paso en el cálculo es anotar, para cada persona en el estudio, la media del grupo correspondiente, $\bar{Y}_k$. Esto es un poco repetitivo pero no particularmente difícil dado que ya calculamos esas medias de grupo al hacer nuestros estadísticos descriptivos, ver Table 13.4.

Table 13.4: Un ejemplo resuelto...2

group k	**outcome Y_{ik}**	**group mean $\bar{Y}_k$**
placebo	0.5	0.45
placebo	0.3	0.45
placebo	0.1	0.45
anxifree	0.6	0.72
anxifree	0.4	0.72

Ahora que los hemos escrito, necesitamos calcular, nuevamente para cada persona, la desviación de la media del grupo correspondiente. Es decir, queremos restar $Y_{ik} - \bar{Y}_k$. Después de haber hecho eso, necesitamos elevar todo al cuadrado. Cuando hacemos eso, esto es lo que obtenemos (Table 13.5)

[7] O, para ser precisas, imagina que "es 1899 y no tenemos amigos y nada mejor que hacer con nuestro tiempo que hacer algunos cálculos que no habría tenido ningún sentido en 1899 porque ANOVA no existió hasta alrededor de la década de 1920".

Table 13.5: Un ejemplo resuelto...3

group k	**outcome** Y_{ik}	**group mean** $\bar{Y}_k$	**dev. from group mean** $Y_{ik} - \bar{Y}_k$	**squared deviation** $(Y_{ik} - \bar{Y}_k)^2$
placebo	0.5	0.45	0.05	0.0025
placebo	0.3	0.45	-0.15	0.0225
placebo	0.1	0.45	-0.35	0.1225
anxifree	0.6	0.72	-0.12	0.0136
anxifree	0.4	0.72	-0.32	0.1003

El último paso es igualmente sencillo. Para calcular la suma de cuadrados intra grupo, simplemente sumamos las desviaciones al cuadrado de todas las observaciones:

$$\begin{aligned} SS_w &= 0.0025 + 0.0225 + 0.1225 + 0.0136 + 0.1003 \\ &= 0.2614 \end{aligned}$$

Por supuesto, si realmente quisiéramos obtener la respuesta correcta, tendríamos que hacer esto para las 18 observaciones en el conjunto de datos, no solo para las primeras cinco. Podríamos continuar con los cálculos de lápiz y papel si quisiéramos, pero es bastante tedioso. Como alternativa, no es demasiado difícil hacer esto en una hoja de cálculo como OpenOffice o Excel. Inténtalo y hazlo tú misma. El que hice yo, en Excel, está en el archivo clinictrial_anova.xls. Cuando lo hagas, deberías terminar con un valor de suma de cuadrados intra grupo de $ 1.39 $.

Bueno. Ahora que hemos calculado la variabilidad intra grupos, SS_w, es hora de centrar nuestra atención en la suma de cuadrados entre grupos, SS_b. Los cálculos aquí son muy similares. La principal diferencia es que en lugar de calcular las diferencias entre una observación Yik y una media de grupo $\bar{Y}_k$ para todas las observaciones, calculamos las diferencias entre las medias de grupo $\bar{Y}_k$ y la media general $\bar{Y}$ (en este caso 0.88) para todos los grupos (Table 13.6).

Table 13.6: Un ejemplo resuelto...4

group k	**group mean** $\bar{Y}_k$	**grand mean** $\bar{Y}$	**deviation** $\bar{Y}_k - \bar{Y}$	**squared deviation** $(\bar{Y}_k - \bar{Y})^2$
placebo	0.45	0.88	-0.43	0.19
anxifree	0.72	0.88	-0.16	0.03
joyzepam	1.48	0.88	0.60	0.36

Sin embargo, para los cálculos entre grupos necesitamos multiplicar cada una de estas desviaciones al cuadrado por N_k, el número de observaciones en el grupo. Hacemos esto porque cada observación en el grupo (todas las N_k) está asociada con una diferencia

entre grupos. Así, si hay seis personas en el grupo de placebo y la media del grupo de placebo difiere de la media general en \$0,19 \$, entonces la variación total entre grupos asociada con estas seis personas es \$6 $\times$0,19 = 1,14 \$. Así que tenemos que ampliar nuestra tabla de cálculos (Table 13.7).

Table 13.7: Un ejemplo resuelto...5

group k	**...**	**squared deviations** $(\bar{Y}_k - \bar{Y})^2$	**sample size** N_k	**weighted squared dev** $N_k(\bar{Y}_k - \bar{Y})^2$
placebo	...	0.19	6	1.14
anxifree	...	0.03	6	0.18
joyzepam	...	0.36	6	2.16

Y ahora nuestra suma de cuadrados entre grupos se obtiene sumando estas "desviaciones cuadráticas ponderadas" sobre los tres grupos en el estudio:

$$\begin{aligned} SS_b &= 1.14 + 0.18 + 2.16 \\ &= 3.48 \end{aligned}$$

Como puedes ver, los cálculos entre grupos son mucho más cortos[8]. Ahora que hemos calculado nuestros valores de sumas de cuadrados, SS_b y SS_w, el resto del ANOVA es bastante sencillo. El siguiente paso es calcular los grados de libertad. Como tenemos $G = 3$ grupos y $N = 18$ observaciones en total, nuestros grados de libertad se pueden calcular mediante una simple resta:

$$\begin{aligned} df_b &= G - 1 = 2 \\ df_w &= NG = 15 \end{aligned}$$

A continuación, dado que ahora hemos calculado los valores de las sumas de cuadrados y los grados de libertad, tanto para la variabilidad intra grupos como para la variabilidad entre grupos, podemos obtener los valores de las medias cuadráticas dividiendo uno por el otro:

$$\begin{aligned} MS_b &= \frac{SS_b}{df_b} = \frac{3.48}{2} = 1.74 \\ MS_w &= \frac{SS_w}{df_w} = \frac{1.39}{15} = 0.09 \end{aligned}$$

Ya casi hemos terminado. Las medias cuadráticas se pueden usar para calcular el valor F, que es la prueba estadística que nos interesa. Hacemos esto dividiendo el valor de MC entre grupos por el valor de MC intra grupos.

[8] En el ensayo clínico de Excel anova.xls, el valor de SCb resultó ser ligeramente diferente, $3,45$, que el que se muestra en el texto anterior (¡redondeando errores!)

$$
\begin{aligned}
F &= \frac{MS_b}{MS_w} = \frac{1.74}{0.09} \\
&= 19,3
\end{aligned}
$$

¡Guauuu! Esto es muy emocionante, ¿verdad? Ahora que tenemos nuestra prueba estadística, el último paso es averiguar si la prueba en sí nos da un resultado significativo. Como se discutió en Chapter 9 en los "viejos tiempos", lo que haríamos sería abrir un libro de texto de estadística o pasar a la sección posterior que en realidad tendría una tabla de búsqueda enorme y encontraríamos el valor umbral F correspondiente a un valor particular de alfa (la región de rechazo de la hipótesis nula), por ejemplo $0,05$, $0,01$ o $0,001$, para 2 y 15 grados de libertad. Hacerlo de esta manera nos daría un valor umbral de F para un alfa de 0.001 de 11.34. Como esto es menor que nuestro valor F calculado, decimos que $p < 0.001$. Pero esos era antes, y ahora el sofisticado software de estadística calcula el valor p exacto por ti. De hecho, el valor p exacto es 0.000071. Entonces, a menos que estemos siendo *extremadamente* conservadores con respecto a nuestra tasa de error Tipo I, estamos prácticamente seguras de que podemos rechazar la hipótesis nula.

En este punto, básicamente hemos terminado. Habiendo completado nuestros cálculos, es tradicional organizar todos estos números en una tabla ANOVA como la de la Tabla 13.1. Para los datos de nuestro ensayo clínico, la tabla ANOVA se vería como Table 13.8.

Table 13.8: La tabla de resultados de ANOVA

	df	**sum of squares**	**mean squares**	**F-statistic**	**p-value**
between groups	2	3.48	1.74	19.3	0.000071
within groups	15	1.39	0.09	-	-

En estos días, probablemente no querrás construir una de estas tablas tú misma, pero encontrarás que casi todo el software estadístico (incluido jamovi) tiende a organizar la salida de un ANOVA en una tabla como esta, por lo que es una buena idea para acostumbrarse a leerlas. Sin embargo, aunque el software generará una tabla ANOVA completa, casi nunca se incluye la tabla completa en tu redacción. Una forma bastante estándar de informar del apartado de estadística sería escribir algo como esto:

> *ANOVA de un factor mostró un efecto significativo de la droga en el estado de ánimo (F(2,15) = 19.3, p < .001).*

Ains. Tanto trabajo para una frase corta.

13.3 Ejecutando un ANOVA en jamovi

Estoy bastante segura de saber lo que estás pensando después de leer la última sección, especialmente si seguiste mi consejo e hiciste todo eso con lápiz y papel (es decir, en una hoja de cálculo) tú misma. Hacer los cálculos de ANOVA tú misma apesta. Hay

muchos cálculos que necesitamos hacer en el camino, y sería tedioso tener que hacer esto una y otra vez cada vez que quisieras hacer un ANOVA.

13.3.1 Uso de jamovi para especificar tu ANOVA

Para facilitarte la vida, jamovi puede hacer ANOVA... ¡hurra! Ves a 'ANOVA' - Análisis 'ANOVA' y mueve la variable mood.gain para que esté en el cuadro 'Variable dependiente', y luego mueve la variable de fármaco para que esté en el cuadro 'Factores fijos'. Esto debería dar los resultados como se muestra en Figure 13.3. [9] Ten en cuenta que también marqué la casilla de verificación η^2, pronunciada "eta" al cuadrado, en la opción 'Tamaño del efecto' y esto también se muestra en la tabla de resultados. Volveremos a los tamaños del efecto un poco más tarde.

ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η²
drug	3.45	2	1.73	18.61	0.00009	0.71
Residuals	1.39	15	0.09			

Figure 13.3: tabla de resultados jamovi para ANOVA de aumento del estado de ánimo por fármaco administrado

La tabla de resultados de jamovi te muestra los valores de las sumas de cuadrados, los grados de libertad y un par de otras cantidades que no nos interesan en este momento. Ten en cuenta, sin embargo, que jamovi no usa los nombres "entre grupos" y "intra grupo". En su lugar, intenta asignar nombres más significativos. En nuestro ejemplo particular, la varianza entre grupos corresponde al efecto que el fármaco tiene sobre la variable de resultado, y la varianza intra grupos corresponde a la variabilidad "sobrante", por lo que se denomina residual. Si comparamos estos números con los números que calculé a mano en [Un ejemplo práctico], puedes ver que son más o menos iguales, aparte de los errores de redondeo. La suma de cuadrados entre grupos es $SS_b = 3.45$, la suma de cuadrados intra grupos es $SS_w = 1.39$, y los grados de libertad son 2 y 15 respectivamente. También obtenemos el valor F y el valor p y, nuevamente, estos son más o menos iguales, sumando o restando errores de redondeo, a los números que calculamos nosotras mismas al hacerlo de la manera larga y tediosa.

13.4 Tamaño del efecto

Hay algunas formas diferentes de medir el tamaño del efecto en un ANOVA, pero las medidas más utilizadas son η^2 (eta al cuadrado) y η^2 parcial. Para un análisis de varianza unifactorial, son idénticos entre sí, así que por el momento solo explicaré η^2 . La definición de η^2 es realmente muy simple

[9] los resultados de jamovi son más precisos que los del texto anterior, debido a errores de redondeo.

$$\eta^2 = \frac{SS_b}{SS_{total}}$$

Eso es todo. Entonces, cuando miro la tabla ANOVA en Figure 13.3, veo que $SS_b = 3,45$ y $SS_tot = 3,45 + 1,39 = 4,84$. Así obtenemos un valor de η^2 de

$$\eta^2 = \frac{3.45}{4.84} = 0.71$$

La interpretación de η^2 es igualmente sencilla. Se refiere a la proporción de la variabilidad en la variable de resultado (mood.gain) que se puede explicar en términos del predictor (fármaco). Un valor de $\eta^2 = 0$ significa que no existe ninguna relación entre los dos, mientras que un valor de $\eta^2 = 1$ significa que la relación es perfecta. Mejor aún, el valor de η^2 está muy relacionado con R^2, como se explicó anteriormente en Section 12.6.1, y tiene una interpretación equivalente. Aunque muchos libros de texto de estadística sugieren η^2 como la medida predeterminada del tamaño del efecto en ANOVA, hay una <a href="https://daniellakens.blogspot.com/2015/06/why-you-should-use" interesante. -omega-squared.html" target="_blank">entrada de blog de Daniel Lakens que sugiere que eta cuadrado quizás no sea la mejor medida del tamaño del efecto en el análisis de datos del mundo real, porque puede ser un estimador sesgado. Acertadamente, también hay una opción en jamovi para especificar omega-cuadrado (ω^2), que es menos sesgado, junto con eta-cuadrado.

13.5 Comparaciones múltiples y pruebas post hoc

Cada vez que ejecutes un ANOVA con más de dos grupos y termines con un efecto significativo, lo primero que probablemente querrás preguntar es qué grupos son realmente diferentes entre sí. En nuestro ejemplo de fármacos, nuestra hipótesis nula fue que los tres fármacos (placebo, Anxifree y Joyzepam) tienen exactamente el mismo efecto sobre el estado de ánimo. Pero si lo piensas bien, la hipótesis nula en realidad afirma tres cosas diferentes a la vez aquí. En concreto, afirma que:

- El fármaco de tu competidor (Anxifree) no es mejor que un placebo (es decir, $\mu_A = \mu_P$)
- Tu fármaco (Joyzepam) no es mejor que un placebo (es decir, $\mu_J = \mu_P$)
- Anxifree y Joyzepam son igualmente efectivos (es decir, $\mu_J = \mu_A$)

Si alguna de esas tres afirmaciones es falsa, entonces la hipótesis nula también es falsa. Entonces, ahora que hemos rechazado nuestra hipótesis nula, estamos pensando que al menos una de esas cosas no es cierta. ¿Pero cuál? Las tres proposiciones son de interés. Dado que deseas saber si tu nuevo fármaco Joyzepam es mejor que un placebo, sería bueno saber cómo actúa en relación a una alternativa comercial existente (es decir, Anxifree). Incluso sería útil comprobar el rendimiento de Anxifree frente al placebo. Incluso si Anxifree ya ha sido ampliamente probado contra placebos por otros investigadores, aún puede ser muy útil verificar que tu estudio esté produciendo resultados similares a trabajos anteriores.

Cuando caracterizamos la hipótesis nula en términos de estas tres proposiciones distintas, queda claro que hay ocho "estados del mundo" posibles entre los que debemos distinguir (Table 13.9).

Table 13.9: La hipótesis nula y ocho posibles 'estados del mundo'

possibility:	**is** $\mu_P = \mu_A$**?**	**is** $\mu_P = \mu_J$**?**	**is** $\mu_A = \mu_J$**?**	**which hypothesis?**
1	✓	✓	✓	null
2	✓	✓		alternative
3	✓		✓	alternative
4	✓			alternative
5	✓	✓	✓	alternative
6		✓		alternative
7			✓	alternative
8				alternative

Al rechazar la hipótesis nula, hemos decidido que no creemos que el número 1 sea el verdadero estado del mundo. La siguiente pregunta es, ¿cuál de las otras siete posibilidades *creemos* que es correcta? Cuando te enfrentas a esta situación, por lo general ayuda mirar los datos. Por ejemplo, si observamos las gráficas en Figure 13.1, es tentador concluir que Joyzepam es mejor que el placebo y mejor que Anxifree, pero no hay una diferencia real entre Anxifree y el placebo. Sin embargo, si queremos obtener una respuesta más clara sobre esto, podría ser útil realizar algunas pruebas.

13.5.1 Ejecución de pruebas t "por pares"

¿Cómo podríamos solucionar nuestro problema? Dado que tenemos tres pares separados de medias (placebo versus Anxifree, placebo versus Joyzepam y Anxifree versus Joyzepam) para comparar, lo que podríamos hacer es ejecutar tres pruebas t separadas y ver qué sucede. Esto es fácil de hacer en jamovi. Puedes ir a las opciones de ANOVA 'Pruebas post hoc', mueve la variable 'fármaco' al cuadro activo de la derecha y luego haz clic en la casilla de verificación 'Sin corrección'. Esto producirá una tabla ordenada que muestra todas las comparaciones de la prueba t por pares entre los tres niveles de la variable del fármaco, como en Figure 13.4

13.5.2 Correcciones para pruebas múltiples

En la sección anterior, insinué que hay un problema con ejecutar montones y montones de pruebas t. El problema es que, al ejecutar estos análisis, lo que estamos haciendo es una "expedición de pesca". Estamos realizando montones de pruebas sin mucha orientación teórica con la esperanza de que algunas de ellas resulten significativas. Este tipo de búsqueda sin teoría de diferencias entre grupos se conoce como **análisis post hoc** ("post hoc" en latín significa "después de esto").[10]

Está bien ejecutar análisis post hoc, pero hay que tener mucho cuidado. Por ejemplo, se debe evitar el análisis que realicé en la sección anterior, ya que cada prueba t individual

[10]si *tienes* alguna base teórica para querer investigar algunas comparaciones pero no otras, la historia es diferente. En esas circunstancias, en realidad no estás ejecutando análisis "post hoc" en absoluto, estás haciendo "comparaciones planificadas". Hablo de esta situación más adelante en el libro: Section 14.9, pero por ahora quiero mantener las cosas simples.

Post Hoc Tests

Post Hoc Comparisons - drug

Comparison			Mean Difference	SE	df	t	p
drug		drug					
anxifree	-	joyzepam	−0.77	0.18	15.00	−4.36	0.00056
	-	placebo	0.27	0.18	15.00	1.52	0.15021
joyzepam	-	placebo	1.03	0.18	15.00	5.88	0.00003

Note. Comparisons are based on estimated marginal means

>

Figure 13.4: Pruebas t por pares no corregidas como comparaciones post hoc en jamovi

está diseñada para tener una tasa de error Tipo I del 5 % (es decir, $\alpha = .05$) y realicé tres de estas pruebas. Imagina lo que hubiera pasado si mi ANOVA involucrara 10 grupos diferentes, y hubiera decidido ejecutar 45 pruebas t "post hoc" para tratar de averiguar cuáles eran significativamente diferentes entre sí, esperaría que 2 o 3 de ellas resultaran significativas solo por casualidad. Como vimos en Chapter 9, el principio organizador central que subyace a la prueba de hipótesis nula es que buscamos controlar nuestra tasa de error Tipo I, pero ahora que estoy ejecutando muchas pruebas t a la vez para determinar la fuente de mis resultados de ANOVA, mi tasa de error Tipo I real se ha salido completamente de control.

La solución habitual a este problema es introducir un ajuste en el valor p, cuyo objetivo es controlar la tasa de error total en toda la familia de pruebas (ver Shaffer (1995)). Un ajuste de esta forma, que generalmente (pero no siempre) se aplica porque una está haciendo un análisis post hoc, a menudo se denomina ** corrección para comparaciones múltiples **, aunque a veces se denomina "inferencia simultánea". En cualquier caso, hay bastantes formas diferentes de hacer este ajuste. Discutiré algunas de ellas en esta sección y en Section 14.8 el próximo capítulo, pero debes tener en cuenta que existen muchos otros métodos (consulta, por ejemplo, Hsu (1996)).

13.5.3 Correcciones de Bonferroni

El más simple de estos ajustes se llama la **corrección de Bonferroni** (Dunn, 1961), y es muy, muy simple. Supongamos que mi análisis post hoc consta de m pruebas separadas, y quiero asegurarme de que la probabilidad total de cometer *cualquier* error de tipo I sea como máximo α.[11] Si es así, entonces la corrección de Bonferroni simplemente dice "multiplique todos sus valores p sin procesar por m". Si dejamos que p denote el valor p original y que $p_j^{'}$ sea el valor corregido, entonces la corrección de Bonferroni dice que:

$$p_j^{'} = m \times p$$

[11] vale la pena señalar de paso que no todos los métodos de ajuste intentan hacer esto. Lo que he descrito aquí es un enfoque para controlar lo que se conoce como "Family-wise Type I error rate". Sin embargo, existen otras pruebas post hoc que buscan controlar la "tasa de descubrimiento falso", que es algo diferente.

Y por lo tanto, si usas la corrección de Bonferroni, rechazarías la hipótesis nula si $p_j^{'} < \alpha$. La lógica de esta corrección es muy sencilla. Estamos haciendo m pruebas diferentes, por lo que si lo organizamos para que cada prueba tenga una tasa de error de tipo I de $\frac{\alpha}{m}$ como máximo, entonces la tasa de error de tipo I *total* en estas pruebas no puede ser mayor que α. Eso es bastante simple, tanto que en el artículo original, el autor escribe:

> El método dado aquí es tan simple y tan general que estoy seguro de que debe haber sido usado antes. Sin embargo, no lo encuentro, por lo que solo puedo concluir que quizás su misma simplicidad ha impedido que los estadísticos se den cuenta de que es un método muy bueno en algunas situaciones (Dunn (1961), pp 52-53).

Para usar la corrección de Bonferroni en jamovi, simplemente haz clic en la casilla de verificación 'Bonferroni' en las opciones de 'Corrección' y verás otra columna añadida a la tabla de resultados de ANOVA que muestra los valores p ajustados para la corrección de Bonferroni (Table 13.8). Si comparamos estos tres valores p con los de las pruebas t por pares sin corregir, está claro que lo único que ha hecho jamovi es multiplicarlos por 3.

13.5.4 Correcciones de Holm

Aunque la corrección de Bonferroni es el ajuste más simple que existe, no suele ser el mejor. Un método que se usa a menudo es la **corrección de Holm** (Holm, 1979). La idea detrás de la corrección de Holm es pretender que está haciendo las pruebas secuencialmente, comenzando con el valor p más pequeño (sin procesar) y avanzando hacia el más grande. Para el j-ésimo mayor de los valores p, el ajuste es *cualquiera*

$$p_j^{'} = j \times p_j$$

(es decir, el valor de p más grande permanece sin cambios, el segundo valor de p más grande se duplica, el tercer valor de p más grande se triplica, y así sucesivamente), o

$$p_j^{'} = p_{j+1}^{'}$$

el que sea más grande. Esto puede ser un poco confuso, así que hagámoslo un poco más despacio. Esto es lo que hace la corrección de Holm. Primero, ordena todos sus valores p en orden, de menor a mayor. Para el valor p más pequeño, todo lo que tiene que hacer es multiplicarlo por m y listo. Sin embargo, para todos los demás es un proceso de dos etapas. Por ejemplo, cuando pasa al segundo valor p más pequeño, primero lo multiplica por $m - 1$. Si esto produce un número que es mayor que el valor p ajustado que obtuvo la última vez, entonces lo conserva. Pero si es más pequeño que el último, copia el último valor p. Para ilustrar cómo funciona esto, considera Table 13.10 que muestra los cálculos de una corrección de Holm para una colección de cinco valores p.

Esperemos que eso aclare las cosas.

Aunque es un poco más difícil de calcular, la corrección de Holm tiene algunas propiedades muy buenas. Es más potente que Bonferroni (es decir, tiene una tasa de error de tipo II más baja) pero, aunque parezca contradictorio, tiene la misma tasa de error tipo I. Como consecuencia, en la práctica casi nunca se utiliza la corrección de Bonferroni, ya que siempre es superada por la corrección de Holm, un poco más

Table 13.10: valores de p corregidos por Holm

raw p	rank j	p × j	Holm p
.001	5	.005	.005
.005	4	.020	.020
.019	3	.057	.057
.022	2	.044	.057
.103	1	.103	.103

elaborada. Debido a esto, la corrección de Holm debería ser tu *ir a* corrección de comparación múltiple. Figure 13.4 también muestra los valores p corregidos de Holm y, como puedes ver, el valor p más grande (correspondiente a la comparación entre Anxifree y el placebo) no se modifica. Con un valor de .15, es exactamente el mismo que el valor que obtuvimos originalmente cuando no aplicamos ninguna corrección. Por el contrario, el valor de p más pequeño (Joyzepam frente a placebo) se ha multiplicado por tres.

13.5.5 Redacción de la prueba post hoc

Finalmente, después de ejecutar el análisis post hoc para determinar qué grupos son significativamente diferentes entre sí, puedes escribir el resultado de esta manera:

> *Las pruebas post hoc (usando la corrección de Holm para ajustar p) indicaron que Joyzepam produjo un cambio de humor significativamente mayor que Anxifree ($p = .001$) y el placebo ($(p = 9.0 \times 10^{-5})$). No encontramos evidencia de que Anxifree funcionara mejor que el placebo ($p = .15$).*

O, si no te gusta la idea de informar valores p exactos, entonces cambiarías esos números por $p < .01$, $p < .001$ y $p > .05$ respectivamente. De todas formas, la clave es que indiques que utilizaste la corrección de Holm para ajustar los valores p. Y, por supuesto, asumo que en otra parte del artículo has incluido los estadísticos descriptivos relevantes (es decir, las medias del grupo y las desviaciones estándar), ya que estos valores p por sí solos no son muy informativos.

13.6 Los supuestos de ANOVA unifactorial

Como cualquier prueba estadística, el análisis de varianza se basa en algunas suposiciones sobre los datos, específicamente los residuales. Hay tres suposiciones clave que debes tener en cuenta: normalidad, homogeneidad de varianzas e independencia.

[Detalle técnico adicional [12]]

[12] si recuerdas [Un ejemplo práctico], que espero que al menos hayas leído por encima incluso si no lo leíste todo, describí los modelos estadísticos que sustentan ANOVA de esta manera:

$$H_0 : Y_{ik} = \mu + \epsilon_{ik}$$

$$H_1 : Y_{ik} = \mu_k + \epsilon_{ik}$$

En estas ecuaciones μ se refiere a una única media general poblacional que es la misma para todos los grupos, y µk es la media poblacional del k-ésimo grupo. Hasta este punto, nos ha interesado principalmente si nuestros datos se describen mejor en términos de una media general única (la hipótesis

Entonces, ¿cómo verificamos si la suposición sobre los residuales es correcta? Bueno, como indiqué anteriormente, hay tres afirmaciones distintas subyacentes en esta declaración, y las consideraremos por separado.

- **Homogeneidad de varianzas**. Fíjate que solo tenemos un valor para la desviación estándar de la población (es decir, σ), en lugar de permitir que cada grupo tenga su propio valor (es decir, σ_k). Esto se conoce como el supuesto de homogeneidad de varianzas (a veces llamado homocedasticidad). ANOVA asume que la desviación estándar de la población es la misma para todos los grupos. Hablaremos de esto extensamente en la sección Comprobación del supuesto de homogeneidad de varianzas.
- **Normalidad**. Se supone que los residuales se distribuyen normalmente. Como vimos en Section 11.9, podemos evaluar esto mirando gráficos QQ (o ejecutando una prueba de Shapiro-Wilk). Hablaré más sobre esto en un contexto ANOVA en la sección Comprobación del supuesto de normalidad.
- **Independencia**. La suposición de independencia es un poco más complicada. Lo que básicamente significa es que conocer un residual no dice nada sobre ningún otro residual. Se supone que todos los valores de ϵ_{ik} han sido generados sin ninguna "consideración" o "relación con" ninguno de los otros. No hay una manera obvia o simple de probar esto, pero hay algunas situaciones que son violaciones claras de esto. Por ejemplo, si tienes un diseño de medidas repetidas, donde cada participante en tu estudio aparece en más de una condición, entonces la independencia no se cumple. ¡Hay una relación especial entre algunas observaciones, a saber, aquellas que corresponden a la misma persona! Cuando eso sucede, debes usar un ANOVA unifactorial de medidas repetidas.

13.6.1 Comprobación del supuesto de homogeneidad de varianzas

> *¡Hacer la prueba preliminar de las varianzas es como hacerse a la mar en un bote de remos para averiguar si las condiciones son lo suficientemente tranquilas para que un trasatlántico salga del puerto!*
> – Caja de Jorge (G. E. P. Box, 1953)

Hay más de una manera de probar el supuesto de homogeneidad de varianzas. La prueba más utilizada para esto que he visto en la literatura es la prueba de Levene (Levene, 1960), y la prueba de Brown-Forsythe estrechamente relacionada (Brown & Forsythe, 1974).

Independientemente de si estás haciendo la prueba estándar de Levene o la prueba de Brown-Forsythe, la prueba estadística, que a veces se denota F pero también a veces se escribe como W, se calcula exactamente de la misma manera que la F para el ANOVA,

nula) o en términos de diferentes medias específicas de grupo (la hipótesis alternativa). ¡Esto tiene sentido, por supuesto, ya que esa es en realidad la pregunta de investigación importante! Sin embargo, todos nuestros procedimientos de prueba se han basado implícitamente en una suposición específica sobre los residuales, ϵ_ik, a saber, que

$$\epsilon_{ik} \sum Normal(0, \sigma^2)$$

Ningún procedimiento matemático funciona correctamente sin esta suposición. O, para ser precisos, puedes hacer todos los cálculos y terminarás con un estadístico F, pero no tienes ninguna garantía de que esta F realmente mida lo que crees que está midiendo, por lo que cualquier conclusión que puedas sacar en base a la prueba F podría ser incorrecta.

simplemente usando Z_{ik} en lugar de Y_{ik}. Con eso en mente, podemos pasar a ver cómo ejecutar la prueba en jamovi.

[Detalle técnico adicional [13]]

13.6.2 Ejecutando la prueba de Levene en jamovi

Bien, entonces, ¿cómo hacemos la prueba de Levene? Realmente simple: en la opción "Comprobaciones de supuestos" de ANOVA, simplemente haz clic en la casilla de verificación "Pruebas de homogeneidad". Si observamos el resultado, que se muestra en Figure 13.5, vemos que la prueba no es significativa ($F_{2,15} = 1.45, p = .266$), por lo que parece que el supuesto de homogeneidad de varianzas está bien. Sin embargo, ¡las apariencias pueden engañar! Si el tamaño de tu muestra es bastante grande, entonces la prueba de Levene podría mostrar un efecto significativo (es decir, p < .05) incluso cuando el supuesto de homogeneidad varianzas no se viole hasta el punto de afectar la solidez de ANOVA. Este era el punto al que George Box se refería en la cita anterior. De manera similar, si el tamaño de tu muestra es bastante pequeño, es posible que no se satisfaga el supuesto de homogeneidad de varianzas y, sin embargo, una prueba de Levene podría no ser significativa (es decir, p > 0,05). Lo que esto significa es que, junto con cualquier prueba estadística del cumplimiento del supuesto, siempre debes trazar la desviación estándar alrededor de las medias para cada grupo/categoría en el análisis... solo para ver si son bastante similares (es decir, homogeneidad de varianzas) o no.

13.6.3 Eliminar el supuesto de homogeneidad de varianzas

En nuestro ejemplo, el supuesto de homogeneidad de varianzas resultó ser bastante seguro: la prueba de Levene resultó no significativa (a pesar de que también deberíamos mirar el gráfico de las desviaciones estándar), por lo que probablemente no tengamos que preocuparnos. Sin embargo, en la vida real no siempre somos tan afortunados. ¿Cómo realizamos nuestro ANOVA cuando se viola el supuesto de homogeneidad de varianzas? Si recuerdas nuestra discusión sobre las pruebas t, vimos este problema antes. La prueba t de Student asume varianzas iguales, por lo que la solución fue usar la prueba t de Welch, que no lo hace. De hecho, Welch (1951) también mostró cómo

[13] la prueba de Levene es sorprendentemente simple. Supongamos que tenemos nuestra variable de resultado Y_{ik}. Todo lo que hacemos es definir una nueva variable, que llamaré Z_{ik}, correspondiente a la desviación absoluta de la media del grupo

$$Z_{ik} = Y_{ik} - \bar{Y}_k$$

Vale, ¿de qué nos sirve esto? Bueno, pensemos un momento qué es realmente Z_{ik} y qué estamos tratando de probar. El valor de Z_{ik} es una medida de cómo la i-ésima observación en el k-ésimo grupo se desvía de la media de su grupo. Y nuestra hipótesis nula es que todos los grupos tienen la misma varianza, es decir, ¡las mismas desviaciones generales de las medias del grupo! Entonces, la hipótesis nula en una prueba de Levene es que las medias poblacionales de Z son idénticas para todos los grupos. Mmm. Entonces, lo que necesitamos ahora es una prueba estadística de la hipótesis nula de que todas las medias de los grupos son ìguales. ¿Donde hemos visto eso antes? Ah, claro, eso es ANOVA, y todo lo que hace la prueba de Levene es ejecutar un ANOVA en la nueva variable Z_{ik}. ¿Qué pasa con la prueba de Brown-Forsythe? ¿Hace algo particularmente diferente? No. El único cambio con respecto a la prueba de Levene es que construye la variable transformada Z de una manera ligeramente diferente, utilizando desviaciones respecto a las medianas del grupo en lugar de desviaciones respecto a las medias del grupo. Es decir, para la prueba de Brown-Forsythe:

$$Z_{ik} = Y_{ik} - median_k(Y)$$

donde $median_k(Y)$ es la mediana del grupo k.

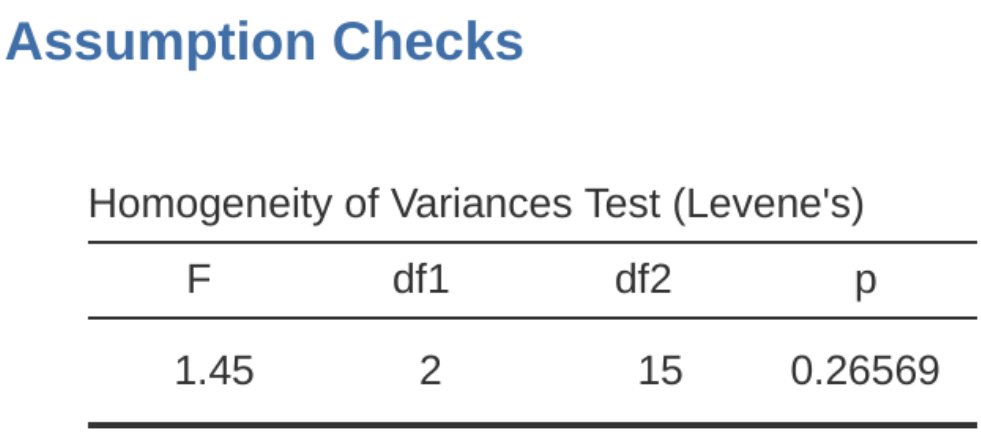


Figure 13.5: salida de prueba de Levene para ANOVA unidireccional en jamovi

podemos resolver este problema para ANOVA (la **prueba unifactorial de Welch**). Se implementa en jamovi utilizando el análisis ANOVA unifactorial. Se trata de un enfoque de análisis específico solo para ANOVA unifactorial, y para ejecutar el ANOVA unifactorial de Welch para nuestro ejemplo, volveríamos a ejecutar el análisis como antes, pero esta vez usamos el comando de análisis jamovi ANOVA - ANOVA unifactorial, y marcas la opción para la prueba de Welch (ver Figure 13.6). Para comprender lo que está sucediendo aquí, comparemos estos números con lo que obtuvimos anteriormente cuando Ejecutando un ANOVA en jamovi originalmente. Para ahorrarte la molestia de retroceder, esto es lo que obtuvimos la última vez: $F(2, 15) = 18,611, p = 0,00009$, que también se muestra como la prueba de Fisher en el ANOVA unifactorial que se muestra en Figure 13.6.

Bien, originalmente nuestro ANOVA nos dio el resultado $F(2, 15) = 18,6$, mientras que la prueba unifactorial de Welch nos dio $F(2, 9,49) = 26,32$. En otras palabras, la prueba de Welch ha reducido los grados de libertad dentro de los grupos de 15 a 9,49 y el valor F ha aumentado de 18,6 a 26,32.

13.6.4 Comprobación del supuesto de normalidad

Probar el supuesto de normalidad es relativamente sencillo. Cubrimos la mayor parte de lo que necesitas saber en Section 11.9. Lo único que realmente necesitamos hacer es dibujar un gráfico QQ y, además, si está disponible, ejecutar la prueba de Shapiro-Wilk. El gráfico QQ se muestra en Figure 13.7 y me parece bastante normal. Si la prueba de Shapiro-Wilk no es significativa (es decir, $p > 0,05$), esto indica que no se viola el supuesto de normalidad. Sin embargo, al igual que con la prueba de Levene, si el tamaño de la muestra es grande, una prueba significativa de Shapiro-Wilk puede ser de hecho un falso positivo, donde la suposición de normalidad no se viola en ningún sentido problemático sustantivo para el análisis. Y, del mismo modo, una muestra muy pequeña puede producir falsos negativos. Por eso es importante una inspección visual del gráfico QQ.

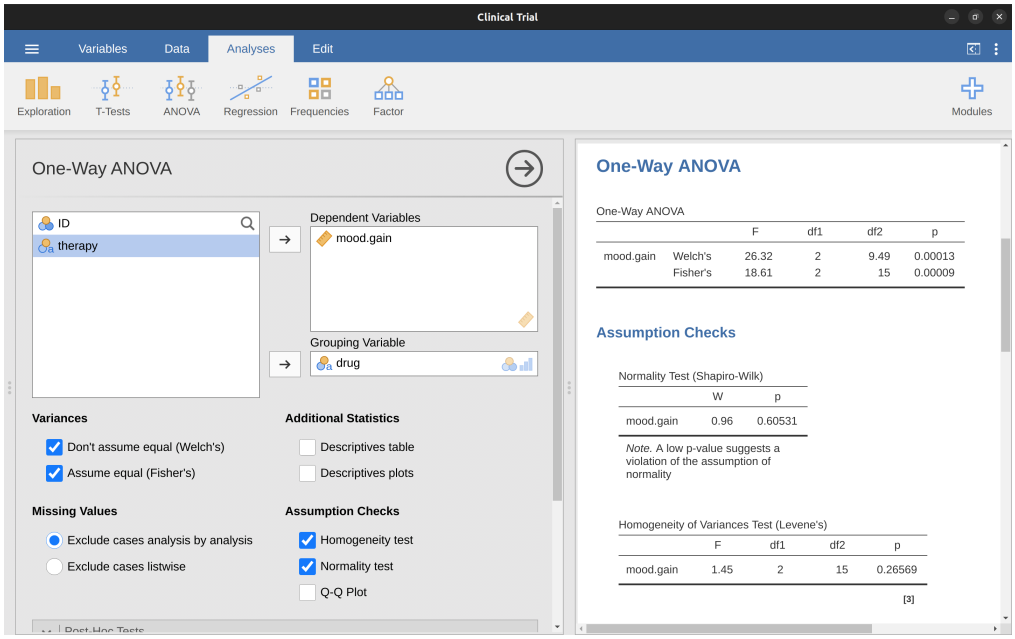


Figure 13.6: prueba de Welch como parte del análisis ANOVA unifactorial en jamovi

Además de inspeccionar el gráfico QQ en busca de desviaciones de la normalidad, la prueba de Shapiro-Wilk para nuestros datos muestra un efecto no significativo, con p = 0,6053 (ver Figure 13.6. Por lo tanto, esto respalda la evaluación del gráfico QQ; ambas comprobaciones no encuentran indicios de que se viole la normalidad.

13.6.5 Eliminando el supuesto de normalidad

Ahora que hemos visto cómo verificar la normalidad, nos preguntamos qué podemos hacer para abordar las violaciones de la normalidad. En el contexto de un ANOVA unifactorial, la solución más fácil probablemente sea cambiar a una prueba no paramétrica (es decir, una que no se base en ningún supuesto particular sobre el tipo de distribución involucrada). Hemos visto pruebas no paramétricas antes, en Chapter 11. Cuando solo tienes dos grupos, la prueba de Mann-Whitney o Wilcoxon proporciona la alternativa no paramétrica que necesitas. Cuando tengas tres o más grupos, puedes usar la **prueba de suma de rangos de Kruskal-Wallis** (Kruskal & Wallis, 1952). Esa es la prueba de la que hablaremos a continuación.

13.6.6 La lógica detrás de la prueba de Kruskal-Wallis

La prueba de Kruskal-Wallis es sorprendentemente similar a ANOVA, en algunos aspectos. En ANOVA comenzamos con Y_{ik}, el valor de la variable de resultado para la i-ésima persona en el k-ésimo grupo. Para la prueba de Kruskal Wallis, lo que haremos es ordenar por rango todos estos valores de Y_{ik} y realizar nuestro análisis en los datos

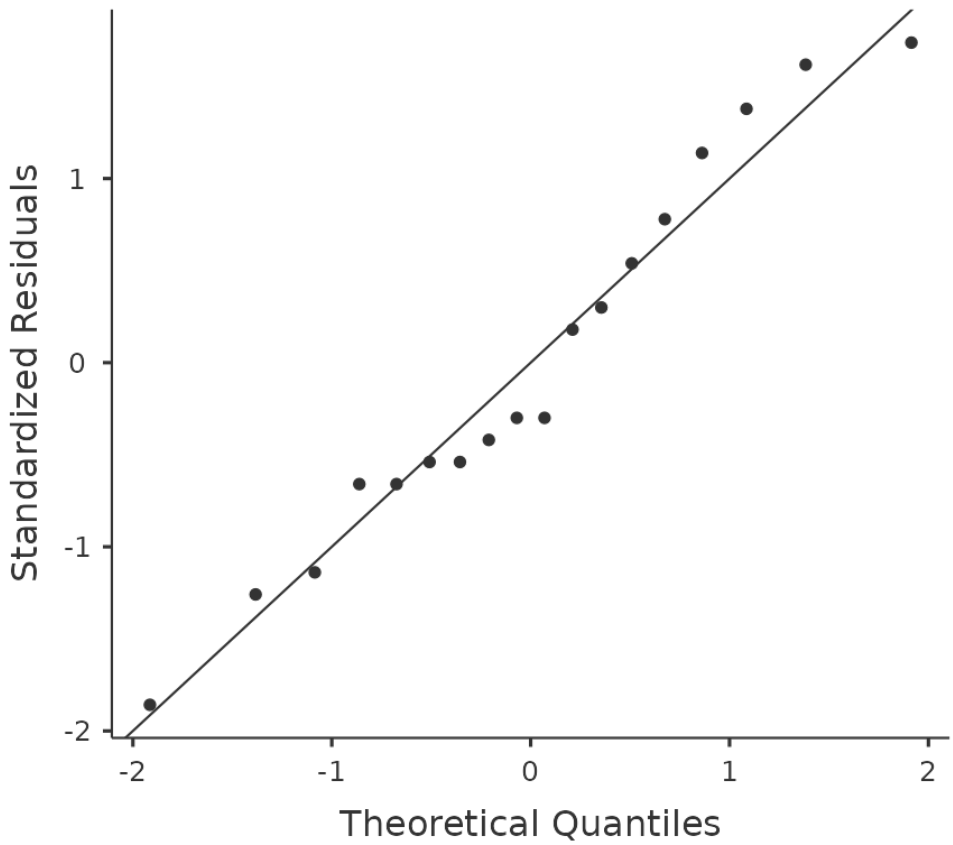


Figure 13.7: gráfico QQ en el análisis ANOVA unifactorial en jamovi

clasificados. [14]

13.6.7 Detalles adicionales

La descripción en la sección anterior ilustra la lógica que subyace a la prueba de Kruskal-Wallis. A nivel conceptual, esta es la forma correcta de pensar en cómo funciona la prueba.[15]

¡Pero espera hay mas! La historia que he contado hasta ahora solo es cierta cuando no hay vínculos en los datos sin procesar. Es decir, si no hay dos observaciones que

[14] Dejemos que R_{ik} se refiera a la clasificación otorgada al i-ésimo miembro del k-ésimo grupo. Ahora, calculemos $\bar{R}_k$, el rango promedio dado a las observaciones en el grupo k-ésimo

$$\bar{R}_k = \frac{1}{N_k} \sum_i R_{ik}$$

y también calcula $\bar{R}$, el rango medio general

$$\bar{R} = \frac{1}{N} \sum_i \sum_k R_{ik}$$

Ahora que hemos hecho esto, podemos calcular las desviaciones al cuadrado del rango medio general $\bar{R}$. Cuando hacemos esto para las puntuaciones individuales, es decir, si calculamos $(R_{ik} - \bar{R})^2$, lo que tenemos es una medida "no paramétrica" de cuánto se desvía la ik-ésima observación del rango medio general . Cuando calculamos la desviación al cuadrado de las medias del grupo de las medias generales, es decir, si calculamos $(R_{ik} - \bar{R})^2$, entonces lo que tenemos es una medida no paramétrica de cuánto el grupo se desvía del rango medio general. Con esto en mente, seguiremos la misma lógica que hicimos con ANOVA y definiremos nuestras medidas de sumas de cuadrados ordenadas, como lo hicimos antes. Primero, tenemos nuestras "sumas de cuadrados totales ordenadas"

$$RSS_{tot} = \sum_k \sum_i (R_{ik} - \bar{R})^2$$

y podemos definir las "sumas de cuadrados ordenadas entre grupos" como este

$$\begin{aligned} RSS_b &= \sum k \sum_i (\bar{R}_k - \bar{R})^2 \\ &= \sum_k N_k (\bar{R}_k - \bar{R})^2 \end{aligned}$$

Entonces, si la hipótesis nula es verdadera y no hay ninguna diferencia verdadera entre los grupos, esperarías que las sumas ordenadas entre grupos RSS_b fueran muy pequeñas, mucho más pequeñas que las sumas ordenadas totales RSS_{tot}. Cualitativamente, esto es muy similar a lo que encontramos cuando construimos el estadístico F de ANOVA, pero por razones técnicas, el estadístico de Kruskal-Wallis, generalmente denominado K, se construye de una manera ligeramente diferente,

$$K = (N-1) \times \frac{RSS_b}{RSS_{tot}}$$

y si la hipótesis nula es verdadera, entonces la distribución muestral de K es aproximadamente ji cuadrada con $G-1$ grados de libertad (donde $G es el número de grupos). Cuanto mayor sea el valor de K, menos consistentes serán los datos con la hipótesis nula, por lo que esta es una prueba unilateral. Rechazamos H_0 cuando K es suficientemente grande.

[15] Sin embargo, desde una perspectiva puramente matemática es innecesariamente complicado. No te mostraré la derivación, pero puedes usar un poco de ingenio algebraicos[b] para ver que la ecuación para K puede ser

$$K = \frac{12}{N(N-1)} \sum_k N_k \bar{R}_k^2 - 3(N+1)$$

Es esta última ecuación la que a veces ves para K. Es mucho más fácil de calcular que la versión que describí en la sección anterior, pero es solo que no tiene sentido para los humanos reales. Probablemente sea mejor pensar en K como lo describí anteriormente, como un análogo de ANOVA basado en rangos. Pero ten en cuenta que la prueba estadística que se calcula termina con un aspecto bastante diferente al que usamos para nuestro ANOVA original. — *b* Un término técnico

tengan exactamente el mismo valor. Si hay empates, entonces tenemos que introducir un factor de corrección a estos cálculos. En este punto, asumo que incluso el lector más diligente ha dejado de preocuparse (o al menos se ha formado la opinión de que el factor de corrección de empates es algo que no requiere su atención inmediata). Así que te diré muy rápidamente cómo se calcula y omitiré los tediosos detalles sobre por qué se hace de esta manera. Supongamos que construimos una tabla de frecuencias para los datos sin procesar y que fj sea el número de observaciones que tienen el j-ésimo valor único. Esto puede sonar un poco abstracto, así que aquí hay un ejemplo concreto de la tabla de frecuencias de mood.gain del conjunto de datos Clinicaltrials.csv (Table 13.11)

Table 13.11: tabla de frecuencias de aumento del estado de ánimo a partir de los datos de Clinicaltrials.csv

0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.8**	**0.9**	**1.1**	**1.2**	**1.3**	**1.4**	**1.7**	**1.8**
1	1	2	1	1	2	1	1	1	1	2	2	1	1

Observando esta tabla, fíjate que la tercera entrada en la tabla de frecuencias tiene un valor de 2. Dado que esto corresponde a una ganancia de estado de ánimo de 0,3, esta tabla nos dice que el estado de ánimo de dos personas aumentó en 0,3. [16]

Y entonces jamovi usa un factor de corrección por empates para calcular el estadístico de Kruskall-Wallis corregido por empates. Y por fin hemos terminado con la teoría de la prueba de Kruskal-Wallis. Estoy segura de que estáis aliviadas de que os haya curado de la ansiedad existencial que surge naturalmente cuando os dais cuenta de que no sabeis cómo calcular el factor de corrección por empates para la prueba de Kruskal-Wallis. ¿Verdad?

13.6.8 Cómo ejecutar la prueba Kruskal-Wallis en jamovi

A pesar del horror por el que hemos pasado al tratar de entender lo que realmente hace la prueba Kruskal Wallis, resulta que ejecutar la prueba es bastante sencillo, ya que jamovi tiene un análisis como parte del conjunto de análisis ANOVA llamado 'No paramétrico' - 'ANOVA unifactorial (Kruskall-Wallis)' La mayoría de las veces tendrás datos como el conjunto de datos clinictrial.csv, en el que tienes tu variable de resultado mood.gain y una variable de agrupación de fármacos. Si es así, puedes continuar y ejecutar el análisis en jamovi. Lo que esto nos da es un Kruskal-Wallis $\chi^2 = 12.076, df = 2, p = 0.00239$, como en Figure 13.8

13.7 ANOVA unifactorial de medidas repetidas

La prueba ANOVA unifactorial de medidas repetidas es un método estadístico para probar diferencias significativas entre tres o más grupos donde se utilizan los mismos participantes en cada grupo (o cada participante se empareja con participantes en otros

[16] Más concretamente, en la notación matemática que introduje anteriormente, esto nos dice que $f_3 = 2$. Hurra. Entonces, ahora que sabemos esto, el factor de corrección por empates (FCE) es:

$$TCF = 1 - \frac{\sum_j f_j^3 - f_j}{N^3 - N}$$

El valor del estadístico de Kruskal-Wallis corregido por empates se obtiene dividiendo el valor de K por esta cantidad. Es esta versión corregida por empates la que calcula jamovi.

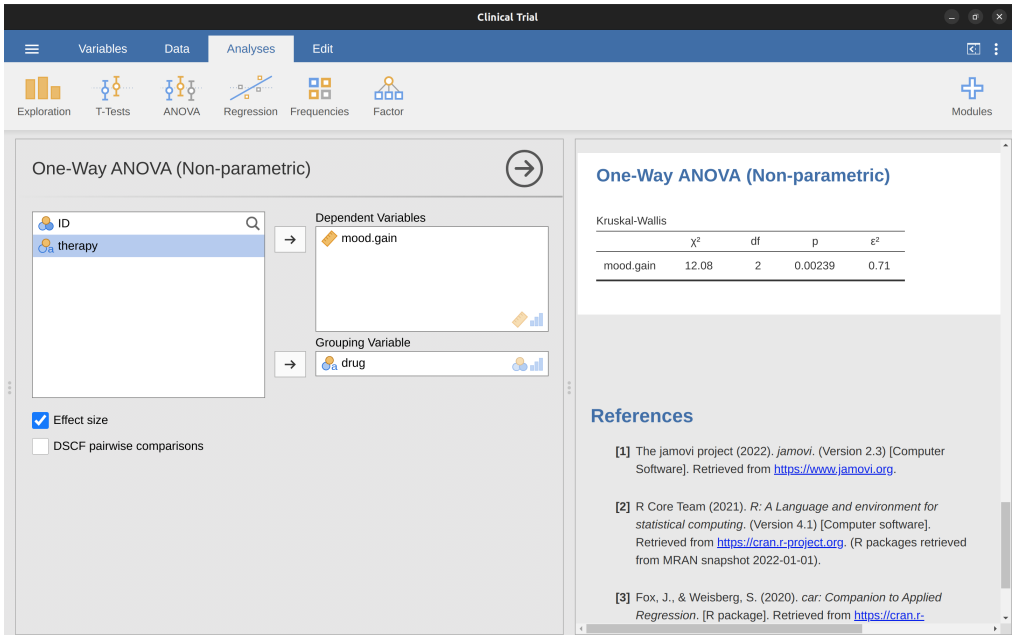


Figure 13.8: ANOVA no paramétrico unifactorial de Kruskall-Wallis en jamovi

grupos experimentales). Por esta razón, siempre debe haber un número igual de puntuaciones (datos) en cada grupo experimental. Este tipo de diseño y análisis también puede denominarse 'ANOVA relacionado' o 'ANOVA intrasujeto'.

La lógica que subyace a un ANOVA de medidas repetidas es muy similar a la de un ANOVA independiente (a veces llamado ANOVA 'entre sujetos'). Recordarás que anteriormente mostramos que en un ANOVA entre sujetos, la variabilidad total se divide en variabilidad entre grupos (SS_b) y variabilidad intra grupos (SS_w), y después de que cada uno se divide por los grados de libertad respectivos para dar MCe y MCi (ver Tabla 13.1) la F se calcula como:

$$F = \frac{MS_b}{MS_w}$$

En un ANOVA de medidas repetidas, la F se calcula de manera similar, pero mientras que en un ANOVA independiente la variabilidad dentro del grupo (SS_w) se usa como base para el denominador MS_w, en un ANOVA de medidas repetidas el SS_w se divide en dos partes. Como estamos usando los mismos sujetos en cada grupo, podemos eliminar la variabilidad debida a las diferencias individuales entre los sujetos (referidos como SCsujetos) de la variabilidad dentro de los grupos. No entraremos en demasiados detalles técnicos sobre cómo se hace esto, pero esencialmente cada sujeto se convierte en un nivel de un factor llamado sujetos. La variabilidad en este factor intra-sujetos se calcula entonces de la misma manera que cualquier factor entre sujetos. Y luego podemos restar SCsujetos de SS_w para proporcionar un término de SCerror más pequeño:

$$\text{ANOVA independiente: } SS_{error} = SS_w$$

$$\text{ANOVA de medidas repetidas: } SS_{error} = SS_w - SS_{sujetos}$$

Este cambio en el término SS_{error} a menudo conduce a una prueba estadística más potente, pero esto depende de si la reducción en el SS_{error} compensa la reducción en los grados de libertad del término de error (a medida que los grados de libertad van de $(n-k)$ [17] a $(n-1)(k-1)$ (teniendo en cuenta que hay más sujetos en el diseño ANOVA independiente).

13.7.1 ANOVA de medidas repetidas en jamovi

Primero, necesitamos algunos datos. Geschwind (1972) ha sugerido que la naturaleza exacta del déficit de lenguaje de un paciente después de un accidente cerebrovascular se puede utilizar para diagnosticar la región específica del cerebro que se ha dañado. Una investigadora está interesada en identificar las dificultades de comunicación específicas experimentadas por seis pacientes que padecen afasia de Broca (un déficit del lenguaje comúnmente experimentado después de un accidente cerebrovascular) (Table 13.12).

Table 13.12: Puntuaciones de tareas de reconocimiento de palabras en pacientes con accidente cerebrovascular

Participant	Speech	Conceptual	Syntax
1	8	7	6
2	7	8	6
3	9	5	3
4	5	4	5
5	6	6	2
6	8	7	4

Los pacientes debían completar tres tareas de reconocimiento de palabras. En la primera tarea (producción del habla), los pacientes debían repetir palabras sueltas leídas en voz alta por la investigadora. En la segunda tarea (conceptual), diseñada para evaluar la comprensión de palabras, los pacientes debían relacionar una serie de imágenes con su nombre correcto. En la tercera tarea (sintaxis), diseñada para evaluar el conocimiento del orden correcto de las palabras, se pidió a los pacientes que reordenaran oraciones sintácticamente incorrectas. Cada paciente completó las tres tareas. El orden en que los pacientes realizaron las tareas fue contrabalanceado entre los participantes. Cada tarea consistió en una serie de 10 intentos. El número de intentos completados con éxito por cada paciente se muestra en Table 13.11. Introduce estos datos en jamovi listos para el análisis (o coge un atajo y carga el archivo broca.csv).

Para realizar un ANOVA relacionado unifactorial en jamovi, abre el cuadro de diálogo ANOVA de medidas repetidas unifactoriales, como en Figure 13.9, a través de ANOVA - ANOVA de medidas repetidas.

Después:

[17](n - k) : (número de sujetos - número de grupos)

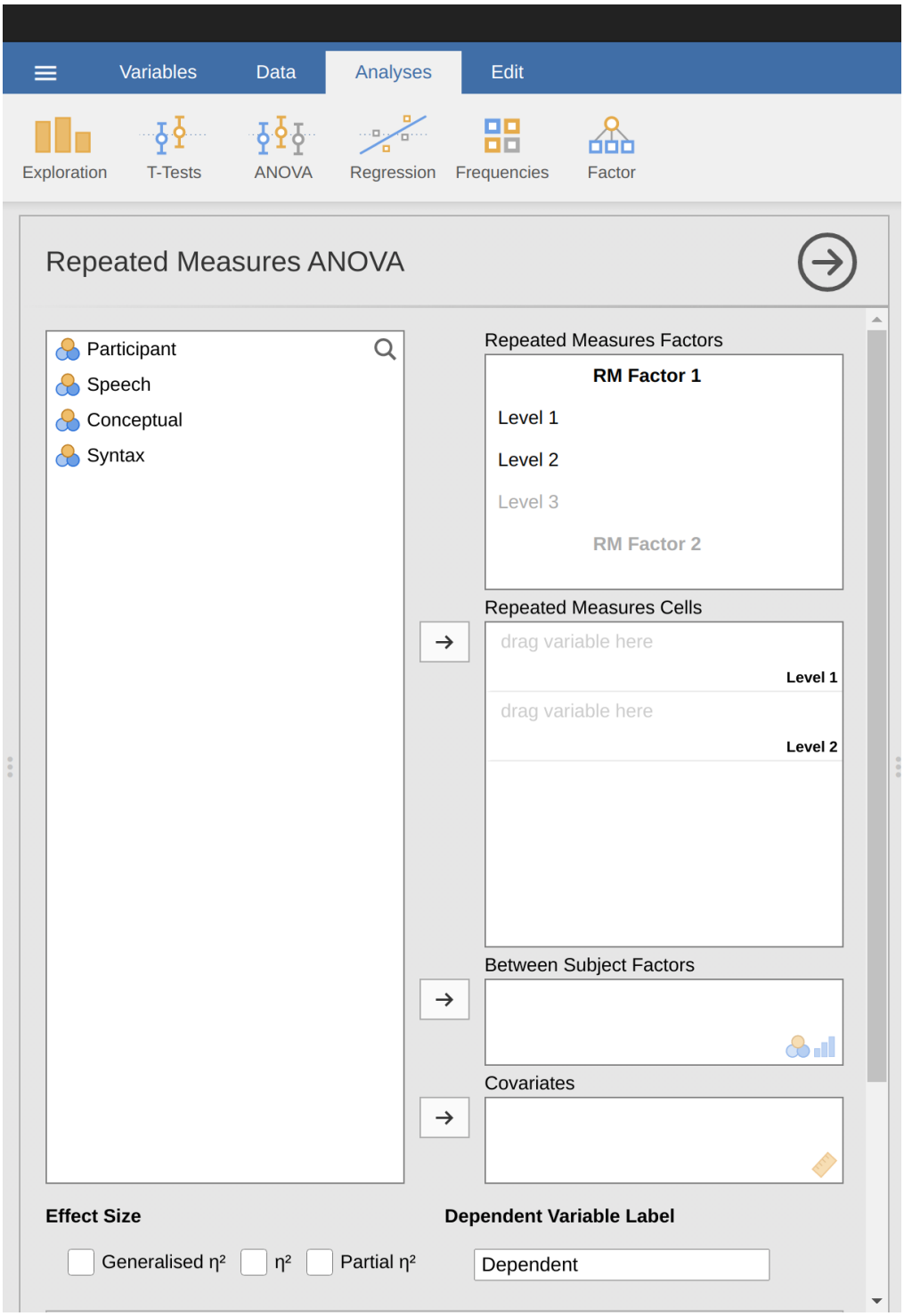


Figure 13.9: cuadro de diálogo ANOVA de medidas repetidas en jamovi

- Introduce un nombre de factor de medidas repetidas. Esta debe ser una etiqueta que elijas para describir las condiciones repetidas por todos los participantes. Por ejemplo, para describir las tareas de habla, conceptuales y sintácticas realizadas por todos los participantes, una etiqueta adecuada sería 'Tarea'. Ten en cuenta que este nuevo nombre del factor representa la variable independiente en el análisis.
- Agrega un tercer nivel en el cuadro de texto Factores de medidas repetidas, ya que hay tres niveles que representan las tres tareas: discurso, conceptual y sintaxis. Cambia las etiquetas de los niveles respectivamente.
- Luego, mueve cada uno de los niveles de las variables al cuadro de texto de la celda de medidas repetidas.
- Finalmente, en la opción Comprobaciones de supuestos, marca el cuadro de texto "Comprobaciones de esfericidad".

La salida jamovi para un ANOVA unifactorial de medidas repetidas se produce como se muestra en Figure 13.10 a Figure 13.13. El primer resultado que debemos observar es la prueba de esfericidad de Mauchly, que prueba la hipótesis de que las varianzas de las diferencias entre las condiciones son iguales (lo que significa que la dispersión de las puntuaciones de la diferencia entre las condiciones del estudio es aproximadamente la misma). En Figure 13.10, el nivel de significación de la prueba de Mauchly es $p = .720$. Si la prueba de Mauchly no es significativa (es decir, p > 0,05, como es el caso en este análisis), entonces es razonable concluir que las varianzas de las diferencias no son significativamente diferentes (es decir, son aproximadamente iguales y se puede asumir la esfericidad).

Assumptions

Tests of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Task	0.85	0.72009	0.87	1.00

Figure 13.10: Salida de ANOVA unifactorial de medidas repetidas - Prueba de esfericidad de Mauchly

Si por el contrario la prueba de Mauchly hubiera sido significativa (p < .05) entonces concluiríamos que existen diferencias significativas entre la varianza de las diferencias, y no se cumple el requisito de esfericidad. En este caso, deberíamos aplicar una corrección al valor F obtenido en el análisis ANOVA relacionado unifactorial:

- Si el valor de Greenhouse-Geisser en la tabla "Pruebas de esfericidad" es > 0,75, debes utilizar la corrección de Huynh-Feldt
- Pero si el valor de Greenhouse-Geisser es < .75, entonces debes usar la corrección de Greenhouse-Geisser.

Ambos valores F corregidos se pueden especificar en las casillas de verificación Correcciones de esfericidad en las opciones de Comprobaciones de supuestos, y los valores F

corregidos se muestran luego en la tabla de resultados, como en la Figura **13.11**.

Repeated Measures ANOVA

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
Task	None	24.78	2	12.39	6.93	0.01296
	Greenhouse-Geisser	24.78	1.74	14.26	6.93	0.01802
	Huynh-Feldt	24.78	2.00	12.39	6.93	0.01296
Residual	None	17.89	10	1.79		
	Greenhouse-Geisser	17.89	8.68	2.06		
	Huynh-Feldt	17.89	10.00	1.79		

Note. Type 3 Sums of Squares

Figure 13.11: Salida de ANOVA unifactorial de medidas repetidas - Pruebas de efectos intrasujetos

En nuestro análisis, vimos que la significación de la prueba de esfericidad de Mauchly fue p = .720 (es decir, p > 0.05). Por lo tanto, esto significa que podemos suponer que se ha cumplido el requisito de esfericidad, por lo que no es necesario corregir el valor F. Por lo tanto, podemos usar los valores de la corrección de esfericidad 'Ninguno' para la medida repetida 'Tarea': $F = 6.93$, $df = 2$, $p = .013$, y podemos concluir que el número de pruebas exitosas completado en cada tarea de lenguaje varió significativamente dependiendo de si la tarea se basaba en el habla, la comprensión o la sintaxis ($F(2, 10) = 6.93$, $p = .013$).

Las pruebas post-hoc también se pueden especificar en jamovi para ANOVA de medidas repetidas de la misma manera que para ANOVA independiente. Los resultados se muestran en Figure 13.12. Estos indican que existe una diferencia significativa entre Habla y Sintaxis, pero no entre otros niveles.

Post Hoc Tests

Post Hoc Comparisons - Task

Comparison			Mean Difference	SE	df	t	p_{tukey}
Task		Task					
Speech	-	Conceptual	1.00	0.68	5.00	1.46	0.38130
	-	Syntax	2.83	0.91	5.00	3.11	0.05814
Conceptual	-	Syntax	1.83	0.70	5.00	2.61	0.10245

Figure 13.12: Pruebas post-hoc en medidas repetidas ANOVA en jamovi

Los estadísticos descriptivos (medias marginales) se pueden revisar para ayudar a interpretar los resultados, producidos en la salida jamovi como en Figure 13.13. La compara-

ción del número medio de intentos completados con éxito por los participantes muestra que las personas con afasia de Broca se desempeñan razonablemente bien en las tareas de producción del habla (media = 7,17) y comprensión del lenguaje (media = 6,17). Sin embargo, su desempeño fue considerablemente peor en la tarea de sintaxis (media = 4.33), con una diferencia significativa en las pruebas post-hoc entre el desempeño de la tarea de habla y sintaxis.

Estimated Marginal Means - Task

			95% Confidence Interval	
Task	Mean	SE	Lower	Upper
Speech	7.17	0.60	5.62	8.71
Conceptual	6.17	0.60	4.62	7.71
Syntax	4.33	0.67	2.62	6.05

Figure 13.13: Salida de ANOVA de medidas repetidas unifactoriales - Estadísticos descriptivos

13.8 La prueba ANOVA no paramétrica de medidas repetidas de Friedman

La prueba de Friedman es una versión no paramétrica de un ANOVA de medidas repetidas y se puede usar en lugar de la prueba de Kruskall-Wallis cuando se prueban las diferencias entre tres o más grupos en los que los mismos participantes están en cada grupo, o cada participante está emparejado con participantes en otras condiciones. Si la variable dependiente es ordinal, o si no se cumple el supuesto de normalidad, se puede utilizar la prueba de Friedman.

Al igual que con la prueba de Kruskall-Wallis, las matemáticas subyacentes son complicadas y no se presentarán aquí. A los fines de este libro, es suficiente señalar que jamovi calcula la versión con corrección de empates de la prueba de Friedman, y en Figure 13.14 hay un ejemplo que usa los datos de Afasia de Broca que ya hemos visto.

Es bastante sencillo ejecutar una prueba de Friedman en jamovi. Simplemente selecciona Análisis - ANOVA - ANOVA de medidas repetidas (no paramétrico), como en Figure 13.14. Luego resalta y transfiere los nombres de las variables de medidas repetidas que deseas comparar (Habla, Conceptual, Sintaxis) al cuadro de texto 'Medidas:'. Para producir estadísticos descriptivos (medias y medianas) para las tres variables de medidas repetidas, haz clic en el botón Descriptivos.

Los resultados de jamovi muestran los estadísticos descriptivos, el valor de ji-cuadrado, los grados de libertad y el valor p (Figure 13.14). Dado que el valor p es menor que el nivel utilizado convencionalmente para determinar la importancia ($p < 0,05$), podemos

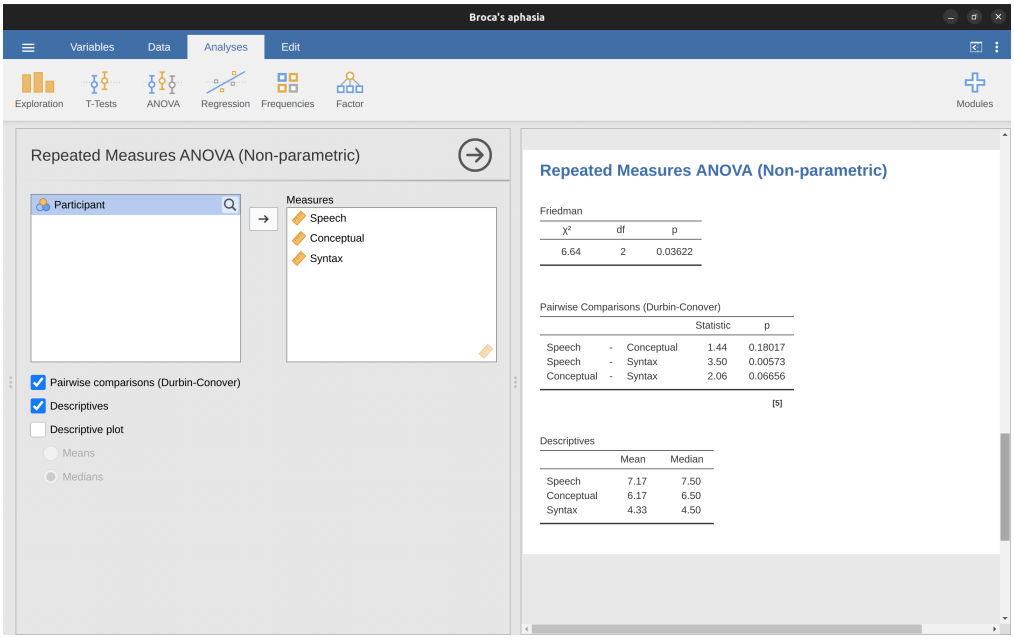


Figure 13.14: el cuadro de diálogo 'ANOVA de medidas repetidas (no paramétrico)' y los resultados en jamovi

concluir que los afásicos de Broca se desempeñan razonablemente bien en las tareas de producción del habla (mediana = 7,5) y comprensión del lenguaje (mediana = 6,5). Sin embargo, su desempeño fue considerablemente peor en la tarea de sintaxis (mediana = 4.5), con una diferencia significativa en las pruebas post-hoc entre el desempeño de la tarea de Habla y Sintaxis.

13.9 Sobre la relación entre ANOVA y la prueba t de Student

Hay una última cosa que quiero señalar antes de terminar. Es algo que mucha gente encuentra un poco sorprendente, pero vale la pena conocerlo. Un ANOVA con dos grupos es idéntico a la prueba t de Student. No es solo que sean similares, sino que en realidad son equivalentes en todos los aspectos relevantes. No intentaré demostrar que esto es cierto, pero os haré una sola demostración concreta. Supongamos que, en lugar de ejecutar un ANOVA en nuestro modelo de fármacos mood.gain ~, lo hacemos usando la terapia como predictor. Si ejecutamos este ANOVA obtenemos un estadístico F de $F(1, 16) = 1,71$ y un valor p $= 0,21$. Dado que solo tenemos dos grupos, en realidad no necesitábamos recurrir a un ANOVA, podríamos haber decidido ejecutar una prueba t de Student. Veamos qué sucede cuando hacemos esto: obtenemos un estadístico t de $t(16) = -1.3068$ y un valor de $p = 0.21$. Curiosamente, los valores de p son idénticos. Nuevamente obtenemos un valor de $p = .21$. Pero, ¿qué pasa con la prueba estadística? Después de ejecutar una prueba t en lugar de un ANOVA, obtenemos una respuesta algo diferente, a saber, $t(16) = -1,3068$. Sin embargo, la relación es bastante

sencilla. Si elevamos al cuadrado el estadístico t, obtenemos el estadístico F de antes: $-1.3068^2 = 1.7077$

13.10 Resumen

Hemos tratado bastante en este capítulo, pero aún falta mucho [18]. Obviamente, no he discutido cómo ejecutar un ANOVA cuando nos interesa más de una variable de agrupación, pero eso se discutirá con mucho detalle en Chapter 14. En términos de lo que hemos discutido, los temas clave fueron:

- La lógica básica que subyace a Cómo funciona ANOVA y [Ejecutar un ANOVA en jamovi]
- Cómo calcular un Tamaño del efecto para un ANOVA.
- Comparaciones múltiples y pruebas post hoc para pruebas múltiples.
- Los supuestos de ANOVA unifactorial
- [Comprobación del supuesto de homogeneidad de varianza] y qué hacer si se infringe: [Eliminación del supuesto de homogeneidad de varianza]
- Comprobación del supuesto de normalidad y qué hacer si se infringe: [Eliminación del supuesto de normalidad]
- ANOVA unifactorial de medidas repetidas y el equivalente no paramétrico, La prueba ANOVA no paramétrica de medidas repetidas de Friedman

[18] Al igual que con todos los capítulos de este libro, me he basado en fuentes diferentes, pero el texto destacado que más me ha influido Sahai & Ageel (2000). No es un libro para principiantes, pero es un libro excelente para lectores más avanzados con interés en comprender las matemáticas que subyacen a ANOVA.

Chapter 14

ANOVA factorial

En el transcurso de los últimos capítulos hemos hecho bastante. Hemos analizado las pruebas estadísticas que puedes usar cuando tienes una variable de predicción nominal con dos grupos (por ejemplo, la prueba t en Chapter 11) o con tres o más grupos (Chapter 13). Chapter 12 introdujo una idea nueva y potente, que consiste en crear modelos estadísticos con múltiples variables predictoras continuas que se usan para explicar una única variable de resultado. Por ejemplo, se podría usar un modelo de regresión para predecir la cantidad de errores que comete un estudiante en una prueba de comprensión lectora en función de la cantidad de horas que estudió para la prueba y su puntuación en una prueba estandarizada de CI.

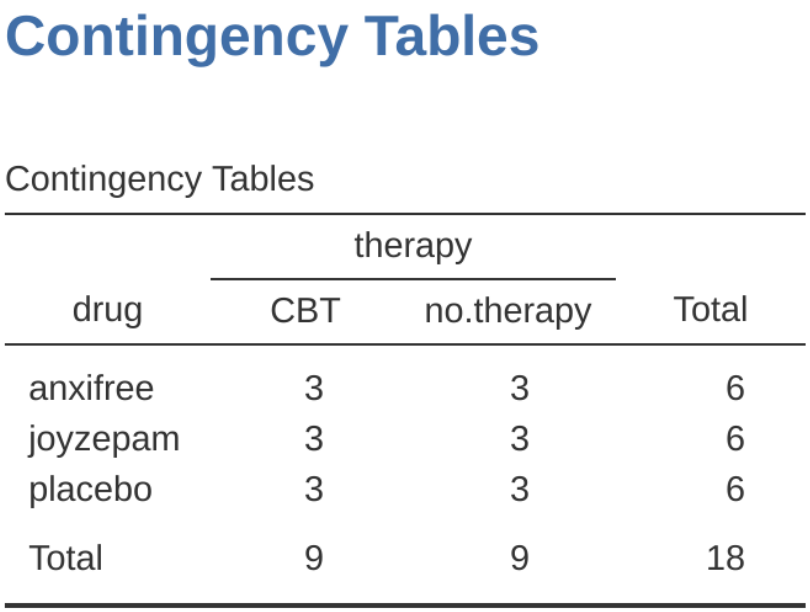
El objetivo de este capítulo es ampliar la idea de utilizar múltiples predictores en el marco ANOVA. Por ejemplo, supongamos que estamos interesadas en usar la prueba de comprensión lectora para medir los logros del alumnado en tres escuelas diferentes, y sospechamos que las niñas y los niños se están desarrollando a ritmos diferentes (y, por lo tanto, se espera que tengan un desempeño diferente en promedio). Cada estudiante se clasifica de dos maneras diferentes: en función de su género y en función de su escuela. Lo que nos gustaría hacer es analizar las puntuaciones de comprensión lectora en términos de estas dos variables de agrupación. La herramienta para hacerlo se denomina genéricamente **ANOVA factorial**. Sin embargo, dado que tenemos dos variables de agrupación, a veces nos referimos al análisis como un ANOVA de dos vías, en contraste con los ANOVA de una vía que ejecutamos en Chapter 13.

14.1 ANOVA factorial 1: diseños balanceados, centrados en los efectos principales

Cuando discutimos el análisis de varianza en Chapter 13, asumimos un diseño experimental bastante simple. Cada persona está en uno de varios grupos y queremos saber si estos grupos tienen puntuaciones medias diferentes en alguna variable de resultado. En esta sección, analizaré una clase más amplia de diseños experimentales conocidos como **diseños factoriales**, en los que tenemos más de una variable de agrupación. Di un ejemplo de cómo podría surgir este tipo de diseño arriba. Otro ejemplo aparece en Chapter 13 en el que estábamos viendo el efecto de diferentes fármacos en el estado de ánimo.ganancia experimentado por cada persona. En ese capítulo encontramos un

efecto significativo del fármaco, pero al final del capítulo también hicimos un análisis para ver si había un efecto de la terapia. No encontramos ninguno, pero hay algo un poco preocupante al tratar de ejecutar dos análisis separados para intentar predecir el mismo resultado. ¿Tal vez en realidad hay un efecto de la terapia sobre el aumento del estado de ánimo, pero no pudimos encontrarlo porque estaba "oculto" por el efecto del fármaco? En otras palabras, vamos a querer ejecutar un único análisis que incluya tanto el fármaco como la terapia como predictores. Para este análisis, cada persona se clasifica en forma cruzada según el fármaco que recibió (un factor con 3 niveles) y la terapia que recibió (un factor con 2 niveles). Nos referimos a esto como un diseño factorial de 3×2.

Si tabulamos de forma cruzada el fármaco por terapia, usando el análisis de 'Frecuencias' - 'Tablas de contingencia' en jamovi (ver Section 6.1), obtenemos la tabla que se muestra en Figure 14.1 .



Contingency Tables

Contingency Tables

	therapy		
drug	CBT	no.therapy	Total
anxifree	3	3	6
joyzepam	3	3	6
placebo	3	3	6
Total	9	9	18

Figure 14.1: tabla de contingencia jamovi de fármaco por tratamiento

Como puedes ver, no solo tenemos participantes correspondientes a todas las combinaciones posibles de los dos factores, lo que indica que nuestro diseño es **completamente cruzado**, resulta que hay un número igual de personas en cada grupo. En otras palabras, tenemos un diseño equilibrado. En esta sección explicaré cómo analizar datos de diseños **equilibrados**, ya que este es el caso más simple. La historia de los diseños desequilibrados es bastante tediosa, así que la dejaremos de lado por el momento.

14.1.1 ¿Qué hipótesis estamos probando?

Al igual que ANOVA unifactorial, ANOVA factorial es una herramienta para probar ciertos tipos de hipótesis sobre las medias de la población. Entonces, una buena forma de comenzar sería explicitar cuáles son realmente nuestras hipótesis. Sin embargo, antes de que podamos llegar a ese punto, es realmente útil tener una notación limpia y simple para describir las medias de la población. Dado que las observaciones se clasifican de forma cruzada en términos de dos factores diferentes, hay muchas medias diferentes en las que podríamos estar interesadas. Para ver esto, comencemos pensando en todas las diferentes medias muestrales que podemos calcular para este tipo de diseño. En primer lugar, está la idea obvia de que podríamos estar interesadas en esta lista de medias grupales (Table 14.1).

Table 14.1: Medias de grupo para grupos de fármacos y terapias en los datos de Clinicaltrial.csv

drug	therapy	mood.gain
placebo	no.therapy	0.30
anxifree	no.therapy	0.40
joyzepam	no.therapy	1.47
placebo	CBT	0.60
anxifree	CBT	1.03
joyzepam	CBT	1.50

Ahora, la siguiente tabla (Table 14.2) muestra una lista de las medias de los grupos para todas las combinaciones posibles de los dos factores (p. ej., personas que recibieron el placebo y ninguna terapia, personas que recibieron el placebo mientras recibían TCC, etc. .). Es útil organizar todos estos números, más las medias marginales y generales, en una sola tabla como esta:

Table 14.2: Medias de grupo y medias totales para los grupos de fármacos y terapias en los datos clintrial.csv

	no therapy	CBT	total
placebo	0.30	0.60	0.45
anxifree	0.40	1.03	0.72
joyzepam	1.47	1.50	1.48
total	0.72	1.04	0.88

Ahora bien, cada una de estas diferentes medias es, por supuesto, un estadístico muestral. Es una cantidad que pertenece a las observaciones específicas que hemos hecho durante nuestro estudio. Sobre lo que queremos hacer inferencias son los parámetros de población correspondientes. Es decir, las verdaderas medias tal como existen dentro de una población más amplia. Esas medias poblacionales también se pueden organizar en una tabla similar, pero necesitaremos un poco de notación matemática para hacerlo (Table 14.3). Como de costumbre, usaré el símbolo μ para indicar la media de una población. Sin embargo, debido a que hay muchas medias diferentes, tendré que usar subíndices para distinguirlas.

Así es como funciona la notación. Nuestra tabla se define en términos de dos factores. Cada fila corresponde a un nivel diferente del Factor A (en este caso, fármaco), y cada columna corresponde a un nivel diferente del Factor B (en este caso, terapia). Si dejamos que R indique el número de filas en la tabla y C indique el número de columnas, podemos referirnos a esto como un ANOVA factorial $R \times C$. En este caso $R = 3$ y $C = 2$. Usaremos letras minúsculas para referirnos a filas y columnas específicas, por lo que μ_{rc} se refiere a la media poblacional asociada con el nivel r-ésimo del Factor A (es decir, el número de fila r) y el c-ésimo nivel del Factor B (columna número c).[1] Entonces, las medias poblacionales ahora se escriben como en Table 14.1:

Table 14.3: Notación para medias poblacionales en una tabla factorial

	no therapy	**CBT**	**total**
placebo	μ_{11}	μ_{12}	
anxifree	μ_{21}	μ_{22}	
joyzepam	μ_{31}	μ_{32}	
total			

Bien, ¿qué pasa con las entradas restantes? Por ejemplo, ¿cómo deberíamos describir el aumento promedio del estado de ánimo en toda la población (hipotética) de personas que podrían recibir Joyzepam en un experimento como este, independientemente de si estaban en TCC? Usamos la notación "punto" para expresar esto. En el caso de Joyzepam, fíjate que estamos hablando de la media asociada con la tercera fila de la tabla. Es decir, estamos promediando las medias de dos celdas (es decir, μ_{31} y μ_{32}). El resultado de este promedio se denomina media marginal y se denotaría $\mu_{3.}$ en este caso. La **media marginal** para la TCC corresponde a la media poblacional asociada a la segunda columna de la tabla, por lo que usamos la notación porque es la media obtenida al promediar (marginalizar[2]) sobre ambas. Entonces, nuestra tabla completa de medias poblacionales se puede escribir como en Table 14.4.

Table 14.4: Notación para las medias poblacionales y totales en una tabla factorial

	no therapy	**CBT**	**total**
placebo	μ_{11}	μ_{12}	$\mu_{1.}$
anxifree	μ_{21}	μ_{22}	$\mu_{2.}$
joyzepam	μ_{31}	μ_{32}	$\mu_{3.}$
total	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

Ahora que tenemos esta notación, es sencillo formular y expresar algunas hipótesis.

[1] lo bueno de la notación de subíndices es que se generaliza muy bien. Si nuestro experimento hubiera involucrado un tercer factor, entonces podríamos simplemente agregar un tercer subíndice. En principio, la notación se extiende a tantos factores como desees incluir, pero en este libro rara vez consideraremos análisis que involucren más de dos factores y nunca más de tres.

[2] técnicamente, la marginalización no es exactamente idéntica a una media normal. Es un promedio ponderado en el que se tiene en cuenta la frecuencia de los diferentes eventos sobre los que se está promediando. Sin embargo, en un diseño equilibrado, todas las frecuencias de nuestras celdas son iguales por definición, por lo que las dos son equivalentes. Discutiremos los diseños desequilibrados más adelante, y cuando lo hagamos, verás que todos nuestros cálculos se convierten en un verdadero dolor de cabeza. Pero ignoremos esto por ahora.

Supongamos que el objetivo es averiguar dos cosas. Primero, ¿la elección del fármaco tiene algún efecto sobre el estado de ánimo? Y segundo, ¿la TCC tiene algún efecto sobre el estado de ánimo? Por supuesto, estas no son las únicas hipótesis que podríamos formular, y veremos un ejemplo realmente importante de un tipo diferente de hipótesis en la sección [ANOVA factorial 2: diseños balanceados, interacciones permitidas], pero estas son las dos hipótesis más simples para poner a prueba, así que empezaremos por ahí. Considera la primera prueba. Si el fármaco no tiene efecto entonces esperaríamos que todas las medias de la fila fueran idénticas, ¿verdad? Así que esa es nuestra hipótesis nula. Por otro lado, si el fármaco sí importa, deberíamos esperar que estas medias de fila sean diferentes. Formalmente, escribimos nuestras hipótesis nula y alternativa en términos de igualdad de medias marginales:

Hipótesis nula, H_0: las medias de las filas son las mismas, es decir, $\mu_{1.} = \mu_{2.} = \mu_{3.}$

Hipótesis alternativa, H_1: la media de al menos una fila es diferente

Vale la pena señalar que estas son exactamente las mismas hipótesis estadísticas que formamos cuando ejecutamos un ANOVA unifactorial en estos datos en Chapter 13. En aquel entonces, usé la notación $\mu \times P$ para referirme a la ganancia media en el estado de ánimo del grupo placebo, con μA y $\mu \times J$ correspondientes a las medias del grupo. para los dos fármacos, y la hipótesis nula fue $\mu P = \mu A = \mu J$. Entonces, en realidad estamos hablando de la misma hipótesis, solo que el ANOVA más complicado requiere una notación más cuidadosa debido a la presencia de múltiples variables de agrupación, por lo que ahora nos referimos a esta hipótesis como $\mu_{1.} = \mu_{2.} = \mu_{3.}$. Sin embargo, como veremos en breve, aunque la hipótesis es idéntica, la prueba de esa hipótesis es sutilmente diferente debido al hecho de que ahora estamos reconociendo la existencia de la segunda variable de agrupación.

Hablando de la otra variable de agrupación, no te sorprenderás al descubrir que nuestra segunda prueba de hipótesis está formulada de la misma manera. Sin embargo, dado que estamos hablando de terapia psicológica en lugar de fármacos, nuestra hipótesis nula ahora corresponde a la igualdad de las medias de la columna:

Hipótesis nula, H_0: las medias de las columnas son las mismas, es decir, $\mu_{.1} = \mu_{.2}$

Hipótesis alternativa, H_1: las medias de las columnas son diferentes, es decir, $\mu_{.1} \neq \mu_{.2}$

14.1.2 Ejecutando el análisis en jamovi

Las hipótesis nula y alternativa que describí en la última sección deberían parecer terriblemente familiares. Son básicamente las mismas que las hipótesis que estábamos probando en nuestros ANOVA unifactoriales más simples en Chapter 13. Por lo tanto, probablemente estés esperando que las pruebas de hipótesis que se utilizan en ANOVA factorial sean esencialmente las mismas que la prueba F de Chapter 13. Esperas ver referencias a sumas de cuadrados (SC), medias cuadráticas (MC), grados de libertad (gl) y, finalmente, un estadístico F que podemos convertir en un valor p, ¿verdad? Bueno, tienes toda la razón. Tanto es así que voy a apartarme de mi enfoque habitual. A lo

largo de este libro, generalmente he tomado el enfoque de describir la lógica (y hasta cierto punto las matemáticas) que sustentan un análisis particular primero y solo luego introducir el análisis en jamovi. Esta vez lo haré al revés y te mostraré cómo hacerlo primero en jamovi. La razón para hacer esto es que quiero resaltar las similitudes entre la herramienta ANOVA unifactorial simple que discutimos en Chapter 13, y el enfoque más complicado que vamos a usar en este capítulo.

Si los datos que estás tratando de analizar corresponden a un diseño factorial balanceado, entonces ejecutar tu análisis de varianza es fácil. Para ver lo fácil que es, comencemos reproduciendo el análisis original de Chapter 13. En caso de que lo hayas olvidado, para ese análisis usamos un solo factor (es decir, fármaco) para predecir nuestra variable de resultado (es decir, estado de ánimo.ganancia), y obtuvimos los resultados que se muestran en Figure 14.2.

ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η²
drug	3.45	2	1.73	18.61	0.00009	0.71
Residuals	1.39	15	0.09			

Figure 14.2: jamovi anova unifactorial de estado de ánimo.ganancia por fármaco

Ahora, supongamos que también tengo curiosidad por saber si la terapia tiene una relación con el aumento del estado de ánimo. A la luz de lo que hemos visto de nuestra discusión sobre la regresión múltiple en Chapter 12, probablemente no te sorprenda que todo lo que tenemos que hacer es agregar la terapia como un segundo 'Factor fijo' en el análisis, ver Figure 14.3.

ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η²	η²p	ω²
drug	3.45	2	1.73	31.71	< .001	0.71	0.84	0.68
therapy	0.47	1	0.47	8.58	0.013	0.10	0.42	0.08
drug ✻ therapy	0.27	2	0.14	2.49	0.125	0.06	0.29	0.03
Residuals	0.65	12	0.05					

Figure 14.3: jamovi bidireccional anova de mood.ganancia por fármacos y terapia

Esta salida es bastante simple de leer también. La primera fila de la tabla informa un valor de suma de cuadrados (SC) entre grupos asociado con el factor de fármaco, junto con un valor de gl entre grupos correspondiente. También calcula un valor de la media cuadrática (MC), un estadístico F y un valor p. También hay una fila que corresponde

al factor de terapia y una fila que corresponde a los residuales (es decir, la variación dentro de los grupos).

No solo todas las cantidades individuales son bastante familiares, sino que las relaciones entre estas diferentes cantidades se han mantenido sin cambios, tal como vimos con el ANOVA unifactorial original. Ten en cuenta que el valor de la media cuadrática se calcula dividiendo SS por el df correspondiente. Es decir, sigue siendo cierto que

$$MS = \frac{SS}{df}$$

independientemente de si estamos hablando de fármacos, terapia o los residuales. Para ver esto, no nos preocupemos por cómo se calculan los valores de las sumas de cuadrados. En su lugar, confiemos en que jamovi ha calculado correctamente los valores de SS e intentemos verificar que el resto de los números tengan sentido. Primero, ten en cuenta que para el factor de fármacos, dividimos 3.45 por 2 y terminamos con un valor de la media cuadrática de 1.73. Para el factor de terapia, solo hay 1 grado de libertad, por lo que nuestros cálculos son aún más simples: dividir 0.47 (el valor de SS) entre 1 nos da una respuesta de 0.47 (el valor de MS).

Volviendo a los estadísticos F y los valores p, fíjate que tenemos dos de cada uno; uno correspondiente al factor fármaco y otro correspondiente al factor terapia. Independientemente de cuál estemos hablando, el estadístico F se calcula dividiendo el valor de la media cuadrática asociado con el factor por el valor de la media cuadrática asociado con los residulaes. Si usamos "A" como notación abreviada para referirnos al primer factor (factor A; en este caso fármaco) y "R" como notación abreviada para referirnos a los residuales, entonces el estadístico F asociado con el factor A se denota como FA, y se calcula de la siguiente manera:

$$F_A = \frac{MS_A}{MS_R}$$

y existe una fórmula equivalente para el factor B (es decir, terapia). Ten en cuenta que este uso de "R" para referirse a los residuales es un poco incómodo, ya que también usamos la letra R para referirnos al número de filas en la tabla, pero solo voy a usar "R" para referirme a los residuales en el contexto de SCR y MCR, así que espero que esto no sea confuso. De todos modos, para aplicar esta fórmula al factor fármacos cogemos la media cuadrática de 1,73 y lo dividimos por el valor de la media cuadrática residual de $0,07$, lo que nos da un estadístico F de 26,15. El cálculo correspondiente para la variable de terapia sería dividir 0.47 por 0.07 lo que da 7.08 como estadístico F. Por supuesto, no sorprende que estos sean los mismos valores que jamovi ha informado en la tabla ANOVA anterior.

También en la tabla ANOVA está el cálculo de los valores de p. Una vez más, no hay nada nuevo aquí. Para cada uno de nuestros dos factores, lo que intentamos hacer es probar la hipótesis nula de que no existe una relación entre el factor y la variable de resultado (seré un poco más precisa sobre esto más adelante). Con ese fin, (aparentemente) hemos seguido una estrategia similar a la que hicimos en el ANOVA unifactorial y hemos calculado un estadístico F para cada una de estas hipótesis. Para convertirlos en valores p, todo lo que debemos hacer es observar que la distribución muestral para

el estadístico F bajo la hipótesis nula (el factor en cuestión es irrelevante) es una distribución F. También ten en cuenta que los valores de los dos grados de libertad son los correspondientes al factor y los correspondientes a los residuales. Para el factor de fármacos, estamos hablando de una distribución F con 2 y 14 grados de libertad (hablaré de los grados de libertad con más detalle más adelante). En cambio, para el factor de terapia la distribución muestral es F con 1 y 14 grados de libertad.

En este punto, espero que puedas ver que la tabla ANOVA para este análisis factorial más complicado debe leerse de la misma manera que la tabla ANOVA para el análisis unifactorial más simple. En resumen, nos dice que el ANOVA factorial para nuestro diseño de $3x2$ encontró un efecto significativo del fármaco ($F_{2,14} = 26,15, p < 0,001$), así como un efecto significativo de la terapia ($F_{1,14} = 7.08, p = .02$). O, para usar la terminología más técnicamente correcta, diríamos que hay dos **efectos principales** del fármaco y la terapia. Por el momento, probablemente parezca un poco redundante referirse a estos como efectos "principales", pero en realidad tiene sentido. Más adelante, vamos a querer hablar sobre la posibilidad de "interacciones" entre los dos factores, por lo que generalmente hacemos una distinción entre efectos principales y efectos de interacción.

14.1.3 ¿Cómo se calcula la suma de cuadrados?

En el apartado anterior tenía dos objetivos. En primer lugar, mostrarte que el método jamovi necesario para hacer ANOVA factorial es prácticamente el mismo que usamos para un ANOVA unifactorial. La única diferencia es la adición de un segundo factor. En segundo lugar, quería mostrarte cómo es la tabla ANOVA en este caso, para que puedas ver desde el principio que la lógica y la estructura básicas que subyacen al ANOVA factorial son las mismas que sustentan el ANOVA unifactorial. Trata de recordarlo. Es cierto, dado que el ANOVA factorial se construye más o menos de la misma manera que el ANOVA unifactorial más simple. Pero esta sensación de familiaridad comienza a evaporarse una vez que comienzas a profundizar en los detalles. Tradicionalmente, esta sensación de consuelo es reemplazada por un impulso de insultar a los autores de libros de texto de estadística.

Bien, comencemos revisando algunos de esos detalles. La explicación que di en la última sección ilustra el hecho de que las pruebas de hipótesis para los efectos principales (del fármaco y la terapia en este caso) son pruebas F, pero lo que no hace es mostrar cómo se calculan los valores de la suma de cuadrados (SC). Tampoco te dice explícitamente cómo calcular los grados de libertad (valores gl), aunque eso es algo simple en comparación. Supongamos por ahora que solo tenemos dos variables predictoras, Factor A y Factor B. Si usamos Y para referirnos a la variable de resultado, entonces usaríamos Yrci para referirnos al resultado asociado con el i-ésimo miembro del grupo rc (es decir, nivel/fila r para el Factor A y nivel/columna c para el Factor B). Por lo tanto, si usamos $\bar{Y}$ para referirnos a la media de una muestra, podemos usar la misma notación que antes para referirnos a las medias de grupo, medias marginales y medias generales. Es decir, $\bar{Y}_{rc}$ es la media muestral asociada al r-ésimo nivel del Factor A y al c-ésimo nivel del Factor: $\bar{Y}_{r.}$ sería la media marginal para el r-ésimo nivel del Factor A, $\bar{Y}_{.c}$ sería la media marginal para el c-ésimo nivel del Factor B, y $\bar{Y}_{..}$ es la media general. En otras palabras, nuestras medias muestrales se pueden organizar en la misma tabla que las medias poblacionales. Para los datos de nuestro ensayo clínico, esa tabla se muestra en Table 14.5.

Y si observamos las medios muestrales que presenté anteriormente, tenemos $\bar{Y}_{11} = 0,30$,

Table 14.5: Notación para medias muestrales para los datos de ensayos clínicos

	no therapy	**CBT**	**total**
placebo	$\bar{Y}_{11}$	$\bar{Y}_{12}$	$\bar{Y}_{1.}$
anxifree	$\bar{Y}_{21}$	$\bar{Y}_{22}$	$\bar{Y}_{2.}$
joyzepam	$\bar{Y}_{31}$	$\bar{Y}_{32}$	$\bar{Y}_{3.}$
total	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{..}$

$\bar{Y}_{12} = 0,60$, etc. En nuestro ejemplo del ensayo clínico, el factor de fármacos tiene 3 niveles y el factor de terapia tiene 2 niveles, entonces lo que estamos tratando de ejecutar es un ANOVA factorial de 3×2. Sin embargo, seremos un poco más generales y diremos que el Factor A (el factor de fila) tiene niveles R y el Factor B (el factor de columna) tiene C niveles, por tanto lo que estamos ejecutando aquí es $R ×C $ ANOVA factorial.

[Detalle técnico adicional [3]]

[3]Ahora que tenemos nuestra notación correcta, podemos calcular los valores de la suma de cuadrados para cada uno de los dos factores de una manera relativamente familiar. Para el Factor A, nuestra suma de cuadrados entre grupos se calcula evaluando hasta qué punto las medias marginales (fila) $\bar{Y}_{1.}, \bar{Y}_{2.}$, etc., son diferente de la media general $\bar{Y}_{..}$ Hacemos esto de la misma manera que lo hicimos para ANOVA unifactorial: calcula la suma de la diferencia al cuadrado entre los valores $\bar{Y}_{i.}$ y $\bar{Y}_{..}$. Específicamente, si hay N personas en cada grupo, entonces calculamos

$$SS_A = (N \times C) \sum_{r=1}^{R} (\bar{Y}_{r.} - \bar{Y}_{..})^2$$

Al igual que con ANOVA unifactorial, la parte ^{a} es la más interesante de esta fórmula, que corresponde a la desviación al cuadrado asociada con el nivel r. Lo que hace esta fórmula es calcular esta desviación al cuadrado para todos los niveles R del factor, sumarlos y luego multiplicar el resultado por $N \times C$. La razón de esta última parte es que hay múltiples celdas en nuestro diseño que tienen nivel r en el Factor A. De hecho, hay C de ellas, una correspondiente a cada nivel posible del Factor B. Por ejemplo, en nuestro ejemplo hay dos celdas diferentes en el diseño correspondientes al fármaco sin ansiedad: una para personas sin terapia y otra para el grupo de TCC. Y mo solo eso, dentro de cada una de estas celdas hay N observaciones. Entonces, si queremos convertir nuestro valor SC en una cantidad que calcule la suma de cuadrados entre grupos "por observación", tenemos que multiplicar por $N \times C$. La fórmula para el factor B es, por supuesto, la mismo, solo que con algunos subíndices mezclados

$$SS_B = (N \times R) \sum_{c=1}^{C} (\bar{Y}_{.c} - \bar{Y}_{..})^2$$

Ahora que tenemos estas fórmulas, podemos compararlas con la salida jamovi de la sección anterior. Una vez más, unade hoja de cálculo es útil para este tipo de cálculos, así que pruébalo tú misma. También puedes echarle un vistazo a la versión que hice en Excel en el archivo clinictrial_factorialanova.xls. Primero, calculemos la suma de cuadrados asociada con el efecto principal del fármaco. Hay un total de $N = 3$ personas en cada grupo y $C = 2$ diferentes tipos de terapia. O, dicho de otro modo, hay $3 \times 2 = 6$ personas que recibieron algún fármaco en particular. Cuando hacemos estos cálculos en una hoja de cálculo, obtenemos un valor de 3,45 para la suma de cuadrados asociada con el efecto principal del fármaco. No es sorprendente que este sea el mismo número que obtienes cuando buscas el valor SC para el factor de fármacos en la tabla ANOVA que presenté anteriormente, en Figure 14.3. Podemos repetir el mismo tipo de cálculo para el efecto de la terapia. Nuevamente, hay $N = 3$ personas en cada grupo, pero como hay $R = 3$ medicamentos diferentes, esta vez notamos que hay $3 \times 3 = 9$ personas que recibieron TCC y 9 personas adicionales que recibieron el placebo. Así que nuestro cálculo en este caso nos da un valor de 0.47 para la suma de cuadrados asociada con el efecto principal de la terapia. Una vez más, no nos sorprende ver que nuestros cálculos son idénticos a la salida de ANOVA en Figure 14.3. Así es como se calculan los valores SC para los dos efectos principales. Estos valores SC son análogos a los valores de suma de cuadrados entre grupos que calculamos al hacer ANOVA unifactorial en Chapter 13. Sin embargo, ya no es una buena idea pensar en ellos como valores SC entre grupos, porque tenemos dos variables de agrupación diferentes y es fácil confundirse. Sin embargo,

14.1.4 ¿Cuáles son nuestros grados de libertad?

Los grados de libertad se calculan de la misma manera que en el ANOVA unifactorial. Para cualquier factor dado, los grados de libertad son iguales al número de niveles menos 1 (es decir, $R - 1$ para la variable de fila Factor A y $C - 1$ para la variable de columna Factor B). Entonces, para el factor fármaco obtenemos $df = 2$, y para el factor de terapia obtenemos $df = 1$. Más adelante, cuando discutamos la interpretación de ANOVA como un modelo de regresión (ver Section 14.6), aclararé cómo llegamos a este número. Pero por el momento podemos usar la definición simple de grados de libertad, a saber, que los grados de libertad son iguales al número de cantidades que se observan, menos el número de restricciones. Entonces, para el factor fármaco, observamos 3 medias grupales separadas, pero están restringidas por 1 media general y, por lo tanto, los grados de libertad son 2. Para los residuales, la lógica es similar, pero no exactamente igual. El número total de observaciones en nuestro experimento es 18. Las restricciones corresponden a 1 media general, los 2 grupos adicionales significan que introduce el factor fármaco y 1 grupo adicional significa que introduce el factor terapia, por lo que nuestros grados de libertad son 14. Como fórmula, esto es $N - 1 - (R - 1) - (C - 1)$, que se simplifica a $N - R - C + 1$.

---

para construir una prueba F, también necesitamos calcular la suma de cuadrados dentro de los grupos. De acuerdo con la terminología que usamos en Chapter 12 y la terminología que jamovi usa al imprimir la tabla ANOVA, comenzaré a referirme al valor SC dentro de los grupos como la suma de cuadrados residual SC_R. Creo que la manera más fácil de pensar en los valores de la SC residual en este contexto es pensar en ello como la variación sobrante en la variable de resultado después de tener en cuenta las diferencias en las medias marginales (es decir, después de eliminar SC_A y SC_B). Lo que quiero decir con eso es que podemos comenzar calculando la suma de cuadrados total, que etiquetaré como SC_T. La fórmula para esto es más o menos la misma que para ANOVA unifactorial. Cogemos la diferencia entre cada observación Yrci y la media general $\hat{Y}_{..}$, elevamos al cuadrado las diferencias y las sumamos todas

$$SS_T = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{i=1}^{N} (Y_{rci} - \bar{Y}_{..})^2$$

La "suma triple" aquí parece más complicada de lo que es. En las dos primeras sumas, sumamos todos los niveles del Factor A (es decir, todas las filas r posibles de nuestra tabla) y todos los niveles del Factor B (es decir, todas las columnas posibles c). Cada combinación rc corresponde a un solo grupo y cada grupo contiene N personas, por lo que también tenemos que sumar todas esas personas (es decir, todos los valores de i). En otras palabras, todo lo que estamos haciendo aquí es sumar todas las observaciones en el conjunto de datos (es decir, todas las posibles combinaciones de rci). En este punto, conocemos la variabilidad total de la variable de resultado SCT y sabemos cuánto de esa variabilidad se puede atribuir al Factor A (SC_A) y cuánto se puede atribuir al Factor B (SC_B). La suma de cuadrados residual se define así como la variabilidad en Y que no se puede atribuir a ninguno de nuestros dos factores. En otras palabras,

$$SS_R = SS_T - (SS_A + SS_B)$$

Por supuesto, hay una fórmula que puedes usar para calcular la SC residual directamente, pero creo que tiene más sentido conceptual pensarlo así. El objetivo de llamarlo residual es que es la variación sobrante, y la fórmula anterior lo deja claro. También debo señalar que, de acuerdo con la terminología utilizada en el capítulo de regresión, es común referirse a $SC_A + SC_B$ como la varianza atribuible al "modelo ANOVA", denotado SCM, por lo que a menudo decimos que la suma de cuadrados total es igual a la suma de cuadrados modelo más la suma de cuadrados residual. Más adelante en este capítulo veremos que esto no es solo una similitud superficial: ANOVA y la regresión son en realidad lo mismo. En cualquier caso, probablemente valga la pena tomarse un momento para comprobar que podemos calcular SC_R usando esta fórmula y verificar que obtenemos la misma respuesta que produce jamovi en su tabla ANOVA. Los cálculos son bastante sencillos cuando se realizan en una hoja de cálculo (consulta el archivo clinictrial_factorialanova.xls). Podemos calcular la SC total usando las fórmulas anteriores (obteniendo una respuesta de $SCtotal = 4.85$) y luego la SC residual (= 0.92). Una vez más, obtenemos la misma respuesta. — [a]Traducción al inglés: "menos tedioso".

14.1.5 ANOVA factorial versus ANOVAs unifactoriales

Ahora que hemos visto cómo funciona un ANOVA factorial, vale la pena dedicar un momento para compararlo con los resultados de los análisis unifactoriales, porque esto nos mostrará por qué es una buena idea ejecutar el ANOVA factorial. En Chapter 13, ejecuté un ANOVA unifactorial para ver si había alguna diferencia entre los medicamentos y un segundo ANOVA unifactorial para ver si había alguna diferencia entre las terapias. Como vimos en la sección Section 14.1.1, las hipótesis nula y alternativa probadas por los ANOVA de una vía son de hecho idénticas a las hipótesis probadas por el ANOVA factorial. Mirando aún más detenidamente las tablas ANOVA, podemos ver que la suma de cuadrados asociada con los factores es idéntica en los dos análisis diferentes (3,45 para el fármaco y 0,92 para la terapia), al igual que los grados de libertad (2 para el fármaco, 1 para la terapia). ¡Pero no dan las mismas respuestas! En particular, cuando ejecutamos el ANOVA unifactorial para la terapia en Section 13.9 no encontramos un efecto significativo (el valor p fue .21). Sin embargo, cuando observamos el efecto principal de la terapia dentro del contexto del ANOVA de dos vías, obtenemos un efecto significativo (p = 0,019). Los dos análisis claramente no son lo mismo.

¿Por qué sucede eso? La respuesta está en comprender cómo se calculan los residuales. Recuerda que la idea que subyace a una prueba F es comparar la variabilidad que se puede atribuir a un factor en particular con la variabilidad que no se puede explicar (los residuales). Si ejecutas un ANOVA unifactorial para la terapia y, por lo tanto, ignoras el efecto del fármaco, ¡el ANOVA terminará volcando toda la variabilidad inducida por el fármaco en los residuales! Esto tiene el efecto de hacer que los datos parezcan más ruidosos de lo que realmente son, y el efecto de la terapia que se encontró correctamente significativo en el ANOVA de dos vías ahora se vuelve no significativo. Si ignoramos algo realmente importante (p. ej., un fármaco) cuando tratamos de evaluar la contribución de otra cosa (p. ej., una terapia), nuestro análisis se verá distorsionado. Por supuesto, está perfectamente bien ignorar las variables que son genuinamente irrelevantes para el fenómeno de interés. Si hubiéramos registrado el color de las paredes, y resultó ser un factor no significativo en un ANOVA de tres vías, estaría perfectamente bien ignorarlo e informar el ANOVA de dos vías más simple que no incluye este factor irrelevante. ¡Lo que no debes hacer es descartar variables que realmente marcan la diferencia!

14.1.6 ¿Qué tipo de resultados capta este análisis?

El modelo ANOVA del que hemos estado hablando hasta ahora cubre una variedad de patrones diferentes que podemos observar en nuestros datos. Por ejemplo, en un diseño ANOVA de dos vías hay cuatro posibilidades: (a) solo importa el factor A, (b) solo importa el factor B, (c) importan tanto A como B, y (d) ni A ni B importan. Un ejemplo de cada una de estas cuatro posibilidades se representa en Figure 14.4.

14.2 ANOVA factorial 2: diseños balanceados, interpretación de las interacciones

Los cuatro patrones de datos que se muestran en Figure 14.4 son bastante realistas. Hay una gran cantidad de conjuntos de datos que producen exactamente esos patrones. Sin embargo, no son todos los posibles y el modelo ANOVA del que hemos estado

comentando hasta este momento no es suficiente para explicar completamente una tabla de medias de grupo. ¿Por que no? Bueno, hasta ahora tenemos la capacidad de hablar sobre la idea de que los fármacos pueden influir en el estado de ánimo y la terapia puede influir en el estado de ánimo, pero no hay forma de saber si hay una **interacción** entre los dos. Se dice que ocurre una interacción entre A y B si el efecto del Factor A es *diferente*, según el nivel del Factor B del que estemos hablando. En Figure 14.5 se muestran varios ejemplos de un efecto de interacción en el contexto de un ANOVA de 2×2. Para dar un ejemplo más concreto, supongamos que el funcionamiento de Anxifree y Joyzepam se rige por mecanismos fisiológicos bastante diferentes. Una consecuencia de esto es que mientras que Joyzepam tiene más o menos el mismo efecto sobre el estado de ánimo independientemente de si uno está en terapia, Anxifree es en realidad mucho más eficaz cuando se administra junto con la TCC. El ANOVA que desarrollamos en la sección anterior no recoge esta idea. Para tener una idea de si realmente está ocurriendo una interacción aquí, es útil trazar las distintas medias de los grupos. En jamovi, esto se hace a través de la opción ANOVA 'Medias marginales estimadas': simplemente mueve el fármaco y la terapia al cuadro 'Medias marginales' debajo del 'Término 1'. Esto debería parecerse a Figure 14.6. Nuestra principal preocupación se relaciona con el hecho de que las dos líneas no son paralelas. El efecto de la TCC (diferencia entre la línea continua y la línea punteada) cuando el fármaco es Joyzepam (lado derecho) parece ser cercano a cero, incluso menor que el efecto de la TCC cuando se usa un placebo (lado izquierdo). Sin embargo, cuando se administra Anxifree, el efecto de la TCC es mayor que el del placebo (centro). ¿Este efecto es real o es solo una variación aleatoria debida al azar? ¡Nuestro ANOVA original no puede responder a esta pregunta, porque no tenemos en cuenta la idea de que las interacciones existen! En esta sección, solucionaremos este problema.

14.2.1 ¿Qué es exactamente un efecto de interacción?

La idea clave que vamos a presentar en esta sección es la de un efecto de interacción. En el modelo ANOVA que hemos visto hasta ahora, solo hay dos factores involucrados en nuestro modelo (es decir, el fármaco y la terapia). Pero cuando añadimos una interacción, añadimos un nuevo componente al modelo: la combinación de fármaco y terapia. Intuitivamente, la idea que subyace a un efecto de interacción es bastante sencilla. Simplemente significa que el efecto del Factor A es diferente, según el nivel del Factor B del que estemos hablando. Pero, ¿qué significa eso realmente en términos de nuestros datos? La trama en Figure 14.5 muestra varios patrones diferentes que, aunque son bastante diferentes entre sí, contarían como un efecto de interacción. Por lo tanto, no es del todo sencillo traducir esta idea cualitativa en algo matemático con lo que un estadístico pueda trabajar.

[Detalle técnico adicional [4]]

[4] Como consecuencia, la forma en que se formaliza la idea de un efecto de interacción en términos de hipótesis nula y alternativa es un poco difícil, y supongo que muchos de los lectores de este libro probablemente no están tan interesados. Aun así, intentaré ofrecer una idea básica. Para empezar, necesitamos ser un poco más explícitos acerca de nuestros efectos principales. Considera el efecto principal del Factor A (fármaco en nuestro ejemplo). Originalmente formulamos esto en términos de la hipótesis nula de que las dos medias marginales $\mu_{r.}$ son iguales entre si. Obviamente, si son iguales entre sí, entonces también deben ser iguales a la media general $\mu_{..}$, ¿verdad? Entonces, lo que podemos hacer es definir el efecto del Factor A en el nivel r para que sea igual a la diferencia entre la media marginal $\mu_{r.}$ y la media general $\mu_{..}$. Denotemos este efecto por α_r, y observemos que

$$\alpha_r = \mu_{r.} - \mu_{..}$$

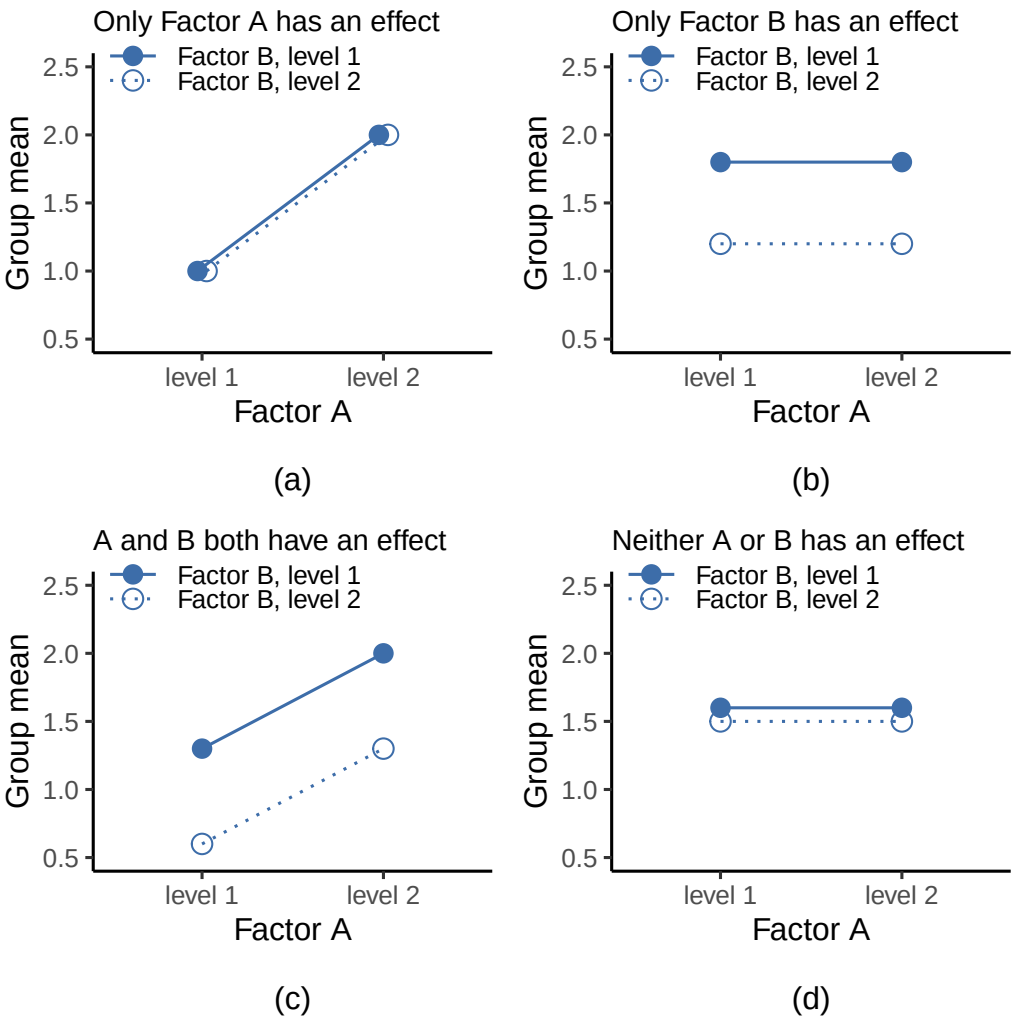


Figure 14.4: Los cuatro resultados diferentes para un ANOVA de 2×2 cuando no hay interacciones presentes. En el panel (a) vemos un efecto principal del Factor A y ningún efecto del Factor B. El panel (b) muestra un efecto principal del Factor B pero ningún efecto del Factor A. El panel (c) muestra los efectos principales tanto del Factor A como del Factor A. Finalmente, el panel (d) muestra ningún efecto de ninguno de los factores

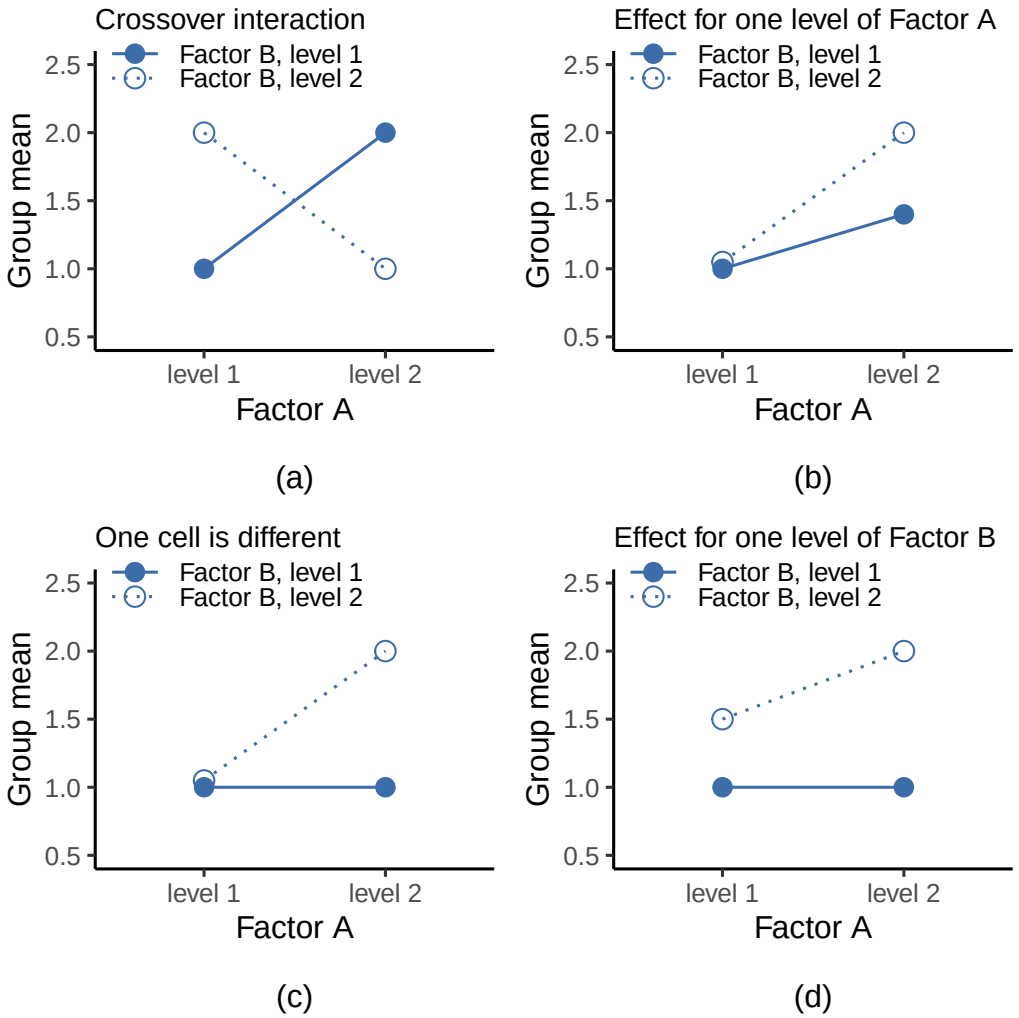


Figure 14.5: Interacciones cualitativamente diferentes para un ANOVA de 2×2

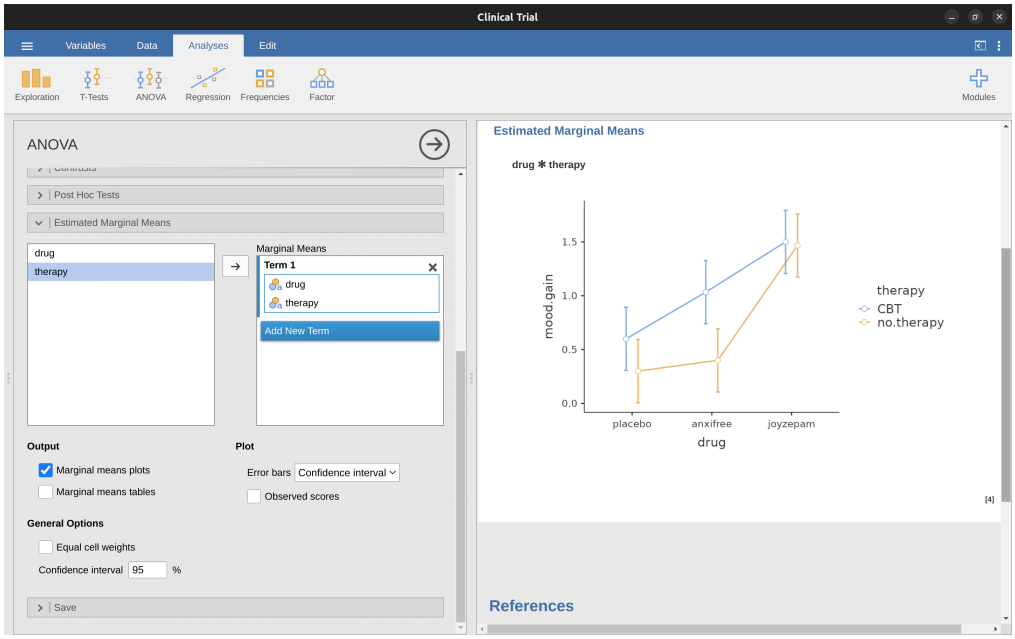


Figure 14.6: pantalla jamovi que muestra cómo generar un gráfico de interacción descriptivo en ANOVA utilizando los datos de ensayos clínicos

---

Ahora, por definición, todos los valores de α_r deben sumar cero, por la misma razón que el promedio de las medias marginales μ_c debe ser la media general $\mu_{..}$. De manera similar, podemos definir el efecto del Factor B en el nivel i como la diferencia entre la media marginal de la columna $\mu_{.c}$ y la media general $\mu_{..}$

$$\beta_c = \mu_{.c} - \mu_{..}$$

y una vez más, estos valores de β_c deben sumar cero. La razón por la que a veces a los estadísticos les gusta hablar de los efectos principales en términos de estos valores α_r y β_c es que les permite ser precisos sobre lo que significa decir que no hay efecto de interacción. Si no hay interacción en absoluto, entonces estos valores α_r y β_c describirán perfectamente las medias del grupo mu_{rc}. Específicamente, significa que

$$\mu_{rc} = \mu_{..} + \alpha_r + \beta_c$$

Es decir, no hay nada especial en las medias grupales que no pudieras predecir conociendo las medias marginales. Y ahí está nuestra hipótesis nula. La hipótesis alternativa es que

$$\mu_{rc} \neq \mu_{..} + \alpha_r + \beta_c$$

para al menos un grupo rc en nuestra tabla. Sin embargo, a los estadísticos a menudo les gusta escribir esto de manera ligeramente diferente. Por lo general, definirán la interacción específica asociada con el grupo rc como un número, torpemente denominado $(\alpha\beta)_{rc}$, y luego dirán que la hipótesis alternativa es que

$$\mu_{rc} = \mu_{..} + \alpha_r + \beta_c + (\alpha\beta)_{rc}$$

donde $(\alpha\beta)_{rc}$ es distinto de cero para al menos un grupo. Esta notación es un poco fea a la vista, pero es útil, como veremos cuando analicemos cómo calcular la suma de cuadrados. ¿Cómo debemos calcular la suma de cuadrados para los términos de interacción, $SS_{A:B}$? Bueno, en primer lugar, es útil notar cómo acabamos de definir el efecto de interacción en términos de en qué medida las medias grupales difieren de lo que esperarías mirando sólo las medias marginales. Por supuesto, todas esas fórmulas se refieren a parámetros poblacionales en lugar de estadísticas muestrales, por lo que en realidad no sabemos cuáles son. Sin embargo, podemos estimarlos usando medias muestrales en lugar de medias poblacionales. Entonces, para el Factor A, una buena manera de estimar el efecto principal en el nivel r es como la diferencia entre la media marginal muestral $\bar{Y}_{rc}$ y la media general muestral $\bar{Y}_{..}$ Es decir,

14.2.2 Grados de libertad para la interacción

Calcular los grados de libertad de la interacción es, una vez más, un poco más complicado que el cálculo correspondiente de los efectos principales. Para empezar, pensemos en el modelo ANOVA como un todo. Una vez que incluimos los efectos de interacción en el modelo, permitimos que cada grupo tenga una media única, mu_{rc}. Para un ANOVA factorial de $R \times C$, esto significa que hay cantidades $R \times C$ de interés en el modelo y solo una restricción: todas las medias de grupo deben promediar la media general. Entonces, el modelo como un todo necesita tener $(R \times C)$ - 1 grado de libertad. Pero el efecto principal del Factor A tiene $R-1$ grados de libertad, y el efecto principal del Factor B tiene $C-1$ grados de libertad. Esto significa que los grados de libertad asociados con la interacción son

$$\begin{aligned} df_{A:B} &= (R \times C - 1) - (R-1) - (C-1) \\ &= RC - R - C + 1 \\ &= (R-1)(C-1) \end{aligned}$$

que es simplemente el producto de los grados de libertad asociados con el factor de fila

---

usaríamos esto como nuestra estimación del efecto

$$\hat{\alpha}_r = bar Y_{r.} - \bar{Y}_{..}$$

De manera similar, nuestra estimación del efecto principal del Factor B en el nivel c se puede definir de la siguiente manera

$$\hat{\beta}_c = \hat{Y}_{.c} - \bar{Y}_{..}$$

Ahora, si vuelves a las fórmulas que usé para describir los valores de SC para los dos efectos principales, notarás que estos términos de efectos son exactamente las cantidades que estábamos elevando al cuadrado y sumando. Entonces, ¿cuál es el análogo de esto para los términos de interacción? La respuesta a esto la podemos encontrar primero reorganizando la fórmula para las medias grupales μ_{rc} bajo la hipótesis alternativa, de modo que obtengamos

$$\begin{aligned} (\alpha\beta)_{rc} &= \mu_{rc} - \mu_{..} - \alpha_r - \beta_c \\ &= \mu_{rc} - \mu_{..} - (\mu_{r.} - \mu_{..}) - (\mu_{.c} - \mu_{..}) \\ &= \mu_{rc} - \mu_{r.} - \mu_{.c} + \mu_{..} \end{aligned}$$

Entonces , una vez más, si sustituimos nuestros estadísticos muestrales en lugar de las medias poblacionales, obtenemos lo siguiente como nuestra estimación del efecto de interacción para el grupo rc, que es

$$\hat{(\alpha\beta)}_{rc} = \bar{Y}_{rc} - \hat{Y}_{r.} - \bar{Y}_{.c} + \bar{Y}_{..}$$

Ahora lo que tenemos hacer es sumar todas estas estimaciones en todos los niveles de R del Factor A y todos los niveles de C del Factor B, y obtenemos la siguiente fórmula para la suma de cuadrados asociados con la interacción como un todo

$$SS_{A:B} = N \sum_{r=1}^{R} \sum_{c=1}^{C} (\bar{Y}_{rc} - \bar{Y}_{r.} - \ bar Y_{.c} + \bar{Y}_{..})^2$$

donde multiplicamos por N porque hay N observaciones en cada uno de los grupos, y queremos que nuestros valores SC reflejen la variación entre observaciones explicada por la interacción, no la variación entre grupos. Ahora que tenemos una fórmula para calcular $SS_{A:B}$, es importante reconocer que el término de interacción es parte del modelo (por supuesto), por lo que la suma de cuadrados total asociada con el modelo, SCM, ahora es igual a la suma de los tres valores SC relevantes, $SC_A + SC_B + SC_{A:B}$. La suma de cuadrados residual SCR se define como la variación sobrante, a saber, $SC_T - SC_M$, pero ahora que tenemos el término de interacción, se convierte en

$$SS_R = SS_T - (SS_A + SS_B + SS_{A:B})$$

Como consecuencia, la suma de cuadrados residual SS_R será menor que en nuestro ANOVA original que no incluía interacciones.

y el factor de columna.

¿Qué pasa con los grados de libertad residuales? Debido a que hemos agregado términos de interacción que absorben algunos grados de libertad, quedan menos grados de libertad residuales. Específicamente, ten en cuenta que si el modelo con interacción tiene un total de $(R \times C) - 1$, y hay N observaciones en su conjunto de datos que están restringidas para satisfacer 1 media general, tus grados de libertad residuales ahora se convierten en $N - (R \times C) - 1 + 1$, o simplemente $N - (R \times C)$.

14.2.3 Ejecución del ANOVA en jamovi

Agregar términos de interacción al modelo ANOVA en jamovi es sencillo. De hecho, es más que sencillo porque es la opción predeterminada para ANOVA. Esto significa que cuando especificas un ANOVA con dos factores, por ejemplo, fármaco y terapia, el componente de interacción (fármaco × terapia) se agrega automáticamente al modelo [5]. Cuando ejecutamos el ANOVA con el término de interacción incluido, obtenemos los resultados que se muestran en Figure 14.7.

ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η²	η²p	ω²
drug	3.45	2	1.73	31.71	0.00002	0.71	0.84	0.68
therapy	0.47	1	0.47	8.58	0.01262	0.10	0.42	0.08
drug ✻ therapy	0.27	2	0.14	2.49	0.12460	0.06	0.29	0.03
Residuals	0.65	12	0.05					

Figure 14.7: Resultados del modelo factorial completo, incluido el componente de interacción fármaco × terapia

Resulta que, aunque tenemos un efecto principal significativo del fármaco ($F_{2,12} = 31,7, p < 0,001$) y el tipo de terapia ($F_{1,12} = 8,6$, p = 0,013), *no hay una interaccin significativa entre los dos* $F_{2,12} = 2,5$, p = 0,125$).

14.2.4 Interpretación de los resultados

Hay un par de cosas muy importantes a considerar al interpretar los resultados del ANOVA factorial. En primer lugar, está el mismo problema que tuvimos con ANOVA unifactorial, que es que si obtienes un efecto principal significativo de (digamos) fármaco, no dice nada sobre qué fármacos son diferentes entre sí. Para averiguarlo, debes realizar análisis adicionales. Hablaremos de algunos análisis que puedes ejecutar en secciones posteriores: Diferentes formas de especificar contrastes y Pruebas post hoc. Lo mismo sucede con los efectos de interacción. Saber que hay una interacción significativa no dice nada sobre qué tipo de interacción existe. Una vez más, deberás ejecutar análisis adicionales.

[5] Es posible que ya hayas notado esto al mirar el análisis de efectos principales en jamovi que describimos anteriormente. Para el propósito de la explicación en este libro, eliminé el componente de interacción del modelo anterior para mantener las cosas limpias y sencillas.

En segundo lugar, existe un problema de interpretación muy peculiar que surge cuando se obtiene un efecto de interacción significativo pero no un efecto principal correspondiente. Esto sucede a veces. Por ejemplo, en la interacción cruzada que se muestra en Figure 14.5 a, esto es exactamente lo que encontrarías. En este caso, ninguno de los efectos principales sería significativo, pero el efecto de interacción sí lo sería. Esta es una situación difícil de interpretar, y la gente a menudo se confunde un poco al respecto. El consejo general que les gusta dar a los estadísticos en esta situación es que no debes prestar mucha atención a los efectos principales cuando hay una interacción. La razón por la que dicen esto es que, aunque las pruebas de los efectos principales son perfectamente válidas desde un punto de vista matemático, cuando hay un efecto de interacción significativo, los efectos principales rara vez prueban hipótesis interesantes. Recuerda de Section 14.1.1 que la hipótesis nula para un efecto principal es que las medias marginales son iguales entre sí, y que una media marginal se forma promediando varios grupos diferentes. Pero si tienes un efecto de interacción significativo, entonces sabes que los grupos que componen la media marginal no son homogéneos, por lo que no está claro por qué te interesarían esas medias marginales.

Esto es lo que quiero decir. Una vez más, sigamos con un ejemplo clínico. Supongamos que tuviéramos un diseño de 2×2 que comparara dos tratamientos diferentes para las fobias (p. ej., desensibilización sistemática frente a inundación) y dos fármacos diferentes para reducir la ansiedad (p. ej., Anxifree frente a Joyzepam). Ahora, supongamos que descubrimos que Anxifree no tuvo efecto cuando el tratamiento fue la desensibilización, y Joyzepam no tuvo efecto cuando el tratamiento fue la inundación. Pero ambos fueron bastante efectivos para el otro tratamiento. Esta es una interacción cruzada clásica, y lo que encontraríamos al ejecutar el ANOVA es que no hay un efecto principal del fármaco, sino una interacción significativa. Ahora bien, ¿qué significa realmente decir que no hay un efecto principal? Bueno, significa que si promediamos los dos tratamientos psicológicos diferentes, entonces el efecto promedio de Anxifree y Joyzepam es el mismo. Pero, ¿por qué a alguien le interesaría eso? Cuando se trata a alguien por fobias, nunca se da el caso de que una persona pueda ser tratada usando un "promedio" de inundación y desensibilización. Eso no tiene mucho sentido. O te quedas con uno o con el otro. Para un tratamiento, un fármaco es eficaz y para el otro tratamiento, el otro fármaco es eficaz. La interacción es lo importante y el efecto principal es algo irrelevante.

Este tipo de cosas suceden a menudo. El efecto principal son las pruebas de las medias marginales, y cuando hay una interacción, a menudo nos damos cuenta de que no estamos muy interesados en las medias marginales porque implican promediar cosas que la interacción nos dice que no deben promediarse. Por supuesto, no siempre es el caso de que un efecto principal no tenga sentido cuando hay una interacción presente. A menudo, puedes obtener un gran efecto principal y una interacción muy pequeña, en cuyo caso aún puedes decir cosas como "el fármaco A es generalmente más efectivo que el fármaco B" (porque hubo un gran efecto del fármaco), pero necesitarías modificarlo un poco agregando que "la diferencia de efectividad fue diferente para diferentes tratamientos psicológicos". En cualquier caso, el punto principal aquí es que cada vez que obtengas una interacción significativa, debes detenerte y pensar qué significa realmente el efecto principal en este contexto. No asumas automáticamente que el efecto principal es interesante.

14.3 Tamaño del efecto

El cálculo del tamaño del efecto para un ANOVA factorial es bastante similar a lo que se utiliza en el ANOVA unidireccional (consulta la sección Tamaño del efecto). Específicamente, podemos usar η^2 (eta-cuadrado) como una forma simple de medir qué tan grande es el efecto general para cualquier término en particular. Como antes, η^2 se define dividiendo la suma de cuadrados asociada con ese término por la suma de cuadrados total. Por ejemplo, para determinar el tamaño del efecto principal del Factor A, usaríamos la siguiente fórmula:

$$\eta_A^2 = \frac{SS_A}{SS_T}$$

Como antes, esto se puede interpretar de la misma manera que R^2 en regresión.[6] Indica la proporción de varianza en la variable de resultado que se puede explicar por el efecto principal de Factor A. Por lo tanto, es un número que va de 0 (ningún efecto) a 1 (considera toda la variabilidad en el resultado). Además, la suma de todos los valores de η^2, cogidos de todos los términos del modelo, sumará el total de R^2 para el modelo ANOVA. Si, por ejemplo, el modelo ANOVA se ajusta perfectamente (es decir, ¡no hay ninguna variabilidad dentro de los grupos!), los valores de η^2 sumarán 1. Por supuesto, eso rara vez sucede en la vida real.

Sin embargo, al hacer un ANOVA factorial, hay una segunda medida del tamaño del efecto que a la gente le gusta informar, conocida como η^2 parcial. La idea que subyace a η^2 parcial (que a veces se denomina p^{η^2} o η_p^2) es que, al medir el tamaño del efecto para un término en particular (digamos, el efecto principal del Factor A), deseas ignorar deliberadamente los otros efectos en el modelo (por ejemplo, el efecto principal del Factor B). Es decir, supondrías que el efecto de todos estos otros términos es cero y luego calcularías cuál habría sido el valor de η^2. En realidad, esto es bastante fácil de calcular. Todo lo que tienes que hacer es quitar la suma de cuadrados asociada con los otros términos del denominador. En otras palabras, si deseas el η^2 parcial para el efecto principal del Factor A, el denominador es solo la suma de los valores de SC para el Factor A y los residuales

$$\text{parcial}\eta_A^2 = \frac{SS_A}{SS_A + SS_R}$$

Esto siempre te dará un número mayor que η^2, que la cínica en mí sospecha que explica la popularidad de η^2 parcial. Y una vez más obtienes un número entre 0 y 1, donde 0 representa ningún efecto. Sin embargo, es un poco más complicado interpretar lo que significa un gran valor de η^2 parcial. En particular, ¡no puedes comparar los valores de η^2 parcial entre términos! Supongamos, por ejemplo, que no hay ninguna variabilidad dentro de los grupos: si es así, $SC_R = 0$. Lo que eso significa es que cada término tiene un valor de η^2 parcial de 1. Pero eso no significa que todos los términos en tu modelo sean igualmente importantes, o que sean igualmente grandes. Todo lo que significa

[6] este capítulo parece estar estableciendo un nuevo récord por la cantidad de cosas diferentes que puede representar la letra R. Hasta ahora tenemos R refiriéndose al paquete de software, el número de filas en nuestra tabla de medias, los residuales en el modelo y ahora el coeficiente de correlación en una regresión. Lo siento. Claramente no tenemos suficientes letras en el alfabeto. Sin embargo, me he esforzado mucho para dejar claro a qué se refiere R en cada caso.

es que todos los términos en tu modelo tienen tamaños de efecto que son grandes en relación con la variación residual. No es comparable entre términos.

Para ver lo que quiero decir con esto, es útil ver un ejemplo concreto. Primero, echemos un vistazo a los tamaños del efecto para el ANOVA original (Table 14.6) sin el término de interacción, de Figure 14.3.

Table 14.6: tamaños del efecto cuando el término de interacción **no** está incluido en el modelo ANOVA

	eta.sq	**partial.eta.sq**
drug	0.71	0.79
therapy	0.10	0.34

Mirando primero los valores de η^2, vemos que el fármaco representa el 71 % de la varianza (es decir, $\eta^2 = 0,71$) en el aumento del estado de ánimo, mientras que la terapia solo representa el 10 %. Esto deja un total de 19% de la variación sin contabilizar (es decir, los residuales constituyen el 19% de la variación en el resultado). En general, esto implica que tenemos un efecto muy grande [7] del fármaco y un efecto modesto de la terapia.

Ahora veamos los valores de η^2 parcial, que se muestran en Figure 14.3. Debido a que el efecto de la terapia no es tan grande, controlarlo no genera mucha diferencia, por lo que el η^2 parcial para el fármaco no aumenta mucho y obtenemos un valor de $p^{\eta^2} = 0,79$. Por el contrario, debido a que el efecto del fármaco fue muy grande, controlarlo provoca una gran diferencia, por lo que cuando calculamos el η^2 parcial para la terapia, puedes ver que aumenta a $\$p\{ 2 \} = 0{,}34$ \$. La pregunta que tenemos que hacernos es, ¿qué significan realmente estos valores de η^2 parcial? La forma en que generalmente interpreto el η^2 parcial para el efecto principal del Factor A es interpretarlo como una declaración sobre un experimento hipotético en el que solo se varió el Factor A. Así, aunque en este experimento variamos tanto A como B, podemos imaginar fácilmente un experimento en el que solo se varió el Factor A, y el estadístico η^2 parcial te dice cuánto de la varianza en la variable de resultado esperarías ver contabilizado en ese experimento. Sin embargo, debes tenerse en cuenta que esta interpretación, como muchas cosas asociadas con los efectos principales, no tiene mucho sentido cuando hay un efecto de interacción grande y significativo.

Hablando de efectos de interacción, Table 14.7 muestra lo que obtenemos cuando calculamos los tamaños del efecto para el modelo que incluye el término de interacción, como en Figure 14.7. Como puedes ver, los valores de η^2 para los efectos principales no cambian, pero los valores de η^2 parcial sí:

14.3.1 Medias estimadas de los grupos

En muchas situaciones, querrás estimar todas las medias de los grupos en función de los resultados de tu ANOVA, así como los intervalos de confianza asociados con ellos. Puedes usar la opción 'Medias marginales estimadas' en el análisis ANOVA de jamovi para hacer esto, como en Figure 14.8. Si el ANOVA que ejecutaste es un modelo

[7] Inverosímilmente grande, creo. ¡La artificialidad de este conjunto de datos realmente está comenzando a mostrarse!

Table 14.7: tamaños del efecto cuando el término de interacción **se** incluye en el modelo ANOVA

	eta.sq	partial.eta.sq
drug	0.71	0.84
therapy	0.10	0.42
drug*therapy	0.06	0.29

saturado (es decir, contiene todos los efectos principales posibles y todos los efectos de interacción posibles), las estimaciones de las medias de los grupos son en realidad idénticas a las medias muestrales, aunque los intervalos de confianza utilizarán una estimación combinada de los errores estándar en lugar de utilizar uno para cada grupo.

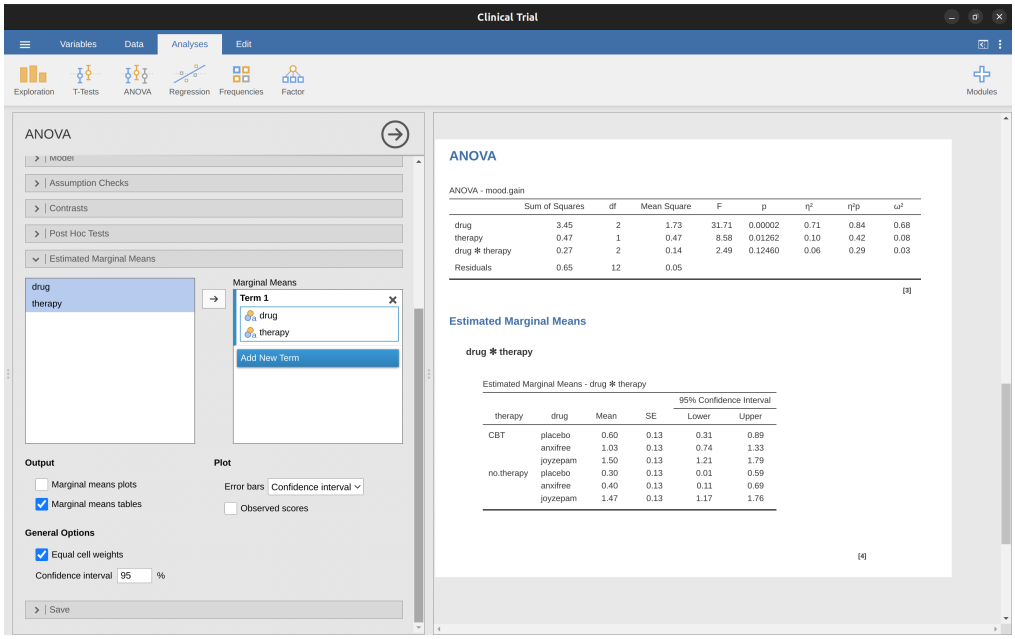


Figure 14.8: captura de pantalla de jamovi que muestra las medias marginales para el modelo saturado, es decir, incluido el componente de interacción, con el conjunto de datos del ensayo clínico

En el resultado, vemos que la mejora media estimada del estado de ánimo para el grupo de placebo sin terapia fue de $0,300$, con un intervalo de confianza de 95% de $0,006$ a $0,594$. Ten en cuenta que estos no son los mismos intervalos de confianza que obtendrías si los calcularas por separado para cada grupo, debido al hecho de que el modelo ANOVA asume la homogeneidad de la varianza y, por lo tanto, utiliza una estimación combinada de la desviación estándar.

Cuando el modelo no contiene el término de interacción, las medias estimadas del grupo serán diferentes de las medias muestrales. En lugar de informar la media muestral, jamovi calculará el valor de las medias del grupo que se esperaría sobre la base de las

medias marginales (es decir, suponiendo que no hay interacción). Usando la notación que desarrollamos anteriormente, la estimación informada para μ_{rc}, la media para el nivel r en el Factor A (fila) y el nivel c en el Factor B (columna) sería $ _{..} + _r + _c ps Si realmente no hay interacciones entre los dos factores, esta es en realidad una mejor estimación de la media poblacional que la media muestral sin procesar. Eliminar el término de interacción del modelo, a través de las opciones 'Modelo' en el análisis ANOVA de jamovi, proporciona las medias marginales para el análisis que se muestra en Figure 14.9.

ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η²	η²p	ω²
drug	3.45	2	1.73	26.15	0.00002	0.71	0.79	0.68
therapy	0.47	1	0.47	7.08	0.01866	0.10	0.34	0.08
Residuals	0.92	14	0.07					

[3]

Estimated Marginal Means

drug ✻ therapy

Estimated Marginal Means - drug ✻ therapy

				95% Confidence Interval	
therapy	drug	Mean	SE	Lower	Upper
CBT	placebo	0.61	0.12	0.35	0.87
	anxifree	0.88	0.12	0.62	1.14
	joyzepam	1.64	0.12	1.38	1.90
no.therapy	placebo	0.29	0.12	0.03	0.55
	anxifree	0.56	0.12	0.30	0.82
	joyzepam	1.32	0.12	1.06	1.58

Figure 14.9: captura de pantalla de jamovi que muestra las medias marginales para el modelo no saturado, es decir, sin el componente de interacción, con el conjunto de datos del ensayo clínico

14.4 Comprobación de supuestos

Al igual que con el ANOVA unifactorial, los supuestos clave del ANOVA factorial son la homogeneidad de la varianza (todos los grupos tienen la misma desviación estándar), la normalidad de los residuales y la independencia de las observaciones. Los dos primeros son cosas que podemos verificar. El tercero es algo que debes evaluar tú misma preguntándote si existe alguna relación especial entre las diferentes observaciones, por ejemplo, medidas repetidas en las que la variable independiente es el tiempo, por lo que existe una relación entre las observaciones en el momento uno y en el momento dos: las observaciones en momentos diferentes son de las mismas personas. Además, si no estás

utilizando un modelo saturado (por ejemplo, si has omitido los términos de interacción), también estás suponiendo que los términos omitidos no son importantes. Por supuesto, puedes verificar esto último ejecutando un ANOVA con los términos omitidos incluidos y ver si son significativos, por lo que es bastante fácil. ¿Qué pasa con la homogeneidad de la varianza y la normalidad de los residuales? Son bastante fáciles de verificar. No es diferente a las comprobaciones que hicimos en un ANOVA unifactorial.

14.4.1 Homogeneidad de varianzas

Como se mencionó en Section 13.6.1 en el último capítulo, es una buena idea inspeccionar visualmente una gráfica de las desviaciones estándar comparadas entre diferentes grupos/categorías, y también ver si la prueba de Levene es consistente con la inspección visual. La teoría que subyace a la prueba de Levene se discutió en Section 13.6.1, por lo que no la discutiré nuevamente. Esta prueba espera que tengas un modelo saturado (es decir, que incluya todos los términos relevantes), porque la prueba se ocupa principalmente de la varianza dentro del grupo, y realmente no tiene mucho sentido calcular esto de otra manera que con respecto al modelo completo. La prueba de Levene se puede especificar en la opción ANOVA 'Comprobaciones de supuestos' - 'Pruebas de homogeneidad' en jamovi, con el resultado que se muestra en Figure 14.10. El hecho de que la prueba de Levene no sea significativa significa que, siempre que sea consistente con una inspección visual de la gráfica de desviaciones estándar, podemos asumir con seguridad que no se viola el supuesto de homogeneidad de varianzas.

14.4.2 Normalidad de los residuales

Al igual que con el ANOVA unifactorial, podemos probar la normalidad de los residuales de manera directa (consulta Section 13.6.4). No obstante, generalmente es una buena idea examinar los residuales gráficamente utilizando un gráfico QQ. Ver Figure 14.10.

14.5 Análisis de covarianza (ANCOVA)

Una variación en ANOVA es cuando tienes una variable continua adicional que crees que podría estar relacionada con la variable dependiente. Esta variable adicional se puede agregar al análisis como una covariable, en el acertadamente llamado análisis de covarianza (ANCOVA).

En ANCOVA, los valores de la variable dependiente se "ajustan" por la influencia de la covariable, y luego las medias de puntuación "ajustadas" se prueban entre grupos de la manera habitual. Esta técnica puede aumentar la precisión de un experimento y, por lo tanto, proporcionar una prueba más "poderosa" de la igualdad de las medias de grupo en la variable dependiente. ¿Cómo hace esto ANCOVA? Bueno, aunque la covariable en sí no suele tener ningún interés experimental, el ajuste de la covariable puede disminuir la estimación del error experimental y, por lo tanto, al reducir la varianza del error, se aumenta la precisión. Esto significa que es menos probable un fallo inapropiada para rechazar la hipótesis nula (falso negativo o error de tipo II).

A pesar de esta ventaja, ANCOVA corre el riesgo de deshacer las diferencias reales entre grupos, y esto debe evitarse. Mira Figure 14.11, por ejemplo, que muestra un gráfico de la ansiedad estadística en relación a la edad y muestra dos grupos distintos: estudiantes que tienen antecedentes o preferencias en Artes o Ciencias. ANCOVA con

Assumption Checks

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
0.22	5	12	0.94731

[3]

Normality Test (Shapiro-Wilk)

Statistic	p
0.96	0.53290

Q-Q Plot

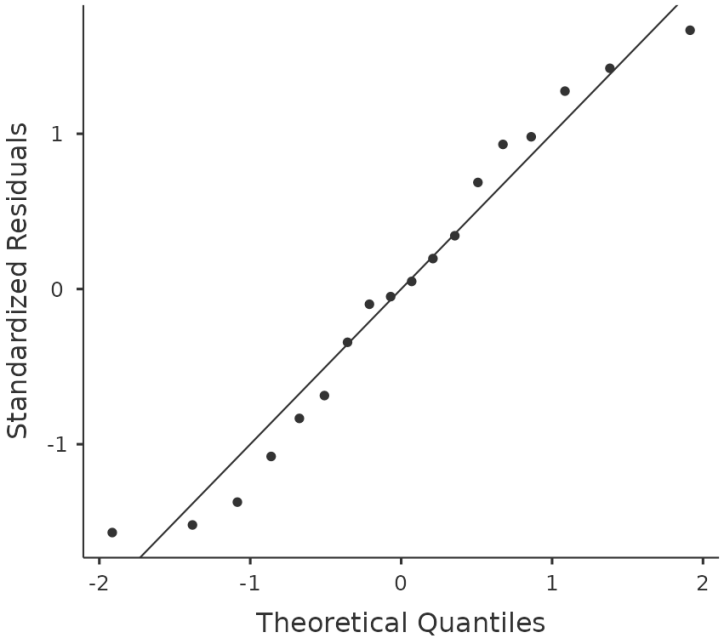


Figure 14.10: Comprobación de supuestos en un modelo ANOVA

la edad como covariable podría llevar a la conclusión de que la ansiedad estadística no difiere en los dos grupos. ¿Sería razonable esta conclusión? Probablemente no porque las edades de los dos grupos no se superponen y el análisis de varianza esencialmente "se ha extrapolado a una región sin datos" (Everitt (1996), p. 68).

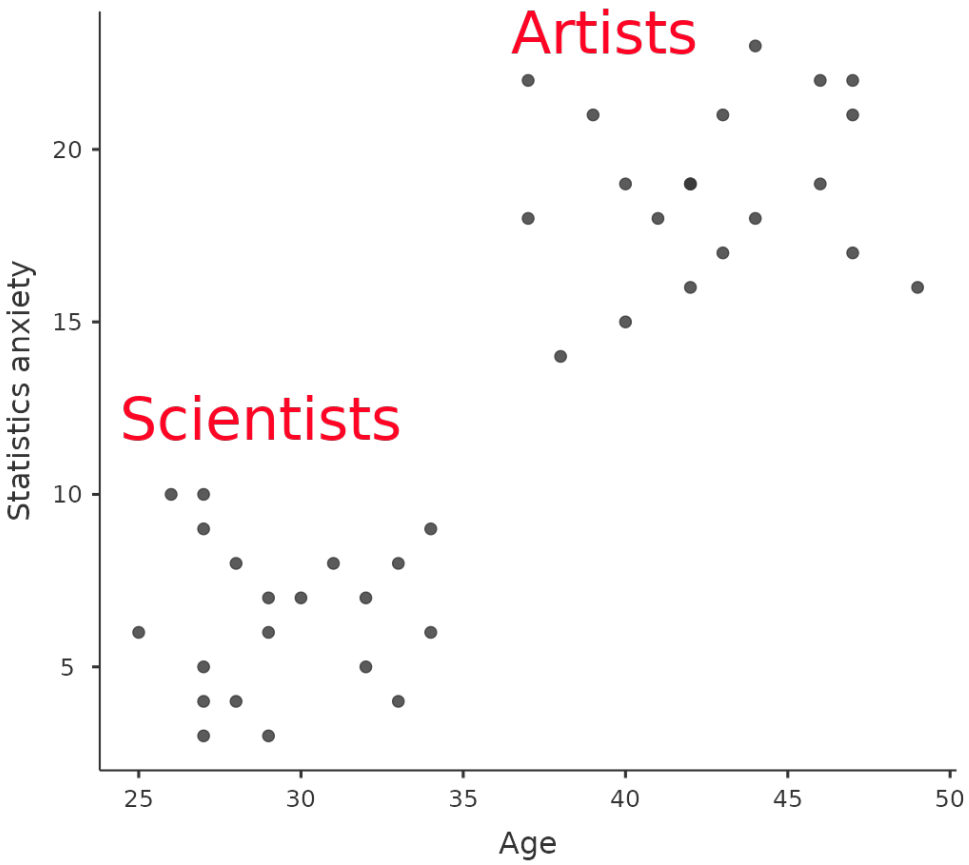


Figure 14.11: Gráfica de la ansiedad estadística frente a la edad para dos grupos distintos

Claramente, se debe pensar detenidamente en un análisis de covarianza con grupos distintos. Esto se aplica tanto a los diseños unifactoriales como a los factoriales, ya que ANCOVA se puede utilizar con ambos.

14.5.1 Ejecución de ANCOVA en jamovi

Un psicólogo de la salud estaba interesado en el efecto de la rutina de ciclismo y el estrés sobre los niveles de felicidad, con la edad como covariable. Puedes encontrar el conjunto de datos en el archivo ancova.csv. Abre este archivo en jamovi y luego, para realizar un ANCOVA, selecciona Análisis - ANOVA - ANCOVA para abrir la ventana de análisis ANCOVA (Figure 14.12). Resalta la variable dependiente 'felicidad' y transfiérela al cuadro de texto 'Variable dependiente'. Resalta las variables independientes 'estrés' y 'desplazamiento' y muévelas al cuadro de texto 'Factores fijos'. Resalta la covariable

'edad' y transfiérela al cuadro de texto 'Covariables'. Luego haz clic en las medias marginales estimadas para que aparezcan las opciones de diagramas y tablas.

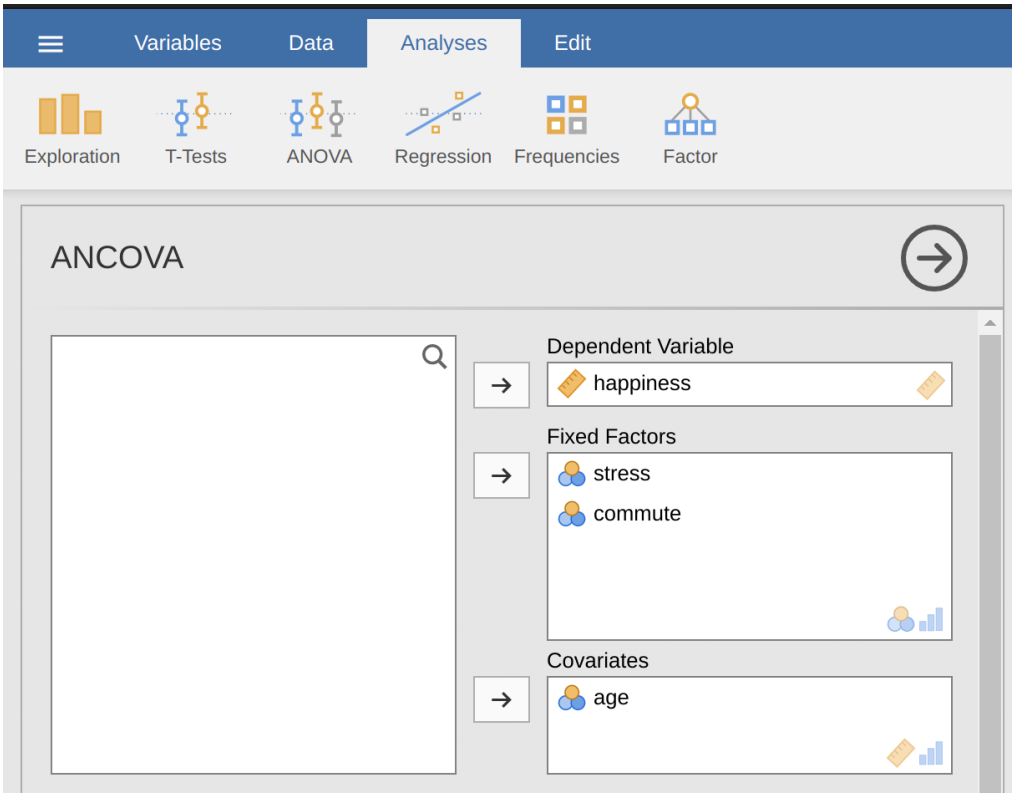


Figure 14.12: La ventana de análisis jamovi ANCOVA

En la ventana de resultados jamovi (Figure 14.13) se genera una tabla ANCOVA que muestra las pruebas de los efectos entre sujetos. El valor de F para la covariable 'edad' es significativo en $p = .023$, lo que sugiere que la edad es un predictor importante de la variable dependiente, la felicidad. Cuando observamos las puntuaciones medias marginales estimadas (Figure 14.14), se han realizado ajustes (en comparación con un análisis sin la covariable) debido a la inclusión de la covariable 'edad' en este ANCOVA. Un gráfico (Figure 14.15) es una buena manera de visualizar e interpretar los efectos significativos.

El valor F para el efecto principal 'estrés' (52.61) tiene una probabilidad asociada de $p < .001$. El valor F para el efecto principal 'desplazamiento' (42.33) tiene una probabilidad asociada de $p < .001$. Dado que ambos son menores que la probabilidad que normalmente se usa para decidir si un resultado estadístico es significativo ($p < .05$), podemos concluir que hubo un efecto principal significativo del estrés ($F(1, 15) = 52.61, p < .001$) y un efecto principal significativo del método de desplazamiento ($F(1, 15) = 42.33, p < .001$). También se encontró una interacción significativa entre el estrés y el método de desplazamiento ($F(1, 15) = 14.15, p = .002$).

En Figure 14.15 podemos ver las puntuaciones de felicidad medias marginales ajustadas

ANCOVA

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p	η^2	η^2p	ω^2
stress	2751.52	1	2751.52	52.61	< .00001	0.40	0.78	0.39
commute	2213.93	1	2213.93	42.33	< .00001	0.32	0.74	0.31
age	334.35	1	334.35	6.39	0.02316	0.05	0.30	0.04
stress ✻ commute	740.12	1	740.12	14.15	0.00188	0.11	0.49	0.10
Residuals	784.45	15	52.30					

Figure 14.13: resultados de jamovi ANCOVA para la felicidad en función del estrés y el método de desplazamiento, con la edad como covariable

Estimated Marginal Means - stress ✻ commute

				95% Confidence Interval	
commute	stress	Mean	SE	Lower	Upper
drive	high	36.11	3.24	29.21	43.02
	low	51.09	3.26	44.13	58.04
cycle	high	43.58	3.84	35.40	51.76
	low	85.82	3.71	77.90	93.74

Figure 14.14: Tabla del nivel medio de felicidad en función del estrés y el método de desplazamiento (ajustado por la covariable edad) con intervalos de confianza del 95 %

cuando la edad es una covariable en un ANCOVA. En este análisis hay un efecto de interacción significativo, por el cual las personas con poco estrés que van en bicicleta al trabajo son más felices que las personas con poco estrés que conducen y las personas con mucho estrés que van en bicicleta o en coche al trabajo. También hay un efecto principal significativo del estrés: las personas con poco estrés son más felices que las que tienen mucho estrés. Y también hay un efecto principal significativo de la conducta de de desplazamiento: las personas que van en bicicleta son más felices, en promedio, que las que conducen al trabajo.

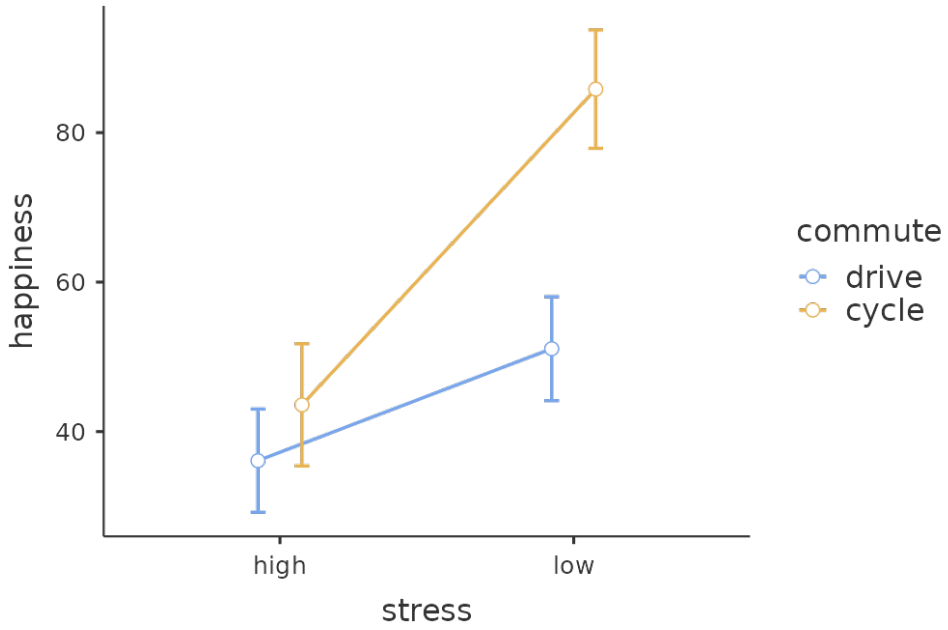


Figure 14.15: gráfico del nivel medio de felicidad en función del estrés y el método de desplazamiento

Una cosa que debes tener en cuenta es que, si estás pensando en incluir una covariable en tu ANOVA, hay una suposición adicional: la relación entre la covariable y la variable dependiente debe ser similar para todos los niveles de la variable independiente. Esto se puede verificar agregando un término de interacción entre la covariable y cada variable independiente en la opción Modelo jamovi - Términos del modelo. Si el efecto de interacción no es significativo, se puede eliminar. Si es significativo, entonces podría ser apropiada una técnica estadística diferente y más avanzada (que está más allá del alcance de este libro, por lo que es posible que desees consultar a un estadístico amigo).

14.6 ANOVA como modelo lineal

Una de las cosas más importantes que hay que entender sobre ANOVA y regresión es que básicamente son lo mismo. A simple vista, tal vez no pensarías que esto es cierto. Después de todo, la forma en que los he descrito hasta ahora sugiere que ANOVA

se ocupa principalmente de probar las diferencias de grupo, y la regresión se ocupa principalmente de comprender las correlaciones entre las variables. Y, hasta donde llega, eso es perfectamente cierto. Pero cuando miras debajo del capó, por así decirlo, la mecánica subyacente de ANOVA y la regresión son terriblemente similares. De hecho, si lo piensas bien, ya has visto evidencia de esto. Tanto ANOVA como la regresión se basan en gran medida en sumas de cuadrados (SC), ambos utilizan pruebas F, etc. Mirando hacia atrás, es difícil escapar de la sensación de que Chapter 12 y Chapter 13 eran un poco repetitivos.

La razón de esto es que ANOVA y la regresión son tipos de **modelos lineales**. En el caso de la regresión, esto es algo obvio. La ecuación de regresión que usamos para definir la relación entre predictores y resultados es la ecuación de una línea recta, por lo que obviamente es un modelo lineal, con la ecuación

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \epsilon_p$$

donde Y_p es el valor de resultado para la p-ésima observación (p. ej., p-ésima persona), X_{1p} es el valor del primer predictor para la p-ésima observación, X_{2p} es el valor del segundo predictor para la p-ésima observación, los términos b_0, b_1 y b_2 son nuestros coeficientes de regresión, y ϵ_p es el p-ésimo residuo. Si ignoramos los residuos ϵ_p y solo nos centramos en la línea de regresión, obtenemos la siguiente fórmula:

$$\hat{Y}_p = b_0 + b_1 X_{1p} + b_2 X_{2p}$$

donde $\hat{Y}_p$ es el valor de Y que la línea de regresión predice para la persona p, a diferencia del valor realmente observado Y_p. Lo que no es inmediatamente obvio es que también podemos escribir ANOVA como un modelo lineal. Sin embargo, en realidad es bastante sencillo hacerlo. Comencemos con un ejemplo realmente simple, reescribiendo un ANOVA factorial de 2×2 como un modelo lineal.

14.6.1 Algunos datos

Para concretar las cosas, supongamos que nuestra variable de resultado es la calificación que recibe un estudiante en mi clase, una variable de escala de razón que corresponde a una nota de 0 a 100. Hay dos variables predictoras de interés: si el estudiante se presentó o no a las clases (la variable de asistencia) y si el estudiante realmente leyó o no el libro de texto (la variable de lectura). Diremos que atiende = 1 si el alumno asistió a clase, y atiende = 0 si no lo hizo. Del mismo modo, diremos que lectura = 1 si el estudiante leyó el libro de texto y lectura = 0 si no lo hizo.

Bien, hasta ahora eso es bastante simple. Lo siguiente que debemos hacer es ajustar algunas matemáticas alrededor de esto (¡lo siento!). Para los propósitos de este ejemplo, permite que Y_p denote la calificación del p-ésimo estudiante en la clase. Esta no es exactamente la misma notación que usamos anteriormente en este capítulo. Anteriormente, usamos la notación Y_{rci} para referirnos a la i-ésima persona en el r-ésimo grupo para el predictor 1 (el factor de fila) y el c-ésimo grupo para el predictor 2 (el factor de columna). Esta notación extendida fue realmente útil para describir cómo se calculan los valores de SC, pero es una molestia en el contexto actual, así que cambiaré la notación aquí. Ahora, la notación Y_p es visualmente más simple que Y_{rci}, ¡pero tiene la desventaja de

que en realidad no realiza un seguimiento de las membresías del grupo! Es decir, si te dijera que $Y_{0,0,3} = 35$, inmediatamente sabrías que estamos hablando de un estudiante (de hecho, el tercer estudiante de este tipo) que no asistió a las clases (es decir, asistió = 0) y no leyó el libro de texto (es decir, lectura = 0), y que terminó suspendiendo la clase (nota = 35). Pero si te digo que $Y_p = 35$, todo lo que sabes es que el p-ésimo estudiante no obtuvo una buena calificación. Aquí hemos perdido información clave. Por supuesto, no se necesita pensar mucho para descubrir cómo solucionar esto. Lo que haremos en su lugar es introducir dos nuevas variables X_{1p} y X_{2p} que realizan un seguimiento de esta información. En el caso de nuestro estudiante hipotético, sabemos que $X_{1p} = 0$ (es decir, asistir = 0) y $X_{2p} = 0$ (es decir, leer = 0). Entonces, los datos podrían verse como Table 14.8.

Table 14.8: Datos de calificación, asistencia y lectura del libro de texto

person, p	**grade,** Y_p	**attendance,** X_{1p}	**reading,** X_{2p}
1	90	1	1
2	87	1	1
3	75	0	1
4	60	1	0
5	35	0	0
6	50	0	0
7	65	1	0
8	70	0	1

Esto no es nada particularmente especial, por supuesto. ¡Es exactamente el formato en el que esperamos ver nuestros datos! Consulta el archivo de datos rtfm.csv. Podemos utilizar el análisis 'Descriptivo' de jamovi para confirmar que este conjunto de datos corresponde a un diseño equilibrado, con 2 observaciones para cada combinación de atención y lectura. De la misma forma también podemos calcular la nota media de cada combinación. Esto se muestra en Figure 14.16. Mirando las puntuaciones medias, una tiene la fuerte impresión de que leer el texto y asistir a la clase importan mucho.

14.6.2 ANOVA con factores binarios como modelo de regresión

Bien, volvamos a hablar de las matemáticas. Ahora tenemos nuestros datos expresados en términos de tres variables numéricas: la variable continua Y y las dos variables binarias X_1 y X_2. Lo que quiero que reconozcas es que nuestro ANOVA factorial de 2×2 es exactamente equivalente al modelo de regresión

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \epsilon_p$$

¡Esta es, por supuesto, exactamente la misma ecuación que usé anteriormente para describir un modelo de regresión de dos predictores! La única diferencia es que X_1 y X_2 ahora son variables binarias (es decir, los valores solo pueden ser 0 o 1), mientras que en un análisis de regresión esperamos que X_1 y X_2 sean continuos. Hay un par de formas en las que podría tratar de convencerte de esto. Una posibilidad sería hacer un

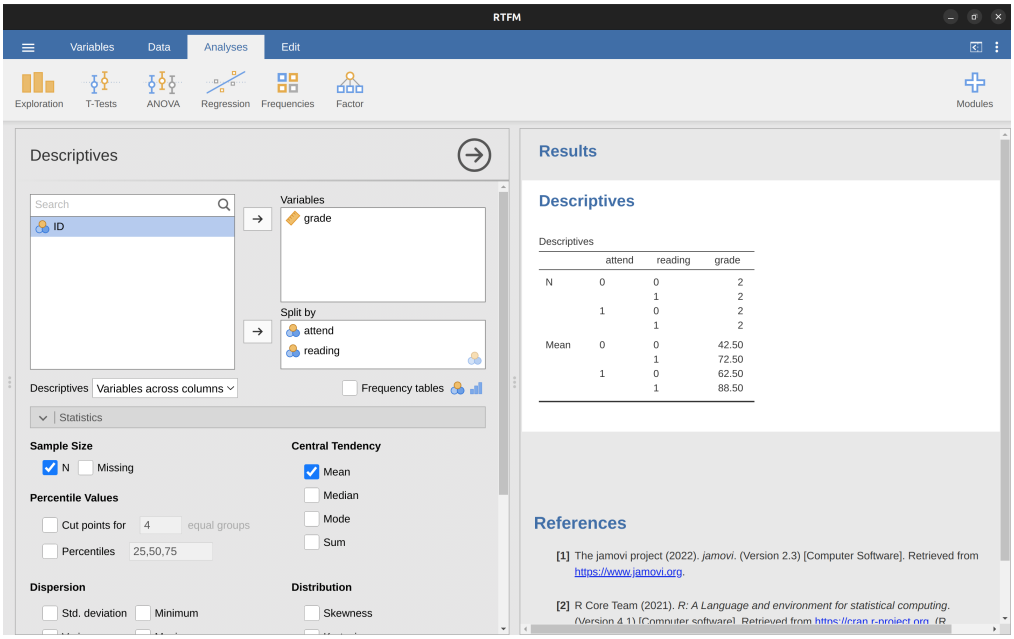


Figure 14.16: descripciones jamovi para el conjunto de datos rtfm

largo ejercicio matemático demostrando que los dos son idénticos. Sin embargo, voy a arriesgarme y supongo que la mayoría de las lectoras de este libro lo encontrarán molesto en lugar de útil. En su lugar, explicaré las ideas básicas y luego confiaré en jamovi para mostrar que los análisis ANOVA y los análisis de regresión no solo son similares, sino que son idénticos a todos los efectos. Comencemos ejecutando esto como un ANOVA. Para hacer esto, usaremos el conjunto de datos rtfm y Figure 14.17 muestra lo que obtenemos cuando ejecutamos el análisis en jamovi.

ANOVA

ANOVA - grade

	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p	ω^2
attend	648.00	1	648.00	21.60	0.00559	0.27	0.81	0.26
reading	1568.00	1	1568.00	52.27	0.00079	0.66	0.91	0.64
Residuals	150.00	5	30.00					

Figure 14.17: ANOVA del conjunto de datos rtfm.csv en jamovi, sin el término de interacción

Entonces, al leer los números clave de la tabla ANOVA y las puntuaciones medias que presentamos anteriormente, podemos ver que los estudiantes obtuvieron una calificación más alta si asistieron a clase ($F_{1,5} = 21.6, p = .0056$) y si leen el libro de texto ($F_{1,5} = 52.3, p = .0008$). Anotemos esos valores p y esos estadísticos F.

Ahora pensemos en el mismo análisis desde una perspectiva de regresión lineal. En el conjunto de datos de rtfm, hemos codificado la asistencia y la lectura como si fueran predictores numéricos. En este caso, esto es perfectamente aceptable. Realmente hay un sentido en el que un estudiante que se presenta a clase (es decir, atiende = 1) de hecho ha tenido "más asistencia" que un estudiante que no lo hace (es decir, atiende = 0). Por lo tanto, no es nada irrazonable incluirlo como predictor en un modelo de regresión. Es un poco inusual, porque el predictor solo tiene dos valores posibles, pero no viola ninguno de los supuestos de la regresión lineal. Y es fácil de interpretar. Si el coeficiente de regresión para asistir es mayor que 0 significa que los estudiantes que asisten a clases obtienen calificaciones más altas. Si es menor que cero, los estudiantes que asisten a clases obtienen calificaciones más bajas. Lo mismo es cierto para nuestra variable de lectura.

Sin embargo, espera un segundo. ¿Por qué es esto cierto? Es algo que es intuitivamente obvio para todos los que han recibido algunas clases de estadísticas y se sienten cómodos con las matemáticas, pero no está claro para todos los demás a primera vista. Para ver por qué esto es cierto, ayuda mirar de cerca a algunos estudiantes específicos. Comencemos por considerar a los estudiantes de 6.º y 7.º en nuestro conjunto de datos (es decir, $p = 6$ y $p = 7$). Ninguno ha leído el libro de texto, por lo que en ambos casos podemos poner lectura = 0. O, para decir lo mismo en nuestra notación matemática, observamos $X_{2,6} = 0$ y $X_{2,7} = 0$. Sin embargo, el estudiante número 7 sí se presentó a las clases (es decir, asistió = 1, $X_{1,7} = 1$) mientras que el estudiante número 6 no lo hizo (es decir, asistió = 0, $X_{1,6} = 0$). Ahora veamos qué sucede cuando insertamos estos números en la fórmula general de nuestra línea de regresión. Para el estudiante número 6, la regresión predice que

$$\begin{aligned}\hat{Y}_6 &= b_0 + b_1 X_{1,6} + b_2 X_{2,6} \\ &= b_0 + (b_1 \times 0) + (b_2 \times 0) \\ &= b_0\end{aligned}$$

Entonces, esperamos que este estudiante obtenga una calificación correspondiente al valor del término de intersección b_0. ¿Qué pasa con el estudiante 7? Esta vez, cuando insertamos los números en la fórmula de la línea de regresión, obtenemos lo siguiente

$$\begin{aligned}\hat{Y}_7 &= b_0 + b_1 X_{1,7} + b_2 X_{2,7} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 0) \\ &= b_0 + b_1\end{aligned}$$

Debido a que este estudiante asistió a clase, la calificación pronosticada es igual al término de intersección b0 más el coeficiente asociado con la variable de asistencia, b_1. Entonces, si b_1 es mayor que cero, esperamos que los estudiantes que asistan a las clases obtengan calificaciones más altas que los estudiantes que no lo hagan. Si este coeficiente es negativo, esperamos lo contrario: los estudiantes que asisten a clase terminan rindiendo mucho peor. De hecho, podemos llevar esto un poco más lejos. ¿Qué pasa con el estudiante número 1, que apareció en clase ($X_{1,1} = 1$) y leyó el libro de texto ($X_{2,1} = 1$)? Si reemplazamos estos números en la regresión obtenemos

$$\begin{aligned}\hat{Y}_1 &= b_0 + b_1 X_{1,1} + b_2 X_{2,1} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 1) \\ &= b_0 + b_1 + b_2\end{aligned}$$

Entonces, si asumimos que asistir a clase te ayuda a obtener una buena calificación (es decir, $b1$ 0) y si asumimos que leer el libro de texto también te ayuda a obtener una buena calificación (es decir, $b2$ 0), entonces nuestra expectativa es que el estudiante 1 obtenga una calificación más alta que el estudiante 6 y el estudiante 7.

Y en este punto no te sorprenderá saber que el modelo de regresión predice que el estudiante 3, que leyó el libro pero no asistió a las clases, obtendrá una calificación de $b_2 + b_0$. No os aburriré con otra fórmula de regresión. En su lugar, lo que haré es mostrarte Table 14.9 con las *calificaciones esperadas*.

Table 14.9: Calificaciones esperadas del modelo de regresión

		read textbook	
		no	yes
attended?	no	β_0	$\beta_0 + \beta_2$
	yes	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2$

Como puedes ver, el término de intercepción b_0 actúa como una especie de calificación "de referencia" que esperaría de aquellos estudiantes que no se toman el tiempo para asistir a clase o leer el libro de texto. De manera similar, b_1 representa el impulso que se espera que obtengas si asistes a clase, y b_2 representa el impulso que proviene de leer el libro de texto. De hecho, si se tratara de un ANOVA, es posible que quieras caracterizar b1 como el efecto principal de la asistencia y b_2 como el efecto principal de la lectura. De hecho, para un ANOVA simple de 2×2, así es exactamente como funciona.

Bien, ahora que realmente comenzamos a ver por qué ANOVA y la regresión son básicamente lo mismo, ejecutemos nuestra regresión usando los datos de rtfm y el análisis de regresión jamovi para convencernos de que esto es realmente cierto. Ejecutar la regresión de la manera habitual da los resultados que se muestran en Figure 14.18.

Hay algunas cosas interesantes a tener en cuenta aquí. Primero, fíjate que el término de intersección es 43,5, que está cerca de la media del "grupo" de 42,5 observada para esos dos estudiantes que no leyeron el texto ni asistieron a clase. En segundo lugar, observa que tenemos el coeficiente de regresión de $b_1 = 18.0$ para la variable de asistencia, lo que sugiere que aquellos estudiantes que asistieron a clase obtuvieron una puntuación un 18% más alta que aquellos que no asistieron. Entonces, nuestra expectativa sería que aquellos estudiantes que asistieron a clase pero no leyeron el libro de texto obtuvieran una calificación de $b_0 + b_1$, que es igual a $43.5 + 18.0 = 61.5$. Puedes comprobar por ti misma que sucede lo mismo cuando miramos a los alumnos que leen el libro de texto.

En realidad, podemos ir un poco más allá al establecer la equivalencia de nuestro ANOVA y nuestra regresión. Mira los valores p asociados con la variable de asistencia y la variable de lectura en el resultado de la regresión. Son idénticos a los que

Model Coefficients - grade

Predictor	Estimate	SE	t	p
Intercept	43.50	3.35	12.97	0.00005
attend	18.00	3.87	4.65	0.00559
reading	28.00	3.87	7.23	0.00079

Figure 14.18: análisis de regresión del conjunto de datos rtfm.csv en jamovi, sin el término de interacción

encontramos anteriormente cuando ejecutamos el ANOVA. Esto puede parecer un poco sorprendente, ya que la prueba utilizada al ejecutar nuestro modelo de regresión calcula un estadístico t y el ANOVA calcula un estadístico F. Sin embargo, si puedes recordar todo el camino de regreso a Chapter 7, mencioné que existe una relación entre la distribución t y la distribución F. Si tienes una cantidad que se distribuye de acuerdo con una distribución t con k grados de libertad y la elevas al cuadrado, entonces esta nueva cantidad al cuadrado sigue una distribución F cuyos grados de libertad son 1 y k. Podemos verificar esto con respecto a los estadísticos t en nuestro modelo de regresión. Para la variable de atención obtenemos un valor de 4,65. Si elevamos al cuadrado este número, obtenemos 21,6, que coincide con el estadístico F correspondiente en nuestro ANOVA.

Finalmente, una última cosa que debes saber. Debido a que jamovi comprende el hecho de que ANOVA y la regresión son ejemplos de modelos lineales, te permite extraer la tabla ANOVA clásica de su modelo de regresión utilizando la 'Regresión lineal' - 'Coeficientes del modelo' - 'Prueba ómnibus' - 'Prueba ANOVA', y esto te dará la tabla que se muestra en Figure 14.19.

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
attend	648.00	1	648.00	21.60	0.00559
reading	1568.00	1	1568.00	52.27	0.00079
Residuals	150.00	5	30.00		

Note. Type 3 sum of squares

Figure 14.19: Resultados de la prueba Omnibus ANOVA del análisis de regresión jamovi

14.6.3 Cómo codificar factores no binarios como contrastes

En este punto, te mostré cómo podemos ver un ANOVA de 2×2 en un modelo lineal. Y es bastante fácil ver cómo esto se generaliza a un ANOVA de $2 \times 2 \times 2$ o un ANOVA de $2 \times 2 \times 2 \times 2$. Es lo mismo, de verdad. Simplemente agrega una nueva variable binaria para cada uno de sus factores. Donde comienza a ser más complicado es cuando consideramos factores que tienen más de dos niveles. Considera, por ejemplo, el ANOVA de 3×2 que ejecutamos anteriormente en este capítulo utilizando los datos de Clinicaltrial.csv. ¿Cómo podemos convertir el factor de fármacos de tres niveles en una forma numérica que sea apropiada para una regresión?

La respuesta a esta pregunta es bastante simple, en realidad. Todo lo que tenemos que hacer es darnos cuenta de que un factor de tres niveles se puede reescribir como dos variables binarias. Supongamos, por ejemplo, que yo fuera a crear una nueva variable binaria llamada druganxifree. Siempre que la variable fármacos sea igual a "anxifree" ponemos druganxifree = 1. De lo contrario, ponemos druganxifree = 0. Esta variable establece un **contraste**, en este caso entre anxifree y los otros dos fármacos. Por sí solo, por supuesto, el contraste druganxifree no es suficiente para capturar completamente toda la información en nuestra variable de fármacos. Necesitamos un segundo contraste, uno que nos permita distinguir entre el joyzepam y el placebo. Para ello, podemos crear un segundo contraste binario, llamado drugjoyzepam, que vale 1 si el fármaco es joyzepam y 0 si no lo es. En conjunto, estos dos contrastes nos permiten discriminar perfectamente entre los tres posibles fármacos. Table 14.10 ilustra esto.

Table 14.10: contrastes binarios para discriminar entre los tres posibles fármacos

drug	druganxifree	drugjoyzepam
"placebo"	0	0
"anxifree"	1	0
"joyzepam"	0	1

Si el fármaco administrado a un paciente es un placebo, las dos variables de contraste serán iguales a 0. Si el fármaco es Anxifree, la variable druganxifree será igual a 1, y la variable drugjoyzepam será 0. Lo contrario es cierto para Joyzepam: drugjoyzepam es 1 y druganxifree es 0.

Crear variables de contraste no es demasiado difícil usando la instrucción calcular nueva variable en jamovi. Por ejemplo, para crear la variable Anxifree, escribe esta expresión lógica en el cuadro de fórmula de calcular nueva variable: IF (drug == 'Anxifree', 1, 0)'. De manera similar, para crear la nueva variable drugjoyzepam usa esta expresión lógica: IF(drug == 'joyzepam', 1, 0). Del mismo modo para la terapia CBT: IF(terapia == 'TCC', 1, 0). Puedes ver estas nuevas variables y las expresiones lógicas correspondientes en el archivo de datos jamoviclinicaltrial2.omv.

Ahora hemos recodificado nuestro factor de tres niveles en términos de dos variables binarias y ya hemos visto que ANOVA y la regresión se comportan de la misma manera para las variables binarias. Sin embargo, existen algunas complejidades adicionales que surgen en este caso, que analizaremos en la siguiente sección.

14.6.4 La equivalencia entre ANOVA y regresión para factores no binarios

Ahora tenemos dos versiones diferentes del mismo conjunto de datos. Nuestros datos originales en los que la variable de fármaco del archivo Clinicaltrial.csv se expresa como un único factor de tres niveles, y los datos expandidos clinicaltrial2.omv en los que se expande en dos contrastes binarios. Una vez más, lo que queremos demostrar es que nuestro ANOVA factorial original de 3×2 es equivalente a un modelo de regresión aplicado a las variables de contraste. Comencemos por volver a ejecutar el ANOVA, con los resultados que se muestran en Figure 14.20.

ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η^2	η^2p	ω^2
drug	3.45	2	1.73	26.15	0.00002	0.71	0.79	0.68
therapy	0.47	1	0.47	7.08	0.01866	0.10	0.34	0.08
Residuals	0.92	14	0.07					

Figure 14.20: resultados de jamovi ANOVA, sin componente de interacción

Obviamente, aquí no hay sorpresas. Ese es exactamente el mismo ANOVA que ejecutamos antes. A continuación, hagamos una regresión usando druganxifree, drugjoyzepam y terapia TCC como predictores. Los resultados se muestran en Figure 14.21.

Model Coefficients - mood.gain

Predictor	Estimate	SE	t	p
Intercept	0.29	0.12	2.38	0.03178
druganxifree	0.27	0.15	1.80	0.09386
drugjoyzepam	1.03	0.15	6.97	< .00001
CBTtherapy	0.32	0.12	2.66	0.01866

Figure 14.21: resultados de regresión jamovi, con variables de contraste druganxifree y drugjoyzepam

Mmm. Este no es el mismo resultado que obtuvimos la última vez. No es sorprendente que la salida de la regresión imprima los resultados de cada uno de los tres predictores por separado, tal como lo hizo cada vez que realizamos un análisis de regresión. Por un lado, podemos ver que el valor p para la variable TCC es exactamente el mismo que el del factor de terapia en nuestro ANOVA original, por lo que podemos estar seguras de que el modelo de regresión está haciendo lo mismo que hizo el ANOVA. Por otro

lado, este modelo de regresión está probando el contraste druganxifree y el contraste drugjoyzepam *por separado*, como si fueran dos variables completamente independientes. Por supuesto, no es sorprendente, porque el pobre análisis de regresión no tiene forma de saber que drugjoyzepam y druganxifree son en realidad los dos contrastes diferentes que usamos para codificar nuestro factor de farmacos de tres niveles. Por lo que se sabe, Drugjoyzepam y Druganxifree no están más relacionados entre sí que Drugjoyzepam y TerapiaTCC. Sin embargo, tú y yo lo sabemos mejor. En este punto no estamos en absoluto interesadas en determinar si estos dos contrastes son individualmente significativos. Solo queremos saber si hay un efecto "general" del fármaco. Es decir, lo que queremos que haga jamovi es ejecutar algún tipo de prueba de "comparación de modelos", una en la que los dos contrastes "relacionados con los fármacos" se agrupan para el propósito de la prueba. ¿Te suenoa? Todo lo que tenemos que hacer es especificar nuestro modelo nulo, que en este caso incluiría el predictor de la terapia TCC y omitiría las dos variables relacionadas con el fármaco, como en Figure 14.22.

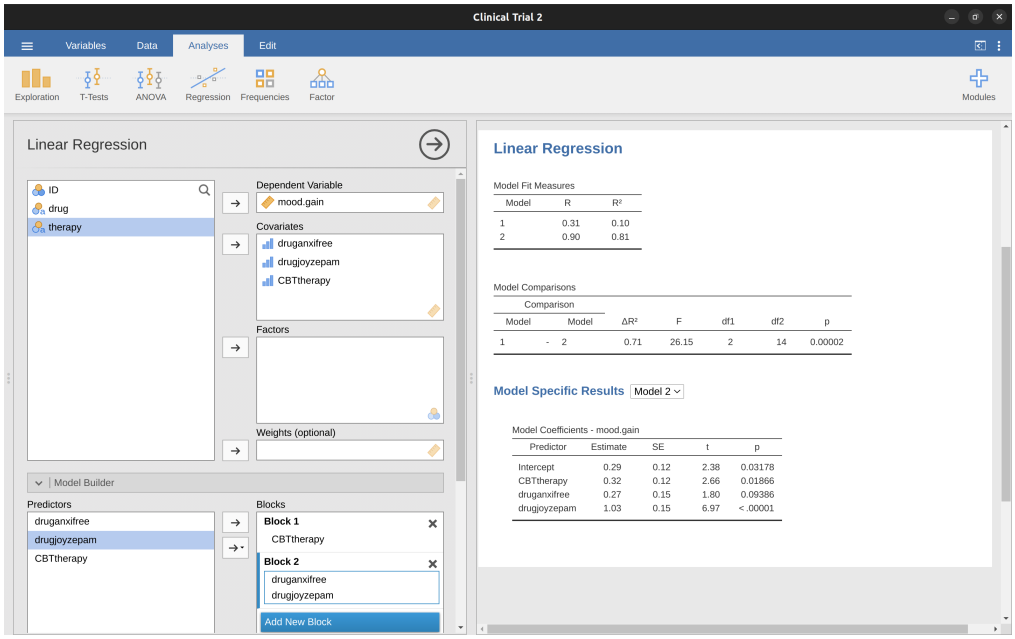


Figure 14.22: Comparación de modelos en la regresión jamovi, modelo nulo 1 vs. modelo de contrastes 2

Ah, eso está mejor. Nuestro estadístico F es 26,15, los grados de libertad son 2 y 14, y el valor p es 0,00002. Los números son idénticos a los que obtuvimos para el efecto principal del fármaco en nuestro ANOVA original. Una vez más vemos que ANOVA y regresión son esencialmente lo mismo. Ambos son modelos lineales y la maquinaria estadística subyacente en ANOVA es idéntica a la maquinaria utilizada en la regresión. La importancia de este hecho no debe ser subestimada. A lo largo del resto de este capítulo vamos a basarnos en gran medida en esta idea.

Aunque analizamos todas las complicaciones de calcular nuevas variables en jamovi para los contrastes druganxifree y drugjoyzepam, solo para mostrar que ANOVA y la regresión son esencialmente lo mismo, en el análisis de regresión lineal de jamovi hay

un ingenioso atajo para obtener estos contrastes, ver Figure 14.23. Lo que jamovi está haciendo aquí es permitirte introducir las variables predictoras que son factores como, espera... ¡factores! Inteligente, eh. También puedes especificar qué grupo usar como nivel de referencia, a través de la opción 'Niveles de referencia'. Hemos cambiado esto a 'placebo' y 'no.terapia', respectivamente, porque tiene más sentido.

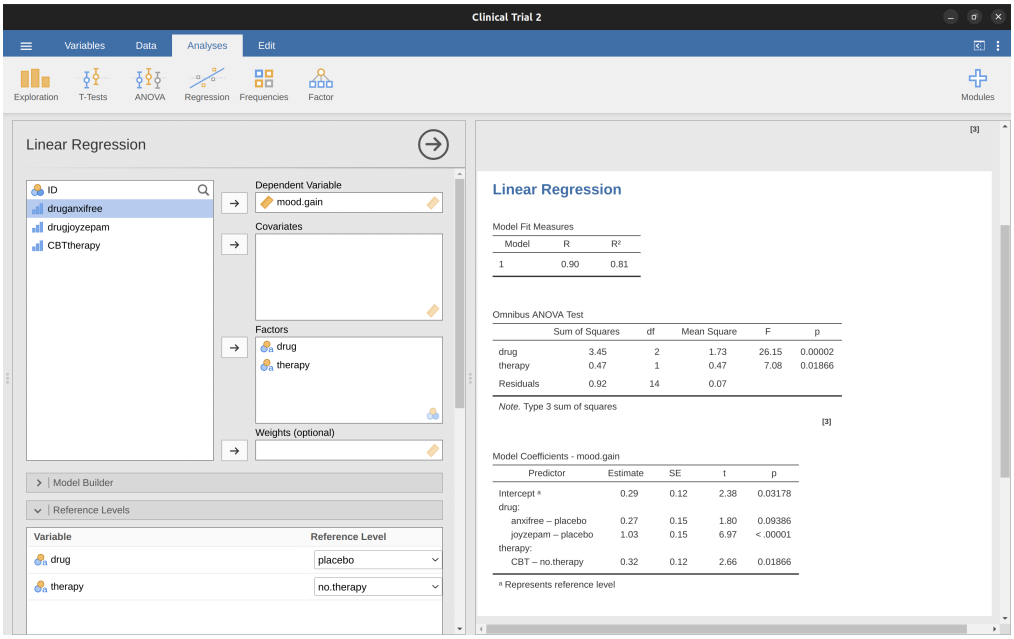


Figure 14.23: análisis de regresión con factores y contrastes en jamovi, incluidos los resultados de la prueba ANOVA ómnibus

Si también haces clic en la casilla de verificación de la prueba 'ANOVA' en la opción 'Coeficientes del modelo' - 'Prueba ómnibus', vemos que el estadístico F es 26,15, los grados de libertad son 2 y 14, y el valor p es 0,00002 (Figure 14.23). Los números son idénticos a los que obtuvimos para el efecto principal del fármaco en nuestro ANOVA original. Una vez más, vemos que ANOVA y regresión son esencialmente lo mismo. Ambos son modelos lineales y la maquinaria estadística subyacente en ANOVA es idéntica a la maquinaria utilizada en la regresión.

14.6.5 Grados de libertad como recuento de parámetros

Por fin, finalmente puedo dar una definición de grados de libertad con la que estoy contenta. Los grados de libertad se definen en términos del número de parámetros que deben estimarse en un modelo. Para un modelo de regresión o ANOVA, el número de parámetros corresponde al número de coeficientes de regresión (es decir, valores b), incluida la intersección. Teniendo en cuenta que cualquier prueba F siempre es una comparación entre dos modelos y el primer gl es la diferencia en la cantidad de parámetros. Por ejemplo, en la comparación de modelos anterior, el modelo nulo (mood.gain ~ terapiaCBT) tiene dos parámetros: hay un coeficiente de regresión para la variable terapiaCBT y otro para la intersección. El modelo alternativo (mood.gain ~ druganxifree + drugjoyzepam

+ therapyCBT) tiene cuatro parámetros: un coeficiente de regresión para cada uno de los tres contrastes y uno más para la intersección. Entonces, los grados de libertad asociados con la diferencia entre estos dos modelos son $df_1 = 4 - 2 = 2$.

¿Qué pasa cuando no parece haber un modelo nulo? Por ejemplo, podrías estar pensando en la prueba F que aparece cuando seleccionas 'Prueba F' en las opciones 'Regresión lineal' - 'Ajuste del modelo'. Originalmente lo describí como una prueba del modelo de regresión en su conjunto. Sin embargo, eso sigue siendo una comparación entre dos modelos. El modelo nulo es el modelo trivial que solo incluye 1 coeficiente de regresión, para el término de intersección. El modelo alternativo contiene $K + 1$ coeficientes de regresión, uno para cada una de las K variables predictoras y uno más para la intersección. Entonces, el valor de gl que ves en esta prueba F es igual a $df_1 = K+1-1 = K$.

¿Qué pasa con el segundo valor de gl que aparece en la prueba F? Esto siempre se refiere a los grados de libertad asociados con los residuales. También es posible pensar en esto en términos de parámetros, pero de una manera un poco contraria a la intuición. Piensa en esto, de esta manera. Supón que el número total de observaciones en todo el estudio es N. Si quieres describir perfectamente cada uno de estos valores N, debes hacerlo usando, bueno… N números. Cuando creas un modelo de regresión, lo que realmente estás haciendo es especificar que algunos de los números deben describir perfectamente los datos. Si tu modelo tiene K predictores y una intersección, entonces has especificado $K + 1$ números. Entonces, sin molestarte en averiguar exactamente cómo se haría esto, ¿cuántos números más crees que se necesitarán para transformar un modelo de regresión de parámetros K '1 en una redescripción perfecta de los datos sin procesar? Si te encuentras pensando que $(K + 1) + (N - K - 1) = N$, por lo que la respuesta tendría que ser $N - K - 1$, ¡bien hecho! Eso es correcto. En principio, puedes imaginar un modelo de regresión absurdamente complicado que incluye un parámetro para cada punto de datos y, por supuesto, proporcionaría una descripción perfecta de los datos. Este modelo contendría N parámetros en total, pero estamos interesadas en la diferencia entre la cantidad de parámetros necesarios para describir este modelo completo (es decir, N) y la cantidad de parámetros utilizados por el modelo de regresión más simple en el que estás realmente interesada (es decir, $K+1$), por lo que el segundo grado de libertad en la prueba F es $df_2 = N-K-1$, donde K es el número de predictores (en un modelo de regresión) o el número de contrastes (en un ANOVA). En el ejemplo anterior, hay ($N = 18$ observaciones en el conjunto de datos y $K + 1 = 4$ coeficientes de regresión asociados con el modelo ANOVA, por lo que los grados de libertad de los residuales son $df_2 = 18 - 4 = 14$.

14.7 Diferentes formas de especificar contrastes

En la sección anterior, te mostré un método para convertir un factor en una colección de contrastes. En el método que te mostré, especificamos un conjunto de variables binarias en las que definimos una tabla como Table 14.11.

Cada fila de la tabla corresponde a uno de los niveles de los factores, y cada columna corresponde a uno de los contrastes. Esta tabla, que siempre tiene una fila más que columnas, tiene un nombre especial. Se llama matriz de contraste. Sin embargo, hay muchas formas diferentes de especificar una matriz de contraste. En esta sección, discuto algunas de las matrices de contraste estándar que usan los estadísticos y cómo puedes

Table 14.11: contrastes binarios para discriminar entre los tres posibles fármacos

drug	druganxifree	drugjoyzepam
”placebo”	0	0
”anxifree”	1	0
”joyzepam”	0	1

usarlas en jamovi. Si planeas leer la sección sobre [ANOVA factorial 3: diseños no balanceados] más adelante, vale la pena leer esta sección detenidamente. Si no, puedes pasarla por alto, porque la elección de los contrastes no importa mucho para los diseños equilibrados.

14.7.1 Contrastes de tratamiento

En el tipo particular de contrastes que he descrito anteriormente, un nivel del factor es especial y actúa como una especie de categoría de “línea base” (es decir, placebo en nuestro ejemplo), frente a la cual se definen los otros dos. El nombre de este tipo de contrastes es contrastes de tratamiento, también conocidos como “codificación ficticia”. En este contraste, cada nivel del factor se compara con un nivel de referencia base, y el nivel de referencia base es el valor de la intersección.

El nombre refleja el hecho de que estos contrastes son bastante naturales y sensibles cuando una de las categorías de su factor es realmente especial porque en realidad representa una línea base. Eso tiene sentido en nuestro ejemplo de ensayo clínico. La condición de placebo corresponde a la situación en la que no le das a la gente ningún fármaco real, por lo que es especial. Las otras dos condiciones se definen en relación con el placebo. En un caso reemplazas el placebo con Anxifree, y en el otro caso lo reemplazas con Joyzepam.

La tabla que se muestra arriba es una matriz de contrastes de tratamiento para un factor que tiene 3 niveles. Pero supongamos que quiero una matriz de contrastes de tratamiento para un factor con 5 niveles. Establecería esto como Table 14.12.

Table 14.12: Matriz de contrastes de tratamiento con 5 niveles

Level	2	3	4	5
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

En este ejemplo, el primer contraste es el nivel 2 comparado con el nivel 1, el segundo contraste es el nivel 3 comparado con el nivel 1, y así sucesivamente. Ten en cuenta que, de forma predeterminada, el *primer* nivel del factor siempre se trata como la categoría de referencia (es decir, es el que tiene todo ceros y no tiene un contraste explícito asociado). En jamovi, puedes cambiar qué categoría es el primer nivel del factor manipulando el orden de los niveles de la variable que se muestra en la ventana ‘Variable de datos’ (haz

doble clic en el nombre de la variable en la columna de la hoja de cálculo para que aparezca la Vista de variables de datos.

14.7.2 Contrastes Helmert

Los contrastes de tratamiento son útiles para muchas situaciones. Sin embargo, tienen más sentido en la situación en la que realmente hay una categoría de referencia y quieres evaluar todos los demás grupos en relación con esa. En otras situaciones, sin embargo, no existe tal categoría de referencia y puede tener más sentido comparar cada grupo con la media de los otros grupos. Aquí es donde nos encontramos con los contrastes de Helmert, generados por la opción 'helmert' en el cuadro de selección jamovi 'ANOVA' - 'Contrastes'. La idea que subyace a los contrastes de Helmert es comparar cada grupo con la media de los "anteriores". Es decir, el primer contraste representa la diferencia entre el grupo 2 y el grupo 1, el segundo contraste representa la diferencia entre el grupo 3 y la media de los grupos 1 y 2, y así sucesivamente. Esto se traduce en una matriz de contraste que se parece a Table 14.13 para un factor con cinco niveles.

Table 14.13: Matriz de contrastes helmert con 5 niveles

1	**-1**	**-1**	**-1**	**-1**
2	1	-1	-1	-1
3	0	2	-1	-1
4	0	0	3	-1
5	0	0	0	4

Algo útil acerca de los contrastes de Helmert es que cada contraste suma cero (es decir, todas las columnas suman cero). Esto tiene como consecuencia que, cuando interpretamos el ANOVA como una regresión, el término de la intersección corresponde a la media general $\mu_{..}$ si estamos usando contrastes de Helmert. Compara esto con los contrastes de tratamiento, en los que el término de intersección corresponde a la media del grupo para la categoría de referencia. Esta propiedad puede ser muy útil en algunas situaciones. Lo que hemos estado asumiendo hasta ahora no es tan importante si tienes un diseño balanceado, pero será importante más adelante cuando consideremos [diseños no balanceados] (ANOVA factorial: diseños no balanceados). De hecho, la razón principal por la que me he molestado en incluir esta sección es que los contrastes se vuelven importantes si quieres entender el ANOVA no balanceado.

14.7.3 Contrastes de suma a cero

La tercera opción que debo mencionar brevemente son los contrastes de "suma a cero", llamados contrastes "simples" en jamovi, que se utilizan para construir comparaciones por pares entre grupos. En concreto, cada contraste codifica la diferencia entre uno de los grupos y una categoría base, que en este caso corresponde al primer grupo (Table 14.14).

Al igual que los contrastes de Helmert, vemos que cada columna suma cero, lo que significa que el término de intersección corresponde a la media general cuando ANOVA se trata como un modelo de regresión. Al interpretar estos contrastes, lo que hay que reconocer es que cada uno de estos contrastes es una comparación por pares entre el grupo 1 y uno de los otros cuatro grupos. Específicamente, el contraste 1 corresponde

Table 14.14: Matriz de contrastes suma a cero con 5 niveles

1	**-1**	**-1**	**-1**	**-1**
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

a una comparación de "grupo 2 menos grupo 1", el contraste 2 corresponde a una comparación de "grupo 3 menos grupo 1", y así sucesivamente.[8]

14.7.4 Contrastes opcionales en jamovi

jamovi también viene con una variedad de opciones que pueden generar diferentes tipos de contrastes en ANOVA. Estos se pueden encontrar en la opción 'Contrastes' en la ventana principal de análisis de ANOVA, donde se enumeran los tipos de contraste en Table 14.15:

14.8 Pruebas post hoc

Es hora de cambiar a un tema diferente. En lugar de comparaciones planificadas previamente que hayas probado utilizando contrastes, supongamos que has realizado tu ANOVA y resulta que obtuviste algunos efectos significativos. Debido al hecho de que las pruebas F son pruebas "ómnibus" que realmente solo prueban la hipótesis nula de que no hay diferencias entre los grupos, la obtención de un efecto significativo no indica qué grupos son diferentes de otros. Discutimos este problema en Chapter 13, y en ese capítulo nuestra solución fue ejecutar pruebas t para todos los pares de grupos posibles, haciendo correcciones para comparaciones múltiples (por ejemplo, Bonferroni, Holm) para controlar la tasa de error de tipo I en todas las comparaciones. Los métodos que usamos en Chapter 13 tienen la ventaja de ser relativamente simples y ser el tipo de herramientas que puedes usar en muchas situaciones diferentes en las que estás probando múltiples hipótesis, pero no son necesariamente las mejores opciones si estás interesada en realizar pruebas post hoc eficientes en un contexto ANOVA. En realidad, hay muchos métodos diferentes para realizar comparaciones múltiples en la literatura estadística (Hsu, 1996), y estaría fuera del alcance de un texto introductorio como este discutirlos todos en detalle.

Dicho esto, hay una herramienta sobre la que quiero llamar tu atención, a saber, la "Diferencia honestamente significativa" de Tukey, o **HSD de Tukey** para abreviar. Por una vez, te ahorraré las fórmulas y me limitaré a las ideas cualitativas. La idea básica en el HSD de Tukey es examinar todas las comparaciones por pares relevantes entre grupos, y solo es realmente apropiado usar el HSD de Tukey si lo que te interesa

[8] ¿Cuál es la diferencia entre el tratamiento y los contrastes simples, te escucho preguntar? Bueno, como ejemplo básico, considera un efecto principal de género, con $m = 0$ y $f = 1$. El coeficiente correspondiente al contraste de tratamientos medirá la diferencia de medias entre hombres y mujeres, y la intersección sería la media de los hombres. Sin embargo, con un contraste simple, es decir, $m = -1$ y $f = 1$, la intersección es el promedio de las medias y el efecto principal es la diferencia de la media de cada grupo con respecto a la intersección.

son las diferencias por pares.[9] Por ejemplo, antes realizaste un ANOVA factorial usando el conjunto de datos clinictrial.csv, y donde especificamos un efecto principal para el fármaco y un efecto principal para la terapia, estaríamos interesados en las siguientes cuatro comparaciones:

- La diferencia en el estado de ánimo de las personas que recibieron Anxifree frente a las personas que recibieron el placebo.
- La diferencia en el estado de ánimo de las personas que recibieron Joyzepam versus las personas que recibieron el placebo.
- La diferencia en el estado de ánimo de las personas que recibieron Anxifree frente a las personas que recibieron Joyzepam.
- La diferencia en el aumento del estado de ánimo para las personas tratadas con TCC y las personas que no recibieron terapia.

Para cualquiera de estas comparaciones, estamos interesadas en la verdadera diferencia entre las medias de los grupos (población). El HSD de Tukey construye intervalos de confianza simultáneos para las cuatro comparaciones. Lo que queremos decir con un intervalo de confianza "simultáneo" del 95 % es que, si tuviéramos que repetir este estudio muchas veces, entonces en el 95 % de los resultados del estudio, los intervalos de confianza contendrían el valor verdadero relevante. Además, podemos usar estos intervalos de confianza para calcular un valor p ajustado para cualquier comparación específica.

La función TukeyHSD en jamovi es bastante fácil de usar. Simplemente especifica el término del modelo ANOVA para el que deseas ejecutar las pruebas post hoc. Por ejemplo, si buscáramos ejecutar pruebas post hoc para los efectos principales pero no para la interacción, abriríamos la opción 'Pruebas Post Hoc' en la pantalla de análisis de ANOVA, moverías las variables del fármaco y la terapia al recuadro de la derecha, y luego seleccionas la casilla de verificación 'Tukey' en la lista de posibles correcciones post hoc que podrían aplicarse. Esto, junto con la tabla de resultados correspondiente, se muestra en Figure 14.24.

El resultado que se muestra en la tabla de resultados de 'Pruebas post hoc' es (espero) bastante sencillo. La primera comparación, por ejemplo, es la diferencia de Anxifree versus placebo, y la primera parte del resultado indica que la diferencia observada en las medias de los grupos es .27. El siguiente número es el error estándar de la diferencia, a partir del cual podríamos calcular el intervalo de confianza del 95 % si quisiéramos, aunque jamovi actualmente no ofrece esta opción. Luego hay una columna con los grados de libertad, una columna con el valor t y finalmente una columna con el valor p. Para la primera comparación, el valor p ajustado es .21. En cambio, si nos fijamos en la siguiente línea, vemos que la diferencia observada entre el joyzepam y el placebo es de 1,03, y este resultado es significativo ($p < 0{,}001$).

Hasta aquí todo bien. ¿Qué pasa si tu modelo incluye términos de interacción? Por ejemplo, la opción predeterminada en jamovi es permitir la posibilidad de que exista una interacción entre el fármaco y la terapia. Si ese es el caso, la cantidad de comparaciones por pares que debemos considerar comienza a aumentar. Como antes, necesitamos

[9] si, por ejemplo, realmente estás interesada en saber si el Grupo A es significativamente diferente de la media del Grupo B y el Grupo C, entonces necesitas usar una herramienta diferente (por ejemplo, el método de Scheffe , que es más conservador y está fuera del alcance de este libro). Sin embargo, en la mayoría de los casos, probablemente estés interesada en las diferencias de grupos por parejas, por lo que es útil conocer el HSD de Tukey.

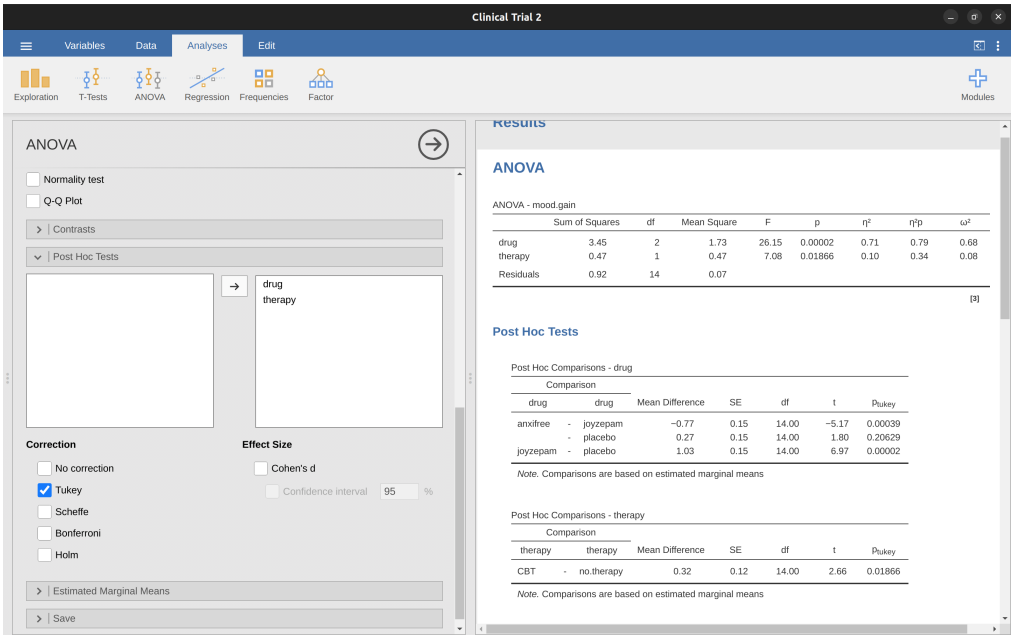


Figure 14.24: prueba post hoc de Tukey HSD en ANOVA factorial jamovi, sin un término de interacción

considerar las tres comparaciones que son relevantes para el efecto principal del fármaco y la única comparación que es relevante para el efecto principal de la terapia. Pero, si queremos considerar la posibilidad de una interacción significativa (y tratar de encontrar las diferencias de grupo que sustentan esa interacción significativa), debemos incluir comparaciones como las siguientes:

- La diferencia en el aumento del estado de ánimo de las personas que recibieron Anxifree y recibieron tratamiento con TCC, en comparación con las personas que recibieron el placebo y recibieron tratamiento con TCC
- La diferencia en el estado de ánimo de las personas que recibieron Anxifree y no recibieron terapia, en comparación con las personas que recibieron el placebo y no recibieron terapia.
- etc

Hay muchas de estas comparaciones que debes considerar. Entonces, cuando ejecutamos el análisis post hoc de Tukey para este modelo ANOVA, vemos que ha realizado muchas comparaciones por pares (19 en total), como se muestra en Figure 14.25. Puedes ver que es bastante similar al anterior, pero con muchas más comparaciones.

14.9 El método de las comparaciones planificadas

Siguiendo con las secciones anteriores sobre contrastes y pruebas post hoc en ANOVA, creo que el método de comparaciones planificadas es lo suficientemente importante como para merecer una breve discusión. En nuestras discusiones sobre comparaciones múltiples, en la sección anterior y en Chapter 13, supuse que las pruebas que deseas ejecutar

Post Hoc Tests

Post Hoc Comparisons - drug

Comparison			Mean Difference	SE	df	t	p_{tukey}
drug		drug					
anxifree	-	joyzepam	−0.77	0.13	12.00	−5.69	0.00027
	-	placebo	0.27	0.13	12.00	1.98	0.15971
joyzepam	-	placebo	1.03	0.13	12.00	7.67	0.00002

Note. Comparisons are based on estimated marginal means

Post Hoc Comparisons - therapy

Comparison			Mean Difference	SE	df	t	p_{tukey}
therapy		therapy					
CBT	-	no.therapy	0.32	0.11	12.00	2.93	0.01262

Note. Comparisons are based on estimated marginal means

Post Hoc Comparisons - drug ✻ therapy

Comparison					Mean Difference	SE	df	t	p_{tukey}
drug	therapy		drug	therapy					
anxifree	CBT	-	anxifree	no.therapy	0.63	0.19	12.00	3.32	0.05298
		-	joyzepam	CBT	−0.47	0.19	12.00	−2.45	0.21392
		-	joyzepam	no.therapy	−0.43	0.19	12.00	−2.27	0.27506
		-	placebo	CBT	0.43	0.19	12.00	2.27	0.27506
		-	placebo	no.therapy	0.73	0.19	12.00	3.85	0.02187
	no.therapy	-	joyzepam	CBT	−1.10	0.19	12.00	−5.77	0.00096
		-	joyzepam	no.therapy	−1.07	0.19	12.00	−5.60	0.00126
		-	placebo	CBT	−0.20	0.19	12.00	−1.05	0.89172
		-	placebo	no.therapy	0.10	0.19	12.00	0.52	0.99401
joyzepam	CBT	-	joyzepam	no.therapy	0.03	0.19	12.00	0.17	0.99997
		-	placebo	CBT	0.90	0.19	12.00	4.72	0.00507
		-	placebo	no.therapy	1.20	0.19	12.00	6.30	0.00044
	no.therapy	-	placebo	CBT	0.87	0.19	12.00	4.55	0.00676
		-	placebo	no.therapy	1.17	0.19	12.00	6.12	0.00057
placebo	CBT	-	placebo	no.therapy	0.30	0.19	12.00	1.57	0.62800

Note. Comparisons are based on estimated marginal means

Figure 14.25: prueba post hoc de Tukey HSD en ANOVA factorial jamovi con un término de interacción

son genuinamente post hoc. Por ejemplo, en nuestro ejemplo de fármacos anterior, tal vez pensaste que todos los fármacos tendrían efectos diferentes en el estado de ánimo (es decir, planteaste la hipótesis de un efecto principal del fármaco), pero no tenías ninguna hipótesis específica sobre cómo serían las diferencias, ni tenías una idea real sobre qué comparaciones por pares valdría la pena mirar. Si ese es el caso, entonces realmente tiens que recurrir a algo como el HSD de Tukey para hacer tus comparaciones por pares.

Sin embargo, la situación es bastante diferente si realmente tuvieras hipótesis reales y específicas sobre qué comparaciones son de interés, y nunca tuvieras la intención de ver otras comparaciones además de las que especificaste con anticipación. Cuando esto es cierto, y si te apegas honesta y rigurosamente a tus nobles intenciones de no realizar ninguna otra comparación (incluso cuando los datos parezcan mostrarte efectos deliciosamente significativos para cosas para las que no tenías una prueba de hipótesis), entonces realmente no tiene mucho sentido ejecutar algo como el HSD de Tukey, porque hace correcciones para un montón de comparaciones que nunca te importaron y nunca tuviste la intención de mirar. En esas circunstancias, puedes ejecutar con seguridad una cantidad (limitada) de pruebas de hipótesis sin realizar un ajuste para pruebas múltiples. Esta situación se conoce como método de comparaciones planificadas, y en ocasiones se utiliza en ensayos clínicos. Sin embargo, la consideración adicional está fuera del alcance de este libro introductorio, ¡pero al menos que sepas que este método existe!

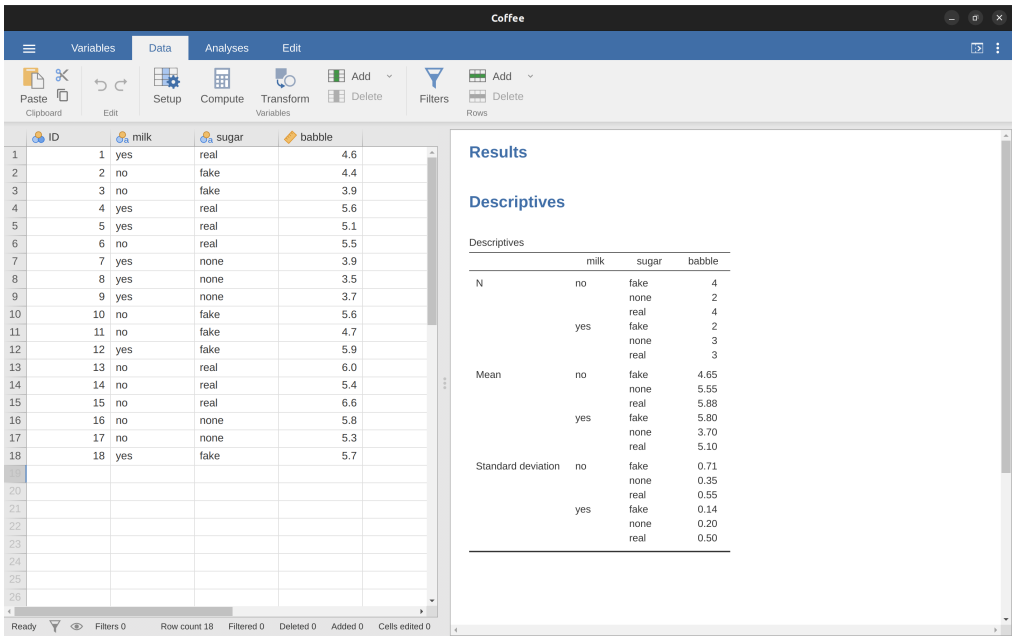
14.10 ANOVA factorial 3: diseños no equilibrados

Es útil conocer el ANOVA factorial. Ha sido una de las herramientas estándar utilizadas para analizar datos experimentales durante muchas décadas, y descubrirás que no puede leer más de dos o tres artículos de psicología sin encontrarte con un ANOVA en alguna parte. Sin embargo, hay una gran diferencia entre los ANOVA que verás en muchos artículos científicos reales y los ANOVA que he descrito hasta ahora. En la vida real, rara vez tenemos la suerte de tener diseños perfectamente equilibrados. Por una razón u otra, es típico terminar con más observaciones en algunas celdas que en otras. O, dicho de otro modo, tenemos un diseño desequilibrado.

Los diseños desequilibrados deben tratarse con mucho más cuidado que los diseños equilibrados, y la teoría estadística que los sustenta es mucho más confusa. Puede ser una consecuencia de este desorden, o puede ser la falta de tiempo, pero mi experiencia ha sido que las clases de métodos de investigación de grado en psicología tienen una desagradable tendencia a ignorar este problema por completo. Muchos libros de texto de estadísticas también tienden a pasarlo por alto. El resultado de esto, creo, es que muchos investigadores activos en el campo en realidad no saben que hay varios "tipos" diferentes de ANOVA desequilibrados, y producen respuestas bastante diferentes. De hecho, al leer la literatura psicológica, me sorprende un poco el hecho de que la mayoría de las personas que informan los resultados de un ANOVA factorial desequilibrado en realidad no ofrecen suficientes detalles para reproducir el análisis. Secretamente sospecho que la mayoría de las personas ni siquiera se dan cuenta de que su paquete de software estadístico está tomando muchas decisiones de análisis de datos sustantivos en su nombre. En realidad, es un poco aterrador cuando lo piensas. Entonces, si quieres evitar entregar el control de tu análisis de datos a un software estúpido, sigue leyendo.

14.10.1 Los datos del café

Como es habitual, nos servirá para trabajar con algunos datos. El archivo coffee.csv contiene un conjunto de datos hipotéticos que produce un ANOVA desequilibrado de 3×2. Supongamos que estuviéramos interesadas en averiguar si la tendencia de las personas a balbucear cuando toman demasiado café es puramente un efecto del café en sí, o si hay algún efecto de la leche y el azúcar que las personas agregan al café. Supongamos que llevamos a 18 personas y les damos un poco de café para beber. La cantidad de café/cafeína se mantuvo constante y variamos si se agregó leche o no, por lo que la leche es un factor binario con dos niveles, "sí" y "no". También variamos el tipo de azúcar involucrado. El café podría contener azúcar "real" o podría contener azúcar "falsa" (es decir, edulcorante artificial) o podría contener "ninguna", por lo que la variable azúcar es un factor de tres niveles. Nuestra variable de resultado es una variable continua que presumiblemente se refiere a alguna medida psicológicamente sensible de la medida en que alguien está "balbuceando". Los detalles realmente no importan para nuestro propósito. Echa un vistazo a los datos en la vista de hoja de cálculo jamovi, como en Figure 14.26.



	ID	milk	sugar	babble
1	1	yes	real	4.6
2	2	no	fake	4.4
3	3	no	fake	3.9
4	4	yes	real	5.6
5	5	yes	real	5.1
6	6	no	real	5.5
7	7	yes	none	3.9
8	8	yes	none	3.5
9	9	yes	none	3.7
10	10	no	fake	5.6
11	11	no	fake	4.7
12	12	yes	fake	5.9
13	13	no	real	6.0
14	14	no	real	5.4
15	15	no	real	6.6
16	16	no	none	5.8
17	17	no	none	5.3
18	18	yes	fake	5.7

	milk	sugar	babble
N	no	fake	4
		none	2
		real	4
	yes	fake	2
		none	3
		real	3
Mean	no	fake	4.65
		none	5.55
		real	5.88
	yes	fake	5.80
		none	3.70
		real	5.10
Standard deviation	no	fake	0.71
		none	0.35
		real	0.55
	yes	fake	0.14
		none	0.20
		real	0.50

Figure 14.26: el conjunto de datos coffee.csv en jamovi, con información descriptiva agregada por niveles del factor

Mirando la tabla de medias en Figure 14.26 tenemos una fuerte impresión de que hay diferencias entre los grupos. Esto es especialmente cierto cuando comparamos estas medias con las desviaciones estándar de la variable balbuceo. Entre los grupos, esta desviación estándar varía de 0,14 a 0,71, que es bastante pequeña en relación con las diferencias en las medias de los grupos.[10] Si bien al principio esto puede parecer un

[10] esta discrepancia en las desviaciones estándar podría (y debería) hacer que te preguntes si tenemos una violación del supuesto de homogeneidad de varianzas. Lo dejaré como un ejercicio para que el

ANOVA factorial sencillo, un problema surge cuando miramos cuántas observaciones tenemos en cada grupo. Fíjate en las diferentes N para los diferentes grupos que se muestran en Figure 14.26. Esto viola una de nuestras suposiciones originales, a saber, que el número de personas en cada grupo es el mismo. Realmente no hemos discutido cómo manejar esta situación.

14.10.2 El "ANOVA estándar" no existe para diseños desequilibrados

Los diseños desequilibrados nos llevan al descubrimiento un tanto inquietante de que en realidad no hay nada a lo que podamos referirnos como un ANOVA estándar. De hecho, resulta que hay tres formas fundamentalmente diferentes[11] en las que es posible que quieras ejecutar un ANOVA en un diseño desequilibrado. Si tienes un diseño equilibrado, las tres versiones producen resultados idénticos, con sumas de cuadrados, valores F, etc., todos conformes a las fórmulas que di al comienzo del capítulo. Sin embargo, cuando tu diseño está desequilibrado, no dan los mismos resultados. Además, no todos son igualmente apropiados para cada situación. Algunos métodos serán más apropiados para tu situación que otros. Dado todo esto, es importante comprender cuáles son los diferentes tipos de ANOVA y en qué se diferencian entre sí.

El primer tipo de ANOVA se conoce convencionalmente como **suma de cuadrados tipo I**. Estoy segura de que puedes adivinar cómo se llaman los otros dos. La parte de "suma de cuadrados" del nombre fue introducida por el paquete de software estadístico SAS y se ha convertido en una nomenclatura estándar, pero es un poco engañosa en algunos aspectos. Creo que la lógica para referirse a ellos como diferentes tipos de suma de cuadrados es que, cuando miras las tablas ANOVA que producen, la diferencia clave en los números son los valores SC. Los grados de libertad no cambian, los valores de MC aún se definen como SC dividido por df, etc. Sin embargo, la terminología es incorrecta porque oculta la razón *por la cual* los valores de SC son diferentes entre sí. Con ese fin, es mucho más útil pensar en los tres tipos diferentes de ANOVA como tres *estrategias de prueba de hipótesis* diferentes. Estas diferentes estrategias conducen a diferentes valores de SC, sin duda, pero lo importante aquí es la estrategia, no los valores de SC en sí mismos. Recuerda de la sección ANOVA como modelo lineal que cualquier prueba F en particular se considera mejor como una comparación entre dos modelos lineales. Entonces, cuando miras una tabla ANOVA, es útil recordar que cada una de esas pruebas F corresponde a un par de modelos que se están comparando. Por supuesto, esto lleva naturalmente a la pregunta de qué par de modelos se está comparando. Esta es la diferencia fundamental entre ANOVA Tipos I, II y III: cada uno corresponde a una forma diferente de elegir los pares de modelos para las pruebas.

lector verifique esto usando la opción de prueba de Levene.

[11] En realidad, esto es un poco mentira. Los ANOVA pueden variar de otras maneras además de las que he discutido en este libro. Por ejemplo, he ignorado por completo la diferencia entre los modelos de efectos fijos en los que los niveles de un factor son "fijos" por el experimentador o el mundo, y los modelos de efectos aleatorios en los que los niveles son muestras aleatorias de una población más grande de niveles posibles (este libro solo cubre modelos de efectos fijos). No cometas el error de pensar que este libro, o cualquier otro, te dirá "todo lo que necesitas saber" sobre estadística, más de lo que un solo libro podría decirte todo lo que necesitas saber sobre psicología, física o filosofía. La vida es demasiado complicada para que eso sea cierto. Sin embargo, esto no es motivo de desesperación. La mayoría de los investigadores se las arreglan con un conocimiento práctico básico de ANOVA que no va más allá que este libro. Solo quiero que tengas en cuenta que este libro es solo el comienzo de una historia muy larga, no la historia completa.

14.10.3 Suma de Cuadrados Tipo I

El método Tipo I a veces se denomina suma de cuadrados "secuencial", porque implica un proceso de agregar términos al modelo de uno en uno. Considera los datos del café, por ejemplo. Supongamos que queremos ejecutar el ANOVA factorial completo de 3×2, incluidos los términos de interacción. El modelo completo contiene la variable de resultado balbuceo, las variables predictoras azúcar y leche, y el término de interacción azúcar × leche. Esto se puede escribir como $balbuceo \sum azcar + leche + azcar \times leche$. La estrategia Tipo I construye este modelo secuencialmente, comenzando desde el modelo más simple posible y agregando términos gradualmente.

El modelo más simple posible para los datos sería uno en el que se suponga que ni la leche ni el azúcar tienen ningún efecto sobre el balbuceo. El único término que se incluiría en dicho modelo es la intersección, escrito como balbuceo ~ 1. Esta es nuestra hipótesis nula inicial. El siguiente modelo más simple para los datos sería uno en el que solo se incluye uno de los dos efectos principales. En los datos del café, hay dos opciones diferentes posibles, porque podríamos elegir agregar leche primero o azúcar primero. El orden realmente importa, como veremos más adelante, pero por ahora hagamos una elección arbitraria y escojamos azúcar. Entonces, el segundo modelo en nuestra secuencia de modelos es balbuceo ~ azúcar, y forma la hipótesis alternativa para nuestra primera prueba. Ahora tenemos nuestra primera prueba de hipótesis (Table 14.16).

Esta comparación forma nuestra prueba de hipótesis del efecto principal del azúcar. El siguiente paso en nuestro ejercicio de construcción de modelos es agregar el otro término de efecto principal, por lo que el siguiente modelo en nuestra secuencia es balbuceo ~ azúcar + leche. Luego, la segunda prueba de hipótesis se forma comparando el siguiente par de modelos (Table 14.17).

Esta comparación forma nuestra prueba de hipótesis del efecto principal de la leche. En cierto sentido, este enfoque es muy elegante: la hipótesis alternativa de la primera prueba forma la hipótesis nula de la segunda. Es en este sentido que el método Tipo I es estrictamente secuencial. Cada prueba se basa directamente en los resultados de la última. Sin embargo, en otro sentido es muy poco elegante, porque hay una fuerte asimetría entre las dos pruebas. La prueba del efecto principal del azúcar (la primera prueba) ignora por completo la leche, mientras que la prueba del efecto principal de la leche (la segunda prueba) sí tiene en cuenta el azúcar. En cualquier caso, el cuarto modelo de nuestra secuencia ahora es el modelo completo, balbuceo ~ azúcar + leche + azúcar × leche, y la prueba de hipótesis correspondiente se muestra en Table 14.18.

El método de prueba de hipótesis predeterminado utilizado por jamovi ANOVA es la suma de cuadrados Tipo III, por lo que para ejecutar un análisis de suma de cuadrados Tipo I, debemos seleccionar 'Tipo 1' en el cuadro de selección 'Suma de cuadrados' en las opciones de jamovi 'ANOVA' - Opciones de 'Modelo'. Esto nos da la tabla ANOVA que se muestra en Figure 14.27.

El gran problema con el uso de la suma de cuadrados Tipo I es el hecho de que realmente depende del orden en que ingresas las variables. Sin embargo, en muchas situaciones el investigador no tiene motivos para preferir un orden sobre otro. Este es presumiblemente el caso de nuestro problema de la leche y el azúcar. ¿Deberíamos agregar primero la leche o primero el azúcar? Es exactamente tan arbitrario como una pregunta de análisis de datos que como una pregunta de preparación de café. De hecho, puede haber algunas personas con opiniones firmes sobre el orden, pero es difícil imaginar una respuesta de

ANOVA

ANOVA - babble

	Sum of Squares	df	Mean Square	F	p
sugar	3.56	2	1.78	6.75	0.01086
milk	0.96	1	0.96	3.63	0.08106
sugar ✻ milk	5.94	2	2.97	11.28	0.00175
Residuals	3.16	12	0.26		

Figure 14.27: tabla de resultados de ANOVA utilizando la suma de cuadrados Tipo I en jamovi

principios a la pregunta. Sin embargo, mira lo que sucede cuando cambiamos el orden, como en Figure 14.28.

ANOVA

ANOVA - babble

	Sum of Squares	df	Mean Square	F	p
milk	1.44	1	1.44	5.48	0.03733
sugar	3.07	2	1.53	5.82	0.01708
milk ✻ sugar	5.94	2	2.97	11.28	0.00175
Residuals	3.16	12	0.26		

Figure 14.28: tabla de resultados ANOVA usando la suma de cuadrados Tipo I en jamovi, pero con los factores ingresados en un orden diferente (la leche primero)

Los valores p para ambos términos del efecto principal han cambiado, y de forma bastante drástica. Entre otras cosas, el efecto de la leche se ha vuelto significativo (aunque se debe evitar sacar conclusiones firmes al respecto, como mencioné anteriormente). ¿Cuál de estos dos ANOVA debe informarse? No es obvio de inmediato.

Cuando observas las pruebas de hipótesis que se utilizan para definir el "primer" efecto principal y el "segundo", está claro que son cualitativamente diferentes entre sí. En nuestro ejemplo inicial, vimos que la prueba del efecto principal del azúcar ignora por completo la leche, mientras que la prueba del efecto principal de la leche sí tiene en cuenta el azúcar. Como tal, la estrategia de prueba Tipo I realmente trata el primer efecto principal como si tuviera una especie de primacía teórica sobre el segundo. En mi experiencia, muy rara vez hay primacía teórica de este tipo que justifique tratar cualquiera de los dos efectos principales de forma asimétrica.

La consecuencia de todo esto es que las pruebas de Tipo I rara vez son de mucho interés, por lo que deberíamos pasar a hablar de las pruebas de Tipo II y las pruebas de Tipo III.

14.10.4 Suma de Cuadrados Tipo III

Habiendo terminado de hablar sobre las pruebas de Tipo I, podrías pensar que lo más natural a hacer a continuación sería hablar sobre las pruebas de Tipo II. Sin embargo, creo que en realidad es un poco más natural discutir las pruebas de Tipo III (que son simples y predeterminadas en jamovi ANOVA) antes de hablar de las pruebas de Tipo II (que son más complicadas). La idea básica que subyace a las pruebas de Tipo III es extremadamente simple. Independientemente del término que intentes evaluar, ejecuta la prueba F en la que la hipótesis alternativa corresponde al modelo ANOVA completo según lo especificado por el usuario, y el modelo nulo simplemente elimina ese término que estás probando. Por ejemplo, en el ejemplo del café, en el que nuestro modelo completo era balbuceo ~ azúcar + leche + azúcar × leche, la prueba del efecto principal del azúcar correspondería a una comparación entre los siguientes dos modelos (Table 14.19).

De manera similar, el efecto principal de la leche se evalúa probando el modelo completo contra un modelo nulo que elimina el término leche, como en Table 14.20.

Finalmente, el término de interacción azúcar × leche se evalúa exactamente de la misma manera. Una vez más, probamos el modelo completo con un modelo nulo que elimina el término de interacción azúcar × leche, como en Table 14.21.

La idea básica se generaliza a ANOVA de orden superior. Por ejemplo, supongamos que intentáramos ejecutar un ANOVA con tres factores, A, B y C, y deseáramos considerar todos los efectos principales posibles y todas las interacciones posibles, incluida la interacción de tres vías A × B $×$ C. (Table 14.22) te muestra cómo son las pruebas de Tipo III para esta situación).

Por fea que parezca esa tabla, es bastante simple. En todos los casos, la hipótesis alternativa corresponde al modelo completo que contiene tres términos de efectos principales (p. ej., A), tres interacciones de dos vías (p. ej., A*B) y una interacción de tres vías (p. ej., A*B* C)). El modelo nulo siempre contiene 6 de estos 7 términos, y el que falta es aquel cuyo significado estamos tratando de probar.

A primera vista, las pruebas de Tipo III parecen una buena idea. En primer lugar, eliminamos la asimetría que nos causaba problemas al ejecutar las pruebas de Tipo I. Y como ahora estamos tratando todos los términos de la misma manera, los resultados de las pruebas de hipótesis no dependen del orden en que los especifiquemos. Esto es definitivamente algo bueno. Sin embargo, existe un gran problema al interpretar los resultados de las pruebas, especialmente para los términos de efecto principal. Considera los datos del café. Supongamos que resulta que el efecto principal de la leche no es significativo según las pruebas de Tipo III. Lo que esto nos dice es que balbucear ~ azúcar + azúcar*leche es un modelo mejor para los datos que el modelo completo. Pero, ¿qué significa eso? Si el término de interacción azúcar*leche tampoco fuera significativo, estaríamos tentados a concluir que los datos nos dicen que lo único que importa es el azúcar. Pero supongamos que tenemos un término de interacción significativo, pero un efecto principal no significativo de la leche. En este caso, ¿debemos suponer que realmente hay un "efecto del azúcar", una "interacción entre la leche y el azúcar", pero no

un "efecto de la leche"? Eso parece una locura. La respuesta correcta simplemente debe ser que no tiene sentido[12] hablar sobre el efecto principal si la interacción es significativa. En general, esto parece ser lo que la mayoría de los estadísticos nos aconsejan hacer, y creo que ese es el consejo correcto. Pero si realmente no tiene sentido hablar de efectos principales no significativos en presencia de una interacción significativa, entonces no es del todo obvio por qué las pruebas de Tipo III deben permitir que la hipótesis nula se base en un modelo que incluye la interacción pero omite una de los principales efectos que lo componen. Cuando se caracterizan de esta manera, las hipótesis nulas realmente no tienen mucho sentido.

Más adelante, veremos que las pruebas de Tipo III se pueden canjear en algunos contextos, pero primero echemos un vistazo a la tabla de resultados de ANOVA usando la suma de cuadrados de Tipo III, consulta Figure 14.29.

ANOVA

ANOVA - babble

	Sum of Squares	df	Mean Square	F	p
milk	1.00	1	1.00	3.81	0.07467
sugar	2.13	2	1.07	4.04	0.04543
milk ✻ sugar	5.94	2	2.97	11.28	0.00175
Residuals	3.16	12	0.26		

Figure 14.29: tabla de resultados de ANOVA utilizando la suma de cuadrados Tipo III en jamovi

Pero ten en cuenta que una de las características perversas de la estrategia de prueba de Tipo III es que, por lo general, los resultados dependen de los contrastes que utilizas para codificar tus factores (consulta la sección Diferentes formas de especificar contrastes si has olvidado cuáles son los diferentes tipos de contrastes).[13]

De acuerdo, si los valores de p que normalmente surgen de los análisis de Tipo III (pero no en jamovi) son tan sensibles a la elección de los contrastes, ¿significa eso que las pruebas de Tipo III son esencialmente arbitrarias y no fiables? Hasta cierto punto, eso es cierto, y cuando pasemos a una discusión sobre las pruebas de Tipo II, veremos que los análisis de Tipo II evitan esta arbitrariedad por completo, pero creo que es una conclusión demasiado firme. En primer lugar, es importante reconocer que algunas elecciones de contrastes siempre producirán las mismas respuestas (ah, esto es lo que sucede en jamovi). De particular importancia es el hecho de que si las columnas de nuestra matriz de contraste están todas restringidas para sumar cero, entonces el análisis Tipo III siempre dará las mismas respuestas.

[12] O, como mínimo, rara vez de interés.

[13] Sin embargo, en jamovi los resultados para el ANOVA de suma de cuadrados Tipo III son los mismos independientemente del contraste seleccionado, ¡así que jamovi obviamente está haciendo algo diferente!

En las pruebas de Tipo II veremos que los análisis de Tipo II evitan esta arbitrariedad por completo, pero creo que es una conclusión demasiado fuerte. En primer lugar, es importante reconocer que algunas elecciones de contrastes siempre producirán las mismas respuestas (ah, esto es lo que sucede en jamovi). De particular importancia es el hecho de que si las columnas de nuestra matriz de contraste están todas restringidas para sumar cero, entonces el análisis Tipo III siempre dará las mismas respuestas.

14.10.5 Suma de Cuadrados Tipo II

Bien, ahora hemos visto las pruebas Tipo I y III, y ambas son bastante sencillas. Las pruebas de tipo I se realizan agregando gradualmente los términos uno a la vez, mientras que las pruebas de tipo III se realizan tomando el modelo completo y observando qué sucede cuando eliminas cada término. Sin embargo, ambos pueden tener algunas limitaciones. Las pruebas de tipo I dependen del orden en que ingresas los términos, y las pruebas de tipo III dependen de cómo codifiques tus contrastes. Las pruebas de tipo II son un poco más difíciles de describir, pero evitan ambos problemas y, como resultado, son un poco más fáciles de interpretar.

Las pruebas de tipo II son muy similares a las pruebas de tipo III. Comienzas con un modelo "completo" y pruebas un término en particular eliminándolo de ese modelo. Sin embargo, las pruebas de Tipo II se basan en el principio de marginalidad que establece que no debes omitir un término de orden inferior de tu modelo si hay términos de orden superior que dependen de él. Entonces, por ejemplo, si tu modelo contiene la interacción bidireccional A × B (un término de segundo orden), entonces realmente deberías contener los efectos principales A y B (términos de primer orden). De manera similar, si contiene un término de interacción triple A × B × C, entonces el modelo también debe incluir los efectos principales A, B y C, así como las interacciones más simples A × B, A × C y B × C. Las pruebas de tipo III violan rutinariamente el principio de marginalidad. Por ejemplo, considera la prueba del efecto principal de A en el contexto de un ANOVA de tres vías que incluye todos los términos de interacción posibles. De acuerdo con las pruebas Tipo III, nuestros modelos nulo y alternativo están en Table 14.23.

Fíjate que la hipótesis nula omite A, pero incluye A × B, A × C y A × B × C como parte del modelo. Esto, de acuerdo con las pruebas de Tipo II, no es una buena elección de hipótesis nula. En cambio, lo que deberíamos hacer, si queremos probar la hipótesis nula de que A no es relevante para nuestro resultado, es especificar la hipótesis nula que es el modelo más complicado que no se basa en ninguna forma de A, incluso como una interacción. La hipótesis alternativa corresponde a este modelo nulo más un término de efecto principal de A. Esto está mucho más cerca de lo que la mayoría de la gente pensaría intuitivamente como un "efecto principal de A", y produce lo siguiente como nuestra prueba Tipo II del efecto principal de A (Table 14.24). [14]

De todos modos, solo para darte una idea de cómo se desarrollan las pruebas Tipo II, aquí está la tabla completa (Table 14.25) de las pruebas que se aplicarían en un ANOVA factorial de tres vías:

[14] Ten en cuenta, por supuesto, que esto depende del modelo que especificó el usuario. Si el modelo ANOVA original no contiene un término de interacción para $B \times C$, obviamente no aparecerá ni en el valor nulo ni en el alternativo. Pero eso es cierto para los Tipos I, II y III. Nunca incluyen ningún término que no hayas incluido, pero toman decisiones diferentes sobre cómo construir pruebas para los que sí incluiste.

En el contexto del ANOVA de dos vías que hemos estado usando en los datos del café, las pruebas de hipótesis son aún más simples. El efecto principal del azúcar corresponde a una prueba F que compara estos dos modelos (Table 14.26).

La prueba del efecto principal de la leche está en Table 14.27.

Finalmente, la prueba para la interacción azúcar × leche está en Table 14.28.

Ejecutar las pruebas vuelve a ser sencillo. Simplemente selecciona 'Tipo 2' en el cuadro de selección 'Suma de cuadrados' en las opciones jamovi 'ANOVA' - 'Modelo'. Esto nos da la tabla ANOVA que se muestra en Figure 14.30.

ANOVA

ANOVA - babble

	Sum of Squares	df	Mean Square	F	p	η^2	$\eta^2 p$	ω^2
milk	0.96	1	0.96	3.63	0.08106	0.07	0.23	0.05
sugar	3.07	2	1.53	5.82	0.01708	0.23	0.49	0.19
milk ✻ sugar	5.94	2	2.97	11.28	0.00175	0.45	0.65	0.40
Residuals	3.16	12	0.26					

Figure 14.30: **?(caption)**

Las pruebas de tipo II tienen algunas ventajas claras sobre las pruebas de tipo I y tipo III. No dependen del orden en que especificas los factores (a diferencia del Tipo I), y no dependen de los contrastes que usas para especificar tus factores (a diferencia del Tipo III). Y aunque las opiniones pueden diferir sobre este último punto, y definitivamente dependerá de lo que intentes hacer con sus datos, creo que es más probable que las pruebas de hipótesis que especificas correspondan a algo que realmente te interese. Como consecuencia, encuentro que por lo general es más fácil interpretar los resultados de una prueba Tipo II que los resultados de una prueba Tipo I o Tipo III. Por esta razón, mi consejo tentativo es que, si no puedes pensar en ninguna comparación de modelos obvia que se corresponda directamente con tus preguntas de investigación, pero aun así deseas ejecutar un ANOVA en un diseño no balanceado, las pruebas de Tipo II son probablemente una mejor opción que las de Tipo I o Tipo III.[15]

[15] me parece divertido notar que el valor predeterminado en R es Tipo I y el valor predeterminado en SPSS y jamovi es Tipo III. Ninguno de estos me atrae tanto. En relación con esto, encuentro deprimente que casi nadie en la literatura psicológica se moleste en informar qué tipo de pruebas realizaron, y mucho menos el orden de las variables (para el Tipo I) o los contrastes utilizados (para el Tipo III). A menudo tampoco informan qué software usaron. La única forma en que puedo entender lo que la gente suele informar es tratar de adivinar a partir de pistas auxiliares qué software estaban usando y asumir que nunca cambiaron la configuración predeterminada. ¡Por favor, no hagas esto! Ahora que conoces estos problemas, asegúrate de indicar qué software usaste y, si estás informando los resultados de ANOVA para datos desequilibrados, especifica qué Tipo de pruebas ejecutaste, especifica la información del orden de los factores si has realizado pruebas Tipo I y especifica contrastes si has hecho pruebas de tipo III. O, mejor aún, ¡haz pruebas de hipótesis que correspondan a las cosas que realmente te importan y luego infórmalas!

14.10.6 Tamaños de los efectos (y sumas de cuadrados no aditivas)

jamovi también proporciona los tamaños de efecto η^2 y η^2 parcial cuando seleccionas estas opciones, como en Figure 14.30. Sin embargo, un diseño desequilibrado involucra un poco de complejidad adicional.

Si recuerdas nuestras primeras discusiones sobre ANOVA, una de las ideas clave que subyacen a los cálculos de sumas de cuadrados es que si sumamos todos los términos SC asociados con los efectos en el modelo, y lo sumamos a la SC residual, se supone que suman la suma de cuadrados total. Y, además de eso, la idea detrás de η^2 es que, debido a que estás dividiendo uno de los términos de SC por el valor total de SC, un valor de η^2 puede interpretarse como la proporción de la varianza explicada por un término particular. Pero esto no es tan sencillo en los diseños desequilibrados porque parte de la varianza "desaparece".

Esto parece un poco extraño al principio, pero he aquí por qué. Cuando tienes diseños desequilibrados, tus factores se correlacionan entre sí, y se vuelve difícil distinguir la diferencia entre el efecto del Factor A y el efecto del Factor B. En el caso extremo, supon que ejecutaríamos un diseño \$2 ×2 \$ en el que el número de participantes en cada grupo había sido como en Table 14.29.

Aquí tenemos un diseño espectacularmente desequilibrado: 100 personas tienen leche y azúcar, 100 personas no tienen leche ni azúcar, y eso es todo. Hay 0 personas con leche y sin azúcar y 0 personas con azúcar pero sin leche. Ahora imagina que, cuando recolectamos los datos, resultó que hay una gran diferencia (y estadísticamente significativa) entre el grupo "leche y azúcar" y el grupo "sin leche y sin azúcar". ¿Es este un efecto principal del azúcar? ¿Un efecto principal de la leche? ¿O una interacción? Es imposible saberlo, porque la presencia de azúcar tiene una asociación perfecta con la presencia de leche. Ahora supongamos que el diseño hubiera sido un poco más equilibrado (Table 14.30).

Esta vez, es técnicamente posible distinguir entre el efecto de la leche y el efecto del azúcar, porque algunas personas tienen uno pero no el otro. Sin embargo, seguirá siendo bastante difícil hacerlo, porque la asociación entre el azúcar y la leche sigue siendo extremadamente fuerte y hay muy pocas observaciones en dos de los grupos. Una vez más, es muy probable que estemos en una situación en la que *sabemos* que las variables predictoras (leche y azúcar) están relacionadas con el resultado (balbuceo), pero no sabemos si la naturaleza de esa relación es el efecto principal de un predictor u otro, o de la interacción.

14.11 Resumen

- [ANOVA factorial 1: diseños balanceados, sin interacciones] y con interacciones incluidas
- Tamaño del efecto, medias estimadas e intervalos de confianza en un ANOVA factorial
- [Comprobación de suposiciones] en ANOVA
- Análisis de Covarianza (ANCOVA)
- Entender [ANOVA como un modelo lineal], incluyendo Diferentes formas de especificar contrastes

- Pruebas post hoc utilizando el HSD de Tukey y un breve comentario sobre El método de las comparaciones planificadas
- [ANOVA factorial 3: diseños desequilibrados]

Table 14.15: tipos de contrastes disponibles en el análisis jamovi ANOVA

Contrast type	
Deviation	Compares the mean of each level (except a reference category) to the mean of all of the levels (grand mean)
Simple	Like the treatment contrasts, the simple contrast compares the mean of each level to the mean of a specified level. This type of contrast is useful when there is a control group. By default the first category is the reference. However, with a simple contrast the intercept is the grand mean of all the levels of the factors.
Difference	Compares the mean of each level (except the first) to the mean of previous levels. (Sometimes called reverse Helmert contrasts)
Helmert	Compares the mean of each level of the factor (except the last) to the mean of subsequent levels
Repeated	Compares the mean of each level (except the last) to the mean of the subsequent level
Polynomial	Compares the linear effect and quadratic effect. The first degree of freedom contains the linear effect across all categories; the second degree of freedom, the quadratic effect. These contrasts are often used to estimate polynomial trends

Table 14.16: Hipótesis nula y alternativa con la variable de resultado 'balbuceo'

Null model:	$babble \sim 1$
Alternative model:	$babble \sim sugar$

Table 14.17: más hipótesis nulas y alternativas con la variable de resultado 'balbuceo'

Null model:	$babble \sim sugar$
Alternative model:	$babble \sim sugar + milk$

Table 14.18: Y más hipótesis nulas y alternativas posibles con la variable de resultado 'balbuceo'

Null model:	$babble \sim sugar + milk$
Alternative model:	$babble \sim sugar + milk + sugar * milk$

Table 14.19: Hipótesis nula y alternativa con la variable de resultado 'balbuceo', con suma de cuadrados Tipo III

Null model:	$babble \sim milk + sugar * milk$
Alternative model:	$babble \sim sugar + milk + sugar * milk$

Table 14.20: Otras hipótesis nulas y alternativas con la variable de resultado 'balbuceo', con suma de cuadrados Tipo III

Null model:	$babble \sim sugar + sugar * milk$
Alternative model:	$babble \sim sugar + milk + sugar * milk$

Table 14.21: Eliminar el término de interacción de las hipótesis con la variable de resultado 'balbuceo', con suma de cuadrados Tipo III

Null model:	$babble \sim sugar + milk$
Alternative model:	$babble \sim sugar + milk + sugar * milk$

Table 14.22: pruebas de tipo III con tres factores y todos los efectos principales y términos de interacción

Term being tested is	**Null model is outcome ...**	**Alternative model is outcome ...**
A	$B + C + A * B + A * C + B * C + A * B * C$	$A + B + C + A * B + A * C + B * C + A * B * C$
B	$A + C + A * B + A * C + B * C + A * B * C$	$A + B + C + A * B + A * C + B * C + A * B * C$
C	$A + B + A * B + A * C + B * C + A * B * C$	$A + B + C + A * B + A * C + B * C + A * B * C$
A*B	$A + B + C + A * C + B * C + A * B * C$	$A + B + C + A * B + A * C + B * C + A * B * C$
A*C	$A + B + C + A * B + B * C + A * B * C$	$A + B + C + A * B + A * C + B * C + A * B * C$
B*C	$A + B + C + A * B + A * C + A * B * C$	$A + B + C + A * B + A * C + B * C + A * B * C$
A*B*C	$A + B + C + A * B + B + A * C + B * C$	$A + B + C + A * B + A * C + B * C + A * B * C$

Table 14.23: pruebas de tipo III para un efecto principal, A, en un ANOVA de tres vías con todos los términos de interacción posibles

Null model:	$outcome \sim B+C+A*B+A*C+B*C+A*B*C$
Alternative model:	$outcome \sim A+B+C+A*B+A*C+B*C+A*B*C$

Table 14.24: pruebas de tipo II para un efecto principal, A, en un ANOVA de tres vías con todos los términos de interacción posibles

Null model:	$outcome \sim B+C+B*C$
Alternative model:	$outcome \sim A+B+C+B*C$

Table 14.25: pruebas de tipo II para un modelo factorial de tres vías

Term being tested is	**Null model is outcome ...**	**Alternative model is outcome ...**
A	$B+C+B*C$	$A+B+C+B*C$
B	$A+C+A*C$	$A+B+C+A*C$
C	$A+B+A*B$	$A+B+C+A*B$
A*B	$A+B+C+A*C+B*C$	$A+B+C+A*B+A*C+B*C$
A*C	$A+B+C+A*B+B*C$	$A+B+C+A*B+A*C+B*C$
B*C	$A+B+C+A*B+A*C$	$A+B+C+A*B+A*C+B*C$
A*B*C	$A+B+C+A*B+A*C+B*C$	$A+B+C+A*B+A*C+B*C+A*B*C$

Table 14.26: Pruebas de tipo II para el efecto principal del azúcar en los datos del café

Null model:	$babble \sim milk$
Alternative model:	$babble \sim sugar + milk$

Table 14.27: Pruebas de tipo II para el efecto principal de la leche en los datos del café

Null model:	$babble \sim sugar$
Alternative model:	$babble \sim sugar + milk$

Table 14.28: Pruebas de tipo II para el término de interacción azúcar x leche

Null model:	$babble \sim sugar + milk$
Alternative model:	$babble \sim sugar+milk+sugar*milk$

Table 14.29: N participantes en un diseño factorial 2 x 2 muy (¡muy!) desequilibrado

	sugar	**no sugar**
milk	100	0
no milk	0	100

Table 14.30: N participantes en un diseño factorial 2 x 2 todavía muy desequilibrado

	sugar	**no sugar**
milk	100	5
no milk	5	100

Chapter 15

Análisis factorial

Los capítulos anteriores han cubierto las pruebas estadísticas para las diferencias entre dos o más grupos. Sin embargo, a veces, cuando realizamos una investigación, es posible que deseemos examinar cómo múltiples variables *co-varían*. Es decir, cómo se relacionan entre sí y si los patrones de relación sugieren algo interesante y significativo. Por ejemplo, a menudo nos interesa explorar si hay factores latentes no observados subyacentes que están representados por las variables observadas, medidas directamente, en nuestro conjunto de datos. En estadística, los factores latentes son inicialmente variables ocultas que no se observan directamente, sino que se infieren (a través del análisis estadístico) de otras variables que se observan (medidas directamente).

En este capítulo consideraremos una serie de análisis factorial diferentes y técnicas relacionadas, comenzando con Análisis factorial exploratorio (AFE). EFA es una técnica estadística para identificar factores latentes subyacentes en un conjunto de datos. Luego cubriremos Análisis de componentes principales (PCA), que es una técnica de reducción de datos que, estrictamente hablando, no identifica los factores latentes subyacentes. En cambio, PCA simplemente produce una combinación lineal de variables observadas. Después de esto, la sección sobre Análisis factorial confirmatorio (CFA) muestra que, a diferencia de EFA, con CFA se comienza con una idea, un modelo, de cómo las variables en sus datos se relacionan entre sí. Luego, prueba tu modelo con los datos observados y evalúa qué tan bueno es el ajuste del modelo. Una versión más sofisticada de CFA es el llamado enfoque [Multi-Trait Multi-Method CFA] en el que tanto el factor latente como la varianza del método se incluyen en el modelo. Esto es útil cuando se utilizan diferentes enfoques metodológicos para la medición y, por lo tanto, la variación del método es una consideración importante. Finalmente, cubriremos un análisis relacionado: Análisis de confiabilidad de consistencia interna prueba cuán consistentemente una escala mide una construcción psicológica.

15.1 Análisis factorial exploratorio

Análisis factorial exploratorio (AFE) es una técnica estadística para revelar cualquier factor latente oculto que se pueda inferir de nuestros datos observados. Esta técnica calcula hasta qué punto un conjunto de variables medidas, por ejemplo, $V1, V2, V3, V4$ y $V5$, pueden representarse como medidas de un factor latente subya-

cente. Este factor latente no puede medirse a través de una sola variable observada sino que se manifiesta en las relaciones que provoca en un conjunto de variables observadas.

En Figure 15.1 cada variable observada V es 'causada' hasta cierto punto por el factor latente subyacente (F), representado por los coeficientes b_1 a b_5 (también llamados factores de carga). Cada variable observada también tiene un término de error asociado, e1 a e5. Cada término de error es la varianza en la variable observada asociada, V_i , que no se explica por el factor latente subyacente.

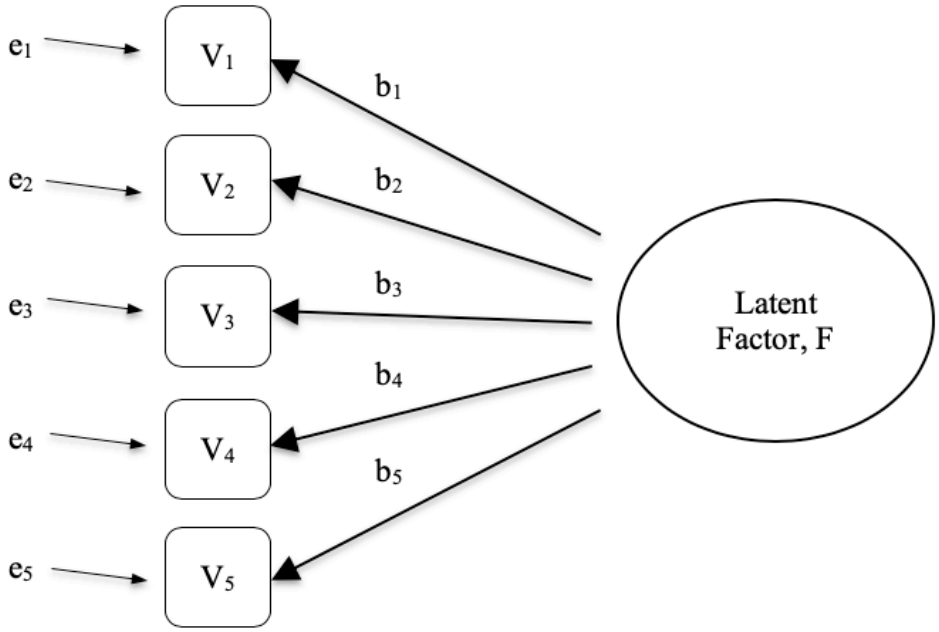


Figure 15.1: factor latente que subyace en la relación entre varias variables observadas

En Psicología, los factores latentes representan fenómenos o construcciones psicológicas que son difíciles de observar o medir directamente. Por ejemplo, personalidad, inteligencia o estilo de pensamiento. En el ejemplo de Figure 15.1, es posible que le hayamos hecho a las personas cinco preguntas específicas sobre su comportamiento o actitudes, y de eso podemos obtener una imagen sobre una construcción de personalidad llamada, por ejemplo, extraversión. Un conjunto diferente de preguntas específicas puede darnos una idea sobre la introversión de un individuo o su escrupulosidad.

Aquí hay otro ejemplo: es posible que no podamos medir directamente la ansiedad estadística, pero podemos medir si la ansiedad estadística es alta o baja con un conjunto de preguntas en un cuestionario. Por ejemplo, "$Q1$: Hacer la tarea para un curso de estadística", "$Q2$: Tratar de comprender las estadísticas descritas en un artículo de revista" y "$Q3$: Solicitar ayuda al profesor para comprender algo del curso", etc., cada uno calificado de baja ansiedad a alta ansiedad. Las personas con alta ansiedad estadística tenderán a dar respuestas igualmente altas en estas variables observadas debido a su alta ansiedad estadística. Del mismo modo, las personas con ansiedad

estadística baja darán respuestas bajas similares a estas variables debido a su ansiedad estadística baja.

En el análisis factorial exploratorio (AFE), esencialmente estamos explorando las correlaciones entre las variables observadas para descubrir cualquier factor subyacente (latente) interesante e importante que se identifique cuando las variables observadas covarían. Podemos usar software estadístico para estimar cualquier factor latente e identificar cuáles de nuestras variables tienen una carga alta[1] (por ejemplo, carga > 0.5) en cada factor, lo que sugiere que son una medida útil o indicador de el factor latente. Parte de este proceso incluye un paso llamado rotación, que para ser honesto es una idea bastante extraña pero afortunadamente no tenemos que preocuparnos por entenderlo; solo necesitamos saber que es útil porque hace que el patrón de cargas en diferentes factores sea mucho más claro. Como tal, la rotación ayuda a ver con mayor claridad qué variables están vinculadas sustancialmente a cada factor. También necesitamos decidir cuántos factores son razonables dados nuestros datos, y útil en este sentido es algo llamado valores propios. Volveremos a esto en un momento, después de que hayamos cubierto algunos de los principales supuestos de la EPT.

15.1.1 Comprobación de supuestos

Hay un par de suposiciones que deben verificarse como parte del análisis. La primera suposición es **esfericidad**, que esencialmente verifica que las variables en su conjunto de datos estén correlacionadas entre sí en la medida en que puedan resumirse potencialmente con un conjunto más pequeño de factores. La prueba de esfericidad de Bartlett verifica si la matriz de correlación observada diverge significativamente de una matriz de correlación cero (o nula). Entonces, si la prueba de Bartlett es significativa ($p < .05$), esto indica que la matriz de correlación observada es significativamente divergente de la nula y, por lo tanto, es adecuada para EFA.

La segunda suposición es la **adecuación del muestreo** y se verifica utilizando la Medida de Adecuación del Muestreo (MSA) de Kaiser-MeyerOlkin (KMO). El índice KMO es una medida de la proporción de varianza entre las variables observadas que podría ser una varianza común. Usando correlaciones parciales, busca factores que carguen solo dos elementos. Rara vez, si acaso, queremos que EFA produzca muchos factores cargando solo dos elementos cada uno. KMO se trata de la adecuación del muestreo porque las correlaciones parciales generalmente se ven con muestras inadecuadas. Si el índice KMO es alto (≈ 1), el EFA es eficiente, mientras que si KMO es bajo (≈ 0), el EFA no es relevante. Los valores de KMO inferiores a $0,5$ indican que EFA no es adecuado y debe haber un valor de KMO de $0,6$ antes de que EFA se considere adecuado. Los valores entre $0,5$ y $0,7$ se consideran adecuados, los valores entre $0,7$ y $0,9$ son buenos y los valores entre $0,9$ y $1,0$ son excelentes.

15.1.2 ¿Para qué sirve la EPT?

Si la EFA ha brindado una buena solución (es decir, un modelo de factores), entonces debemos decidir qué hacer con nuestros nuevos y brillantes factores. Los investigadores a menudo usan EFA durante el desarrollo de escalas psicométricas. Desarrollarán un conjunto de elementos del cuestionario que creen que se relacionan con uno o más constructos psicológicos, usarán EFA para ver qué elementos "van juntos" como factores

[1] muy útil, las cargas factoriales se pueden interpretar como coeficientes de regresión estandarizados

latentes y luego evaluarán si algunos elementos deben eliminarse porque no son útiles. o medir claramente uno de los factores latentes.

De acuerdo con este enfoque, otra consecuencia de EFA es combinar las variables que se cargan en distintos factores en un puntaje de factor, a veces conocido como puntaje de escala. Hay dos opciones para combinar variables en una puntuación de escala:

- Crear una nueva variable con una puntuación ponderada por las cargas factoriales de cada elemento que contribuye al factor.
- Crear una nueva variable a partir de cada ítem que contribuya al factor, pero ponderándolos por igual.

En la primera opción, la contribución de cada ítem a la puntuación combinada depende de qué tan fuertemente se relacione con el factor. En la segunda opción, generalmente solo promediamos todos los elementos que contribuyen sustancialmente a un factor para crear la variable de puntuación de escala combinada. Cuál elegir es una cuestión de preferencia, aunque una desventaja con la primera opción es que las cargas pueden variar bastante de una muestra a otra, y en las ciencias del comportamiento y de la salud, a menudo estamos interesados en desarrollar y usar puntajes de escala de cuestionarios compuestos en diferentes estudios. y diferentes muestras. En cuyo caso, es razonable utilizar una medida compuesta que se base en los elementos sustantivos que contribuyen por igual en lugar de ponderar por cargas específicas de muestra de una muestra diferente. En cualquier caso, entender una medida de variable combinada como un promedio de elementos es más simple e intuitivo que usar una combinación ponderada óptimamente específica de una muestra.

Una técnica estadística más avanzada, que está más allá del alcance de este libro, emprende el modelado de regresión donde los factores latentes se utilizan en modelos de predicción de otros factores latentes. Esto se denomina "modelado de ecuaciones estructurales" y existen programas de software específicos y paquetes R dedicados a este enfoque. Pero no nos adelantemos; en lo que realmente deberíamos centrarnos ahora es en cómo hacer un EFA en jamovi.

15.1.3 EPT en Jamovi

Primero, necesitamos algunos datos. Veinticinco ítems de autoinforme de personalidad (ver Figure 15.2) tomados del International Personality Item Pool fueron incluido como parte de la evaluación de personalidad basada en la web de Evaluación de personalidad de apertura sintética (SAPA) (SAPA: http://sapa-project.org) proyecto. Los 25 ítems están organizados por cinco factores putativos: Amabilidad, Escrupulosidad, Extraversión, Neuroticismo y Apertura.

Los datos de los ítems se recopilaron utilizando una escala de respuesta de 6 puntos:

1. Muy impreciso
2. Moderadamente impreciso
3. Ligeramente impreciso
4. Ligeramente preciso
5. Moderadamente preciso
6. Muy preciso.

Una muestra de $N = 250$ respuestas está contenida en el conjunto de datos bfi_sample.csv. Como investigadores, estamos interesados en explorar los datos para

ver si hay algunos factores latentes subyacentes que se miden razonablemente bien con las variables observadas de 25 en el archivo de datos bfi_sample.csv. Abra el conjunto de datos y verifique que las variables de 25 estén codificadas como variables continuas (técnicamente, son ordinales, aunque para EFA en jamovi en general no importa, excepto si decide calcular puntajes de factores ponderados, en cuyo caso se necesitan variables continuas) . Para realizar EFA en jamovi:

Variable name	Question / Item (short phrases that you should respond to by indicating how accurately the statement describes your typical behaviour or attitudes)	Coding (R: reverse)
A1	Am indifferent to the feelings of others.	R
A2	Inquire about others' well-being.	
A3	Know how to comfort others.	
A4	Love children.	
A5	Make people feel at ease.	
C1	Am exacting in my work.	
C2	Continue until everything is perfect.	
C3	Do things according to a plan.	
C4	Do things in a half-way manner.	R
C5	Waste my time.	R
E1	Don't talk a lot.	R
E2	Find it difficult to approach others.	R
E3	Know how to captivate people.	
E4	Make friends easily.	
E5	Take charge.	
N1	Get angry easily.	
N2	Get irritated easily.	
N3	Have frequent mood swings.	
N4	Often feel blue.	
N5	Panic easily.	
O1	Am full of ideas.	
O2	Avoid difficult reading material.	R
O3	Carry the conversation to a higher level.	
O4	Spend time reflecting on things.	
O5	Will not probe deeply into a subject.	R

Figure 15.2: veinticinco elementos variables observados organizados por cinco factores de personalidad putativos en el conjunto de datos bfi_sample.csv

- Seleccione Factor - Análisis factorial exploratorio en la barra de botones principal de jamovi para abrir la ventana de análisis EFA (Figure 15.3).
- Seleccione las 25 preguntas de personalidad y transfiéralas al cuadro 'Variables'.
- Marque las opciones apropiadas, incluidas las 'Comprobaciones de suposiciones', pero también las opciones de 'Método' de rotación, 'Número de factores' para

extraer y 'Salida adicional'. Consulte Figure 15.3 para ver las opciones sugeridas para este EFA ilustrativo, y tenga en cuenta que el 'Método' de rotación y el 'Número de factores' extraídos normalmente los ajusta el investigador durante el análisis para encontrar el mejor resultado, como se describe a continuación.

Primero, verifique las suposiciones (Figure 15.4). Puede ver que (1) la prueba de esfericidad de Bartlett es significativa, por lo que se cumple esta suposición; y (2) la medida de adecuación del muestreo (MSA) de KMO es de $ 0.81 $ en general, lo que sugiere una buena adecuación del muestreo. No hay problemas aquí entonces!

Lo siguiente que debe verificar es cuántos factores usar (o "extraer" de los datos). Hay tres enfoques diferentes disponibles:

- Una convención es elegir todos los componentes con valores propios mayores que 12 . Esto nos daría cuatro factores con nuestros datos (pruébalo y verás).
- El examen del diagrama de pantalla, como en Figure 15.5, le permite identificar el "punto de inflexión". Este es el punto en el que la pendiente de la curva del pedregal se nivela claramente, por debajo del "codo". Esto nos daría cinco factores con nuestros datos. Interpretar scree plots es un poco un arte: en Figure 15.5 hay un paso notable de 5 a 6 factores, pero en otros scree plots que mire no será tan claro.
- Mediante una técnica de análisis en paralelo, los valores propios obtenidos se comparan con los que se obtendrían a partir de datos aleatorios. El número de factores extraídos es el número con valores propios mayores que los que se encontrarían con datos aleatorios.

El tercer enfoque es bueno según Fabrigar et al. (1999), aunque en la práctica los investigadores tienden a observar los tres y luego emitir un juicio sobre la cantidad de factores que se interpretan de manera más fácil o útil. Esto puede entenderse como el "criterio de significado", y los investigadores normalmente examinarán, además de la solución de uno de los enfoques anteriores, soluciones con uno o dos factores más o menos. Luego adoptan la solución que tiene más sentido para ellos.

Al mismo tiempo, también debemos considerar la mejor manera de rotar la solución final. Hay dos enfoques principales para la rotación: la rotación ortogonal (p. ej., 'varimax') obliga a que los factores seleccionados no estén correlacionados, mientras que la rotación oblicua (p. ej., 'oblimin') permite correlacionar los factores seleccionados. Las dimensiones de interés para los psicólogos y los científicos del comportamiento a menudo no son dimensiones que esperaríamos que fueran ortogonales, por lo que las soluciones oblicuas son posiblemente más sensatas[2]

Prácticamente, si en una rotación oblicua se encuentra que los factores están sustancialmente correlacionados (positivo o negativo, y > 0.3), como en Figure 15.6 donde una correlación entre dos de los factores extraídos es 0.31, entonces esto confirmaría nuestra

[2]las rotaciones oblicuas proporcionan dos matrices de factores, una denominada matriz de estructura y otra denominada matriz de patrón. En jamovi, solo se muestra la matriz de patrones en los resultados, ya que suele ser la más útil para la interpretación, aunque algunos expertos sugieren que ambos pueden ser útiles. En una matriz de estructura, los coeficientes muestran la relación entre la variable y los factores mientras ignoran la relación de ese factor con todos los demás factores (es decir, una correlación de orden cero). Los coeficientes de matriz de patrones muestran la contribución única de un factor a una variable mientras controlan los efectos de otros factores en esa variable (similar al coeficiente de regresión parcial estandarizado). Bajo rotación ortogonal, los coeficientes de estructura y patrón son los mismos.

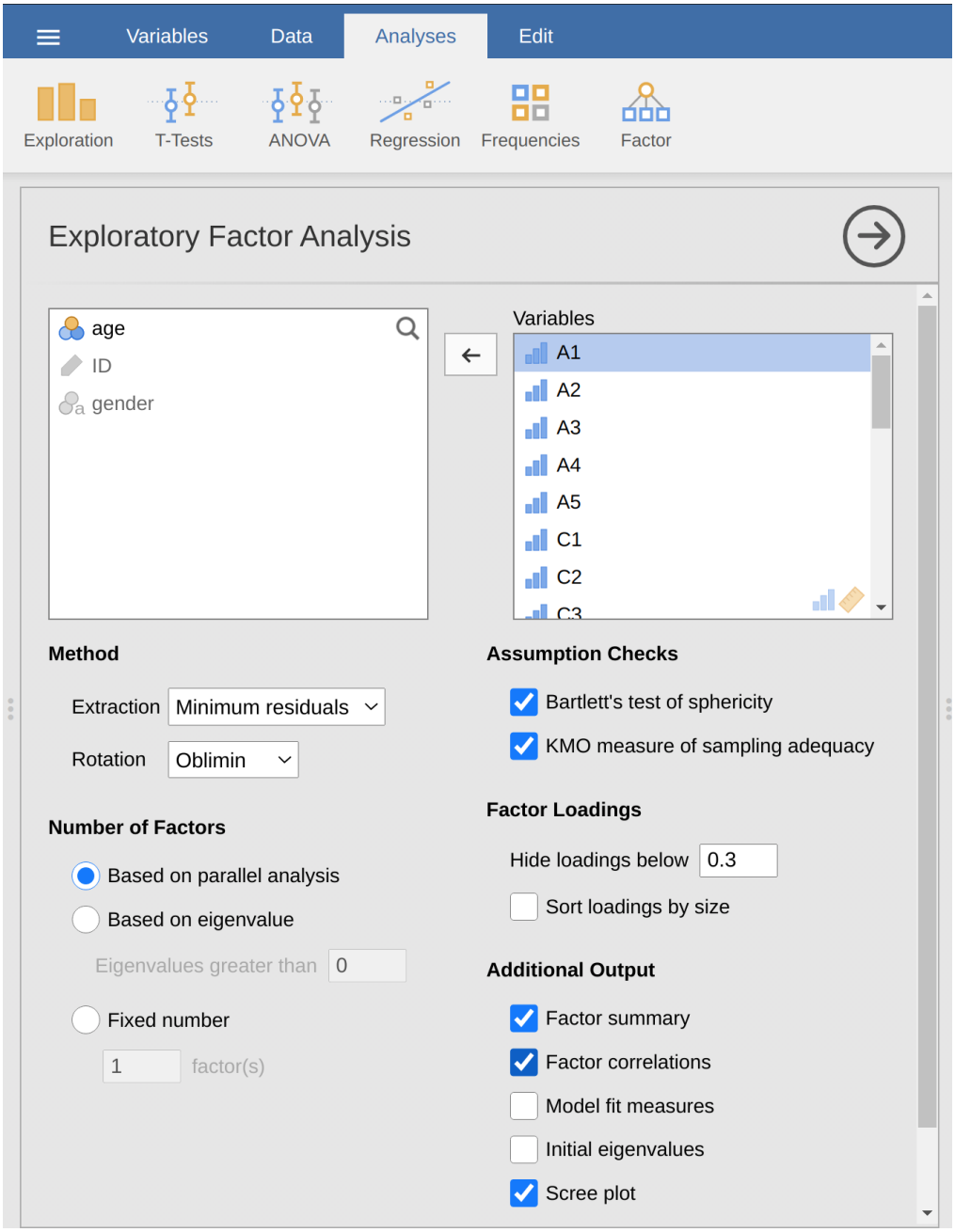


Figure 15.3: La ventana de análisis EFA jamovi

Assumption Checks

Bartlett's Test of Sphericity

χ^2	df	p
2204.28	300	< .00001

KMO Measure of Sampling Adequacy

	MSA
Overall	0.81
A1	0.59
A2	0.84
A3	0.82
A4	0.83
A5	0.86
C1	0.80
C2	0.81
C3	0.75
C4	0.79
C5	0.84
E1	0.83
E2	0.85
E3	0.83
E4	0.87
E5	0.91
N1	0.74
N2	0.71
N3	0.77
N4	0.83
N5	0.80
O1	0.84
O2	0.70
O3	0.82
O4	0.74
O5	0.76

Figure 15.4: jamovi EFA comprueba la suposición de los datos del cuestionario de personalidad

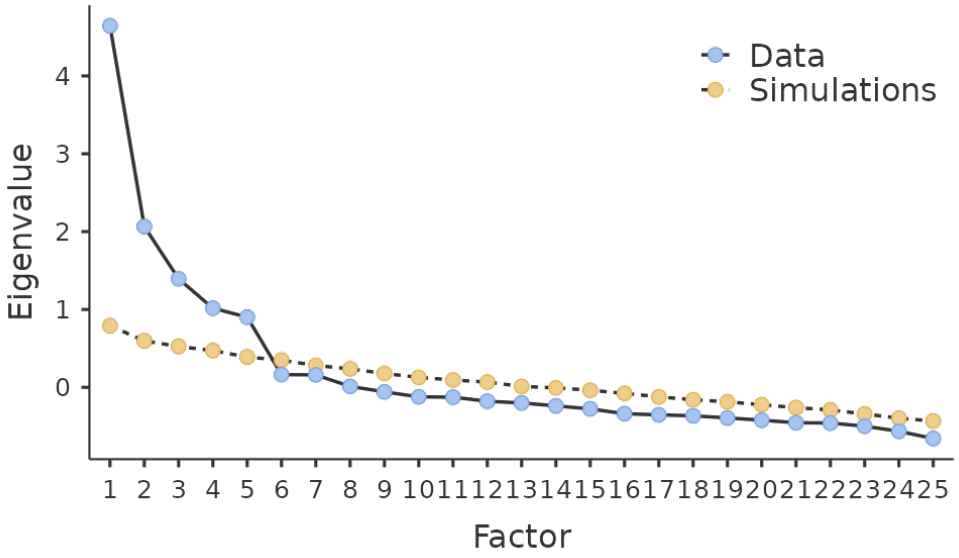


Figure 15.5: gráfico de pantalla de los datos de personalidad en jamovi EFA, que muestra una inflexión notable y se nivela después del punto 5 (el 'codo')

intuición para preferir la rotación oblicua. Si los factores están, de hecho, correlacionados, entonces una rotación oblicua producirá una mejor estimación de los verdaderos factores y una mejor estructura simple que una rotación ortogonal. Y, si la rotación oblicua indica que los factores tienen correlaciones cercanas a cero entre sí, entonces el investigador puede continuar y realizar una rotación ortogonal (que luego debería dar aproximadamente la misma solución que la rotación oblicua).

Al comprobar la correlación entre los factores extraídos, al menos una correlación fue superior a 0,3 (Figure 15.6), por lo que se prefiere una rotación oblicua ("oblimin") de los cinco factores extraídos. También podemos ver en Figure 15.6 que la proporción de la variación general en los datos que se explica por los cinco factores es del 46 %. El factor uno representa alrededor del 10% de la varianza, los factores dos a cuatro alrededor del 9% cada uno y el factor cinco un poco más del 7%. Esto no es genial; Hubiera sido mejor si la solución general explicara una proporción más sustancial de la varianza en nuestros datos.

Tenga en cuenta que en cada EFA podría tener potencialmente la misma cantidad de factores que variables observadas, pero cada factor adicional que incluya agregará una cantidad menor de varianza explicada. Si los primeros factores explican una buena cantidad de la varianza en las 25 variables originales, entonces esos factores son claramente un sustituto útil y más simple para las 25 variables. Puede eliminar el resto sin perder demasiado de la variabilidad original. Pero si se necesitan 18 factores (por ejemplo) para explicar la mayor parte de la variación en esas 25 variables, también podría usar los 25 originales.

Figure 15.7 muestra las cargas factoriales. Es decir, cómo se cargan los 25 elementos de

Factor Statistics

Summary

Factor	SS Loadings	% of Variance	Cumulative %
1	2.61	10.45	10.45
2	2.38	9.52	19.97
3	2.41	9.63	29.60
4	2.21	8.83	38.42
5	1.84	7.34	45.77

Inter-Factor Correlations

	1	2	3	4	5
1	—	−0.16	−0.16	0.02	−0.10
2		—	0.31	0.13	0.19
3			—	0.23	0.22
4				—	0.20
5					—

Figure 15.6: estadísticas de resumen de factores y correlaciones para una solución de cinco factores en jamovi EFA

personalidad diferentes en cada uno de los cinco factores seleccionados. Tenemos cargas ocultas menores a 0.3 (configuradas en las opciones que se muestran en Figure 15.3.

Para los Factores $1, 2, 3$ y 4, el patrón de las cargas factoriales coincide estrechamente con los factores putativos especificados en Figure 15.2. ¡Uf! Y el factor 5 está bastante cerca, con cuatro de las cinco variables observadas que supuestamente miden la "apertura" cargando bastante bien en el factor. Sin embargo, la variable 04 no parece encajar, ya que la solución factorial en Figure 15.7 sugiere que se carga en el factor 4 (aunque con una carga relativamente baja) pero no sustancialmente en el factor 5.

La otra cosa a tener en cuenta es que aquellas variables que se denotaron como "R: codificación inversa" en Figure 15.2 son aquellas que tienen cargas de factores negativas. Eche un vistazo a los ítems A1 ("Soy indiferente a los sentimientos de los demás") y A2 ("Pregunto por el bienestar de los demás"). Podemos ver que una puntuación alta en $A1$ indica baja simpatía, mientras que una puntuación alta en $A2$ (y todas las demás variables "A" para el caso) indica alta simpatía. Por lo tanto, A1 se correlacionará negativamente con las otras variables "A", y es por eso que tiene una carga factorial negativa, como se muestra en Figure 15.7.

También podemos ver en Figure 15.7 la "singularidad" de cada variable. La singularidad es la proporción de varianza que es 'única' para la variable y no explicada por los factores[3]. Por ejemplo, el 72% de la varianza en 'A1' no se explica por los factores de la solución de cinco factores. Por el contrario, 'N1' tiene una varianza relativamente baja que no se explica en la solución factorial (35 %). Nótese que cuanto mayor es la 'singularidad', menor es la relevancia o contribución de la variable en el modelo factorial.

Para ser honesto, es inusual obtener una solución tan clara en EPT. Por lo general, es un poco más complicado que esto y, a menudo, interpretar el significado de los factores es más desafiante. No es frecuente que tenga un grupo de artículos tan claramente delineado. Más a menudo, tendrá un montón de variables observadas que cree que pueden ser indicadores de algunos factores latentes subyacentes, ¡pero no tiene un sentido tan fuerte de qué variables van a ir a dónde!

Por lo tanto, parece que tenemos una solución de cinco factores bastante buena, aunque representa una proporción general relativamente baja de la varianza observada. Supongamos que estamos contentos con esta solución y queremos usar nuestros factores en análisis posteriores. La opción sencilla es calcular una puntuación general (promedio) para cada factor sumando la puntuación de cada variable que se carga sustancialmente en el factor y luego dividiendo por el número de variables (en otras palabras, crear una "puntuación media" para cada persona a través de los ítems para cada escala Para cada persona en nuestro conjunto de datos que implica, por ejemplo, para el factor de Amabilidad, sumando $A1 + A2 + A3 + A4 + A5$, y luego dividiendo por 5. [4] En esencia, el puntaje factorial que hemos calculado se basa en puntajes igualmente ponderados de cada una de las variables/ítems incluidos. Podemos hacer esto en jamovi en dos pasos:

- Recodifique A1 en "A1R" al revertir la puntuación de los valores en la variable (es decir, $6 = 1$; $5 = 2$; $4 = 3$; $3 = 4$; $2 = 5$; $1 = 6$) usando el jamovi comando transform variable (ver Figure 15.8).

[3] a veces se informa en el análisis factorial la "comunalidad", que es la cantidad de varianza en una variable que se explica por la solución factorial. Unicidad es igual a (1 $\sum$ comunalidad)

[4] recordar primero revertir la puntuación de algunas variables si es necesario

Exploratory Factor Analysis

Factor Loadings

	Factor					
	1	2	3	4	5	Uniqueness
A1				−0.48		0.72
A2				0.71		0.48
A3				0.68		0.44
A4				0.43		0.67
A5				0.52		0.53
C1		0.72				0.48
C2		0.68				0.52
C3		0.51				0.75
C4		−0.64				0.47
C5		−0.58				0.52
E1			−0.53			0.70
E2			−0.70			0.35
E3			0.49		0.39	0.48
E4			0.59	0.36		0.43
E5			0.42			0.57
N1	0.79					0.35
N2	0.79					0.40
N3	0.69					0.47
N4	0.50		−0.45			0.45
N5	0.47					0.63
O1					0.57	0.63
O2					−0.49	0.68
O3					0.69	0.41
O4				0.33		0.75
O5					−0.52	0.68

Note. 'Minimum residual' extraction method was used in combination with a 'oblimin' rotation

Figure 15.7: Cargas factoriales para una solución de cinco factores en jamovi EFA

- Calcule una nueva variable, llamada “Agradabilidad”, calculando la media de A1R, A2, A3, A4 y A5. Hágalo con el comando jamovi compute new variable (ver Figure 15.9).

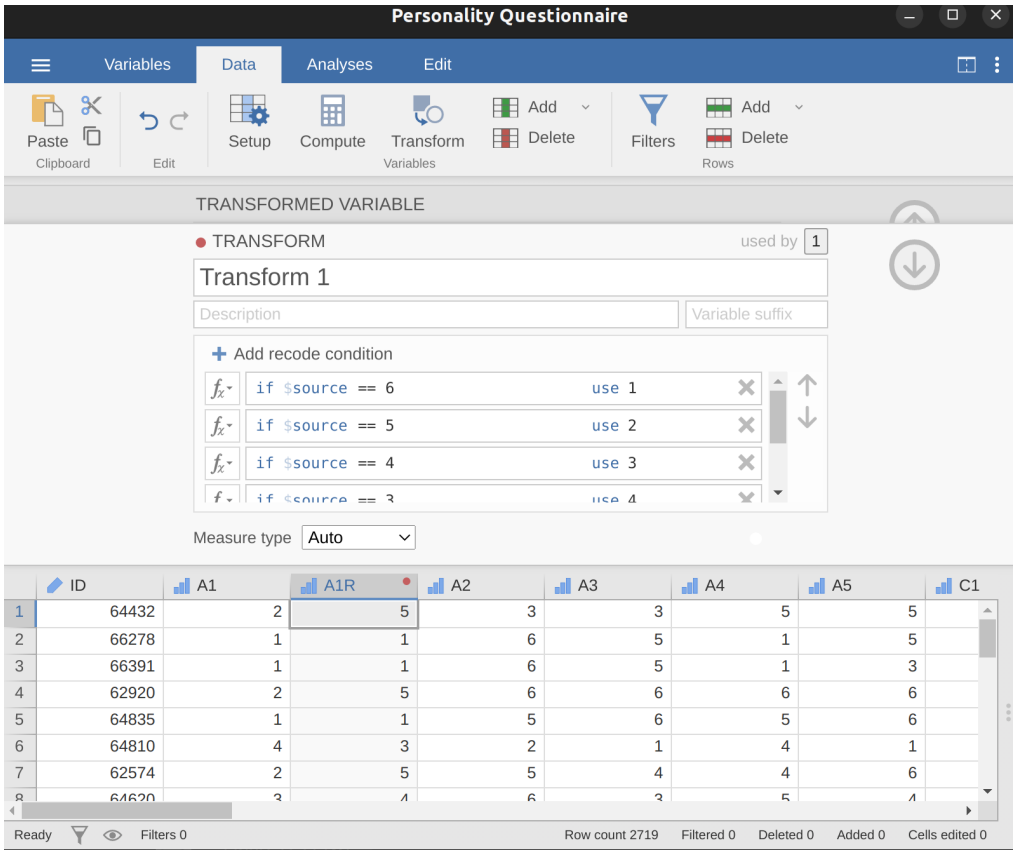


Figure 15.8: Recodificar variable usando el comando Transformar jamovi

Otra opción es crear un índice de puntaje factorial ** ponderado de manera óptima **. Para hacer esto, guarde las puntuaciones de los factores en el conjunto de datos, usando la casilla de verificación ‘Guardar’ - ‘Puntuaciones de los factores’. Una vez hecho esto verás que se han añadido cinco nuevas variables (columnas) a los datos, una por cada factor extraído. Ver Figure 15.10 y Figure 15.11.

Ahora puede continuar y realizar más análisis, utilizando las escalas factoriales basadas en la puntuación media (p. ej., como en Figure 15.9) o utilizando las puntuaciones factoriales ponderadas de forma óptima calculadas por jamovi. ¡Tu elección! Por ejemplo, una cosa que le gustaría hacer es ver si hay diferencias de género en cada una de nuestras escalas de personalidad. Hicimos esto para la puntuación de Amabilidad que calculamos utilizando el enfoque de puntuación media, y aunque la gráfica de la prueba t (Figure 15.12) mostró que los hombres eran menos agradables que las mujeres, esto no fue una diferencia significativa (Mann-Whitney $ U = 5768$, $p = .075$).

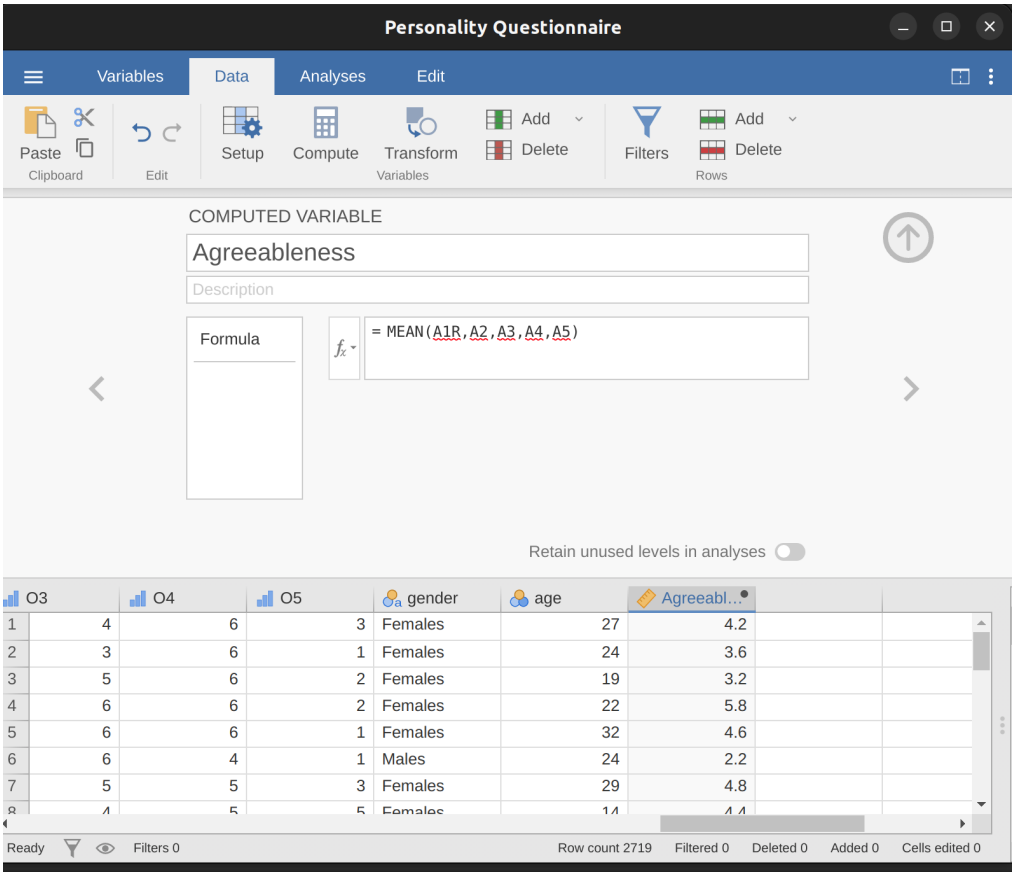


Figure 15.9: Calcule la nueva variable de puntaje de escala usando el comando de variable computada jamovi

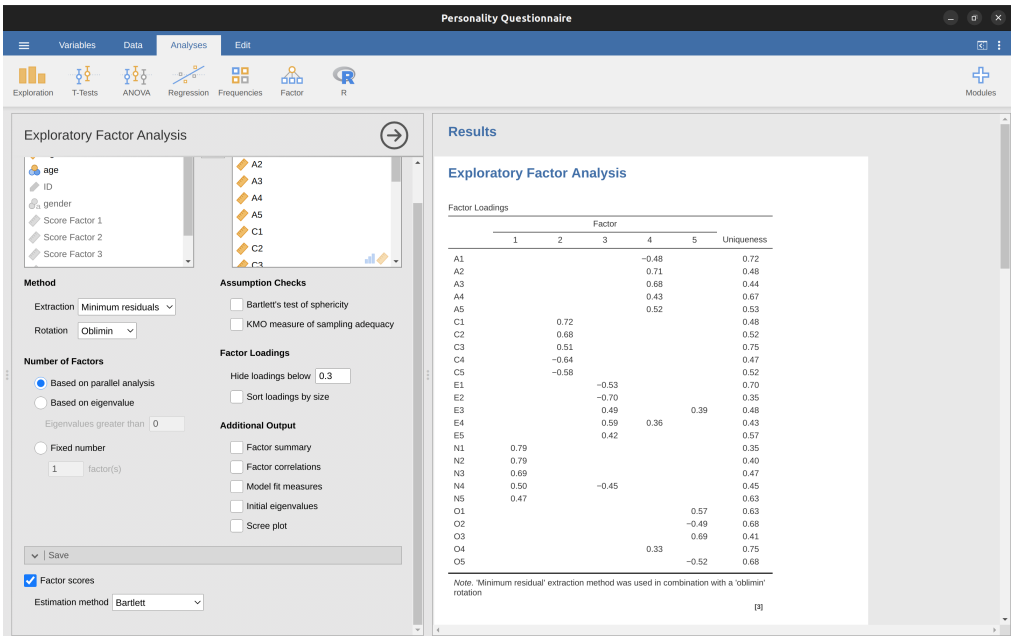


Figure 15.10: opción jamovi para puntajes factoriales para la solución de cinco factores, utilizando el método de ponderación óptima 'Bartlett'

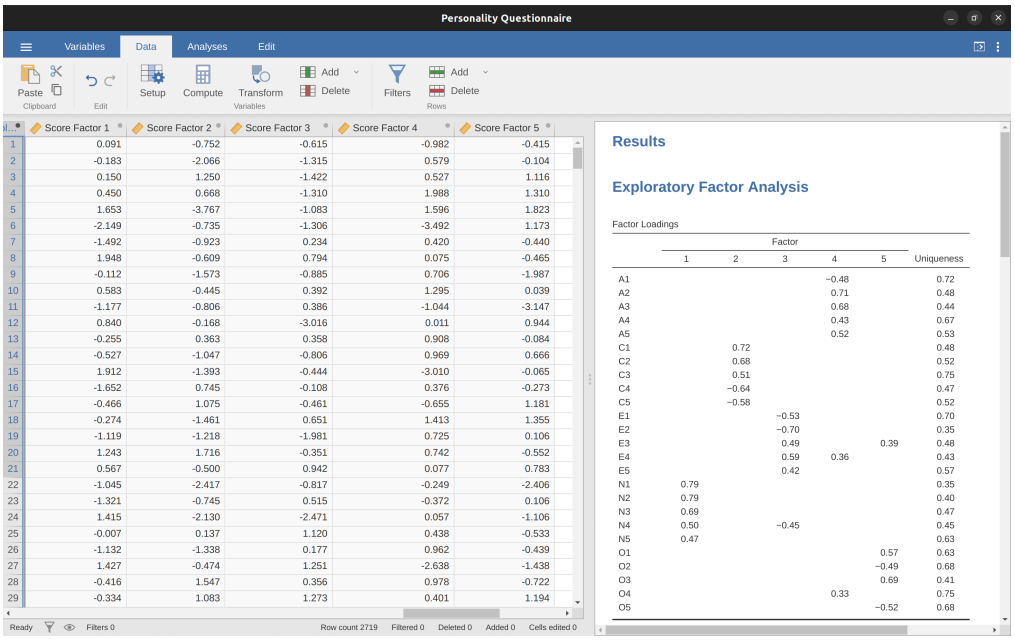


Figure 15.11: vista de hoja de datos que muestra las cinco variables de puntaje factorial recién creadas

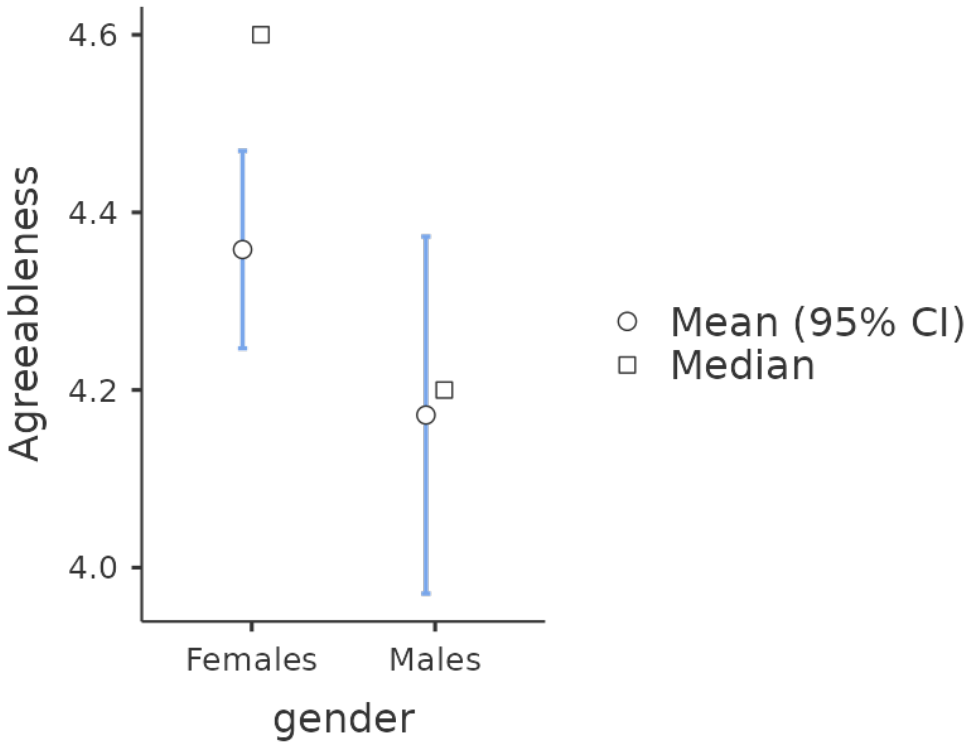


Figure 15.12: Comparación de las diferencias en las puntuaciones basadas en factores de simpatía entre hombres y mujeres

15.1.4 Escribir un EFA

Con suerte, hasta ahora le hemos dado una idea de la EPT y cómo llevar a cabo la EPT en jamovi. Entonces, una vez que haya completado su EFA, ¿cómo lo escribe? No existe una forma estándar formal de redactar una EFA, y los ejemplos tienden a variar según la disciplina y el investigador. Dicho esto, hay algunas piezas de información bastante estándar para incluir en su redacción:

1. ¿Cuáles son los fundamentos teóricos para el área que está estudiando, y específicamente para los constructos que le interesa descubrir a través de EPT?
2. Una descripción de la muestra (por ejemplo, información demográfica, tamaño de la muestra, método de muestreo).
3. Una descripción del tipo de datos utilizados (p. ej., nominales, continuos) y estadísticas descriptivas.
4. Describa cómo hizo para probar los supuestos de la EFA. Se deben informar los detalles sobre los controles de esfericidad y las medidas de adecuación del muestreo.
5. Explique qué método de extracción de FA (p. ej., 'Residuos mínimos' o 'Máxima probabilidad') se utilizó.
6. Explique los criterios y el proceso utilizado para decidir cuántos factores se extrajeron en la solución final y qué elementos se seleccionaron. Explique claramente la justificación de las decisiones clave durante el proceso de EFA.
7. Explique qué métodos de rotación se intentaron, los motivos y los resultados.
8. Las cargas factoriales finales deben informarse en los resultados, en una tabla. Esta tabla también debe informar la singularidad (o comunalidad) de cada variable (en la última columna). Las cargas factoriales deben informarse con etiquetas descriptivas además de los números de artículo. Las correlaciones entre los factores también deben incluirse, ya sea en la parte inferior de esta tabla, en una tabla separada.
9. Deben proporcionarse nombres significativos para los factores extraídos. Es posible que desee utilizar nombres de factores previamente seleccionados, pero al examinar los elementos y factores reales, puede pensar que un nombre diferente es más apropiado.

15.2 Análisis de componentes principales

En la sección anterior vimos que EPT trabaja para identificar factores latentes subyacentes. Y, como vimos, en un escenario, el número más pequeño de factores latentes se puede usar en análisis estadísticos adicionales usando algún tipo de puntajes de factores combinados.

De esta manera, EFA se está utilizando como una técnica de "reducción de datos". Otro tipo de técnica de reducción de datos, a veces vista como parte de la familia EFA, es el **análisis de componentes principales (PCA)**. Sin embargo, PCA no identifica factores latentes subyacentes. En su lugar, crea una puntuación compuesta lineal a partir de un conjunto más grande de variables medidas.

PCA simplemente produce una transformación matemática a los datos originales sin suposiciones sobre cómo las variables co-varían. El objetivo de PCA es calcular algunas combinaciones lineales (componentes) de las variables originales que se pueden usar para resumir el conjunto de datos observados sin perder mucha información. Sin embargo, si la identificación de la estructura subyacente es un objetivo del análisis, entonces se prefiere EFA. Y, como vimos, EFA produce puntajes factoriales que se pueden usar para propósitos de reducción de datos al igual que los puntajes de componentes principales (Fabrigar et al., 1999).

PCA ha sido popular en psicología por varias razones y, por lo tanto, vale la pena mencionarlo, aunque hoy en día EFA es tan fácil de hacer dada la potencia de las computadoras de escritorio y puede ser menos susceptible al sesgo que PCA, especialmente con una pequeña cantidad de factores. y variables. Gran parte del procedimiento es similar a EFA, por lo que, aunque existen algunas diferencias conceptuales, prácticamente los pasos son los mismos, y con muestras grandes y un número suficiente de factores y variables, los resultados de PCA y EFA deberían ser bastante similares.

Para realizar PCA en jamovi, todo lo que necesita hacer es seleccionar 'Factor' - 'Análisis de componentes principales' en la barra de botones principal de jamovi para abrir la ventana de análisis de PCA. Luego puede seguir los mismos pasos de [EFA en jamovi] arriba.

15.3 Análisis factorial confirmatorio

Por lo tanto, nuestro intento de identificar los factores latentes subyacentes utilizando EFA con preguntas cuidadosamente seleccionadas del conjunto de elementos de personalidad pareció tener bastante éxito. El próximo paso en nuestra búsqueda para desarrollar una medida útil de la personalidad es verificar los factores latentes que identificamos en el EFA original con una muestra diferente. Queremos ver si los factores se mantienen, si podemos confirmar su existencia con datos diferentes. Esta es una verificación más rigurosa, como veremos. Y se llama **Análisis factorial confirmatorio (CFA)** ya que, como era de esperar, buscaremos confirmar una estructura de factor latente especificada previamente.[5]

En CFA, en lugar de hacer un análisis en el que vemos cómo los datos van juntos en un sentido exploratorio, imponemos una estructura, como en Figure 15.13, sobre los datos y vemos qué tan bien se ajustan a nuestros datos preespecificados. estructura. En este sentido, estamos realizando un análisis confirmatorio, para ver qué tan bien los datos observados confirman un **modelo** preespecificado.

Por lo tanto, un análisis factorial confirmatorio (CFA) directo de los ítems de personalidad especificaría cinco factores latentes como se muestra en Figure 15.13, cada uno medido por cinco variables observadas. Cada variable es una medida de un factor latente subyacente. Por ejemplo, A1 se predice por el factor latente subyacente Amabilidad. Y debido a que A1 no es una medida perfecta del factor de simpatía, hay un término de

[5]aparte, dado que teníamos una idea bastante firme de nuestros factores "putativos" iniciales, podríamos haber ido directamente a CFA y omitir el paso de EFA. Ya sea que use EFA y luego pase a CFA, o vaya directamente a CFA, es una cuestión de juicio y qué tan seguro está inicialmente de que tiene el modelo correcto (en términos de número de factores y variables). Más temprano en el desarrollo de escalas, o en la identificación de construcciones latentes subyacentes, los investigadores tienden a usar EFA. Más tarde, a medida que se acercan a una escala final, o si quieren verificar una escala establecida en una nueva muestra, CFA es una buena opción.

error, e, asociado con él. En otras palabras, e representa la variación en A1 que no se explica por el factor de simpatía. Esto a veces se denomina **error de medición**.

El siguiente paso es considerar si se debe permitir que los factores latentes se correlacionen en nuestro modelo. Como se mencionó anteriormente, en las ciencias psicológicas y del comportamiento, los constructos a menudo están relacionados entre sí, y también pensamos que algunos de nuestros factores de personalidad pueden estar correlacionados entre sí. Entonces, en nuestro modelo, deberíamos permitir que estos factores latentes covaríen, como lo muestran las flechas de dos puntas en Figure 15.13.

Al mismo tiempo, debemos considerar si existe alguna buena razón sistemática para que algunos de los términos de error estén correlacionados entre sí. Una razón para esto podría ser que existe una característica metodológica compartida para subconjuntos particulares de las variables observadas, de modo que las variables observadas pueden estar correlacionadas por razones metodológicas en lugar de factores latentes sustantivos. Volveremos a esta posibilidad en una sección posterior pero, por ahora, no hay razones claras que podamos ver que justifiquen correlacionar algunos de los términos de error entre sí.

Sin ningún término de error correlacionado, el modelo que estamos probando para ver qué tan bien se ajusta a nuestros datos observados es tal como se especifica en Figure 15.13. Solo se espera encontrar en los datos los parámetros que están incluidos en el modelo, por lo que en CFA todos los demás parámetros posibles (coeficientes) se establecen en cero. Entonces, si estos otros parámetros no son cero (por ejemplo, puede haber una carga sustancial de A1 en el factor latente Extraversión en los datos observados, pero no en nuestro modelo), entonces podemos encontrar un ajuste deficiente entre nuestro modelo y los datos observados. .

Correcto, echemos un vistazo a cómo configuramos este análisis CFA en jamovi.

15.3.1 CFA en Jamovi

Abra el archivo bfi_sample2.csv, verifique que las 25 variables estén codificadas como ordinales (o continuas; no supondrá ninguna diferencia para este análisis). Para realizar CFA en jamovi:

- Seleccione Factor - Análisis factorial confirmatorio en la barra de botones principal de jamovi para abrir la ventana de análisis CFA (Figure 15.14).
- Seleccione las variables 5 A y transfiéralas al cuadro 'Factores' y asígnele la etiqueta "Agradabilidad".
- Cree un nuevo Factor en el cuadro 'Factores' y etiquételo como "Conciencia". Seleccione las 5 variables C y transfiéralas al cuadro 'Factores' debajo de la etiqueta "Consciencia".
- Cree otro Factor nuevo en el cuadro 'Factores' y etiquételo como "Extraversión". Seleccione las 5 variables E y transfiéralas al cuadro 'Factores' debajo de la etiqueta "Extraversión".
- Cree otro Factor nuevo en el cuadro 'Factores' y etiquételo como "Neuroticismo". Seleccione las 5 N variables y transfiéralas al cuadro 'Factores' debajo de la etiqueta "Neuroticismo".

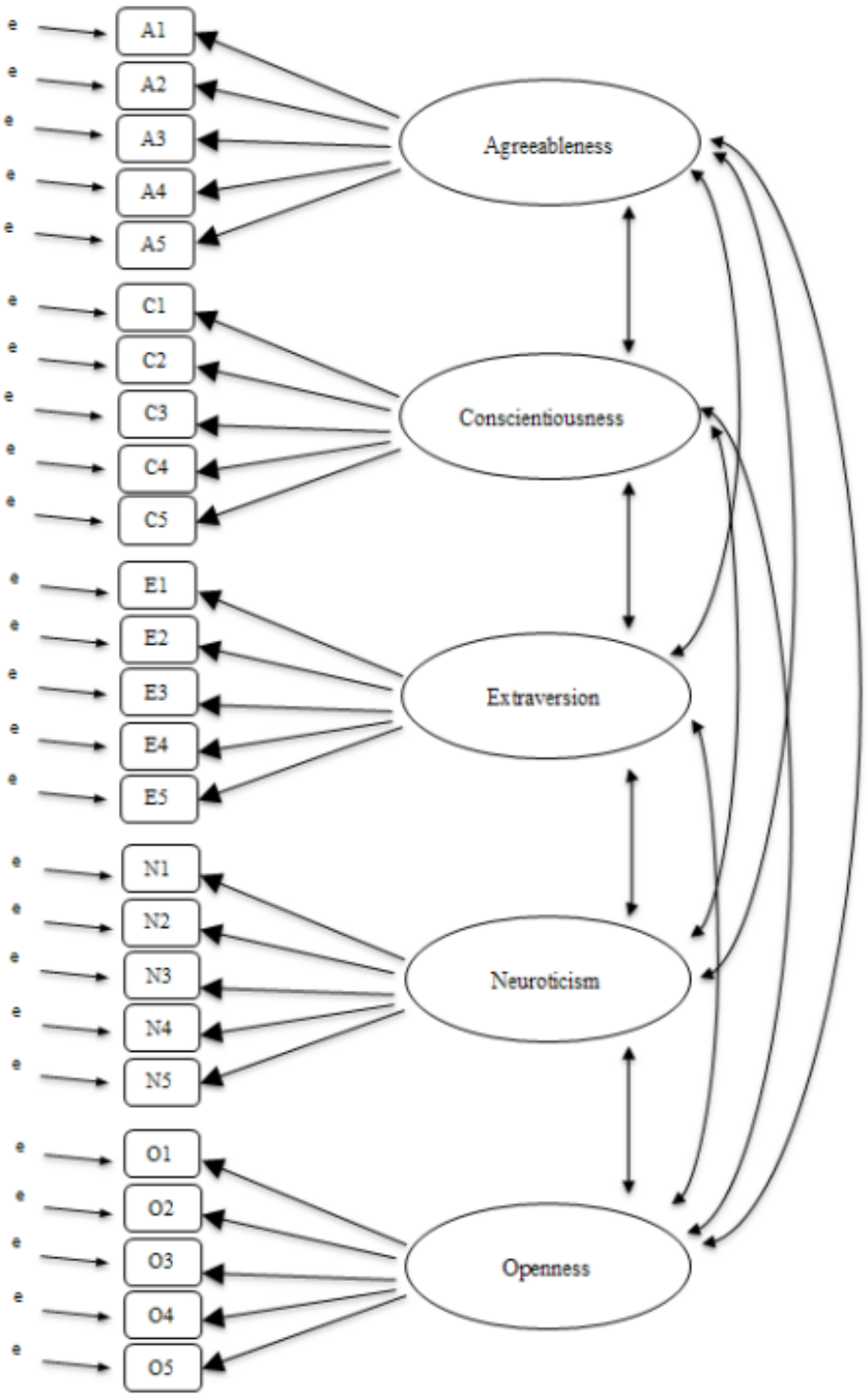


Figure 15.13: Especificación previa inicial de la estructura de factores latentes para las escalas de personalidad de cinco factores, para usar en CFA

- Cree otro Factor nuevo en el cuadro 'Factores' y etiquételo como "Apertura". Seleccione las 5 variables O y transfiéralas al cuadro 'Factores' debajo de la etiqueta "Apertura".
- Verifique otras opciones apropiadas, los valores predeterminados están bien para este trabajo inicial, aunque es posible que desee verificar la opción "Diagrama de ruta" en "Gráficos" para ver que jamovi produce un diagrama (bastante) similar a nuestro Figure 15.13 .

Una vez que hayamos configurado el análisis, podemos dirigir nuestra atención a la ventana de resultados de jamovi y ver qué es qué. Lo primero que hay que mirar es el **ajuste del modelo** (Figure 15.15), ya que nos dice qué tan bien se ajusta nuestro modelo a los datos observados. NB en nuestro modelo solo se estiman las covarianzas preespecificadas, incluidas las correlaciones de factores por defecto. Todo lo demás se pone a cero.

Hay varias formas de evaluar el ajuste del modelo. La primera es una estadística de chi-cuadrado que, si es pequeña, indica que el modelo se ajusta bien a los datos. Sin embargo, la estadística de chi-cuadrado utilizada para evaluar el ajuste del modelo es bastante sensible al tamaño de la muestra, lo que significa que con una muestra grande, un ajuste lo suficientemente bueno entre el modelo y los datos casi siempre produce un chi grande y significativo ($p < .05$). valor cuadrado.

Por lo tanto, necesitamos otras formas de evaluar el ajuste del modelo. En jamovi se proporcionan varios por defecto. Estos son el índice de ajuste comparativo (CFI), el índice de Tucker Lewis (TLI) y el error cuadrático medio de aproximación (RMSEA) junto con el intervalo de confianza del 90 % para el RMSEA. Algunas reglas generales útiles son que un ajuste satisfactorio está indicado por CFI $>$ 0,9, TLI $>$ 0,9 y RMSEA de aproximadamente 0,05 a 0,08. Un buen ajuste es CFI $>$ 0,95, TLI $>$ 0,95 y RMSEA y CI superior para RMSEA $<$ 0,05.

Entonces, mirando Figure 15.15 podemos ver que el valor de chi-cuadrado es grande y altamente significativo. Nuestro tamaño de muestra no es demasiado grande, por lo que esto posiblemente indica un mal ajuste. El CFI es de 0.762 y el TLI es de 0.731, lo que indica un mal ajuste entre el modelo y los datos. El RMSEA es de 0.085 con un intervalo de confianza de 90% de 0.077 a 0.092, de nuevo esto no indica un buen ajuste.

Bastante decepcionante, ¿eh? Pero tal vez no sea demasiado sorprendente dado que en el EFA anterior, cuando ejecutamos con un conjunto de datos similar (consulte la sección Análisis factorial exploratorio), solo alrededor de la mitad de la varianza en los datos fue explicada por el modelo de cinco factores.

Pasemos a ver las cargas factoriales y las estimaciones de la covarianza factorial, que se muestran en Figure 15.16 y Figure 15.17. La estadística Z y el valor p para cada uno de estos parámetros indican que hacen una contribución razonable al modelo (es decir, no son cero), por lo que no parece haber ninguna razón para eliminar ninguna de las rutas de factores variables especificadas. o correlaciones factor-factor del modelo. A menudo, las estimaciones estandarizadas son más fáciles de interpretar y se pueden especificar en la opción 'Estimaciones'. Estas tablas pueden incorporarse de manera útil en un informe escrito o artículo científico.

¿Cómo podríamos mejorar el modelo? Una opción es retroceder algunas etapas y volver a pensar en los elementos/medidas que estamos usando y cómo podrían mejorarse o cam-

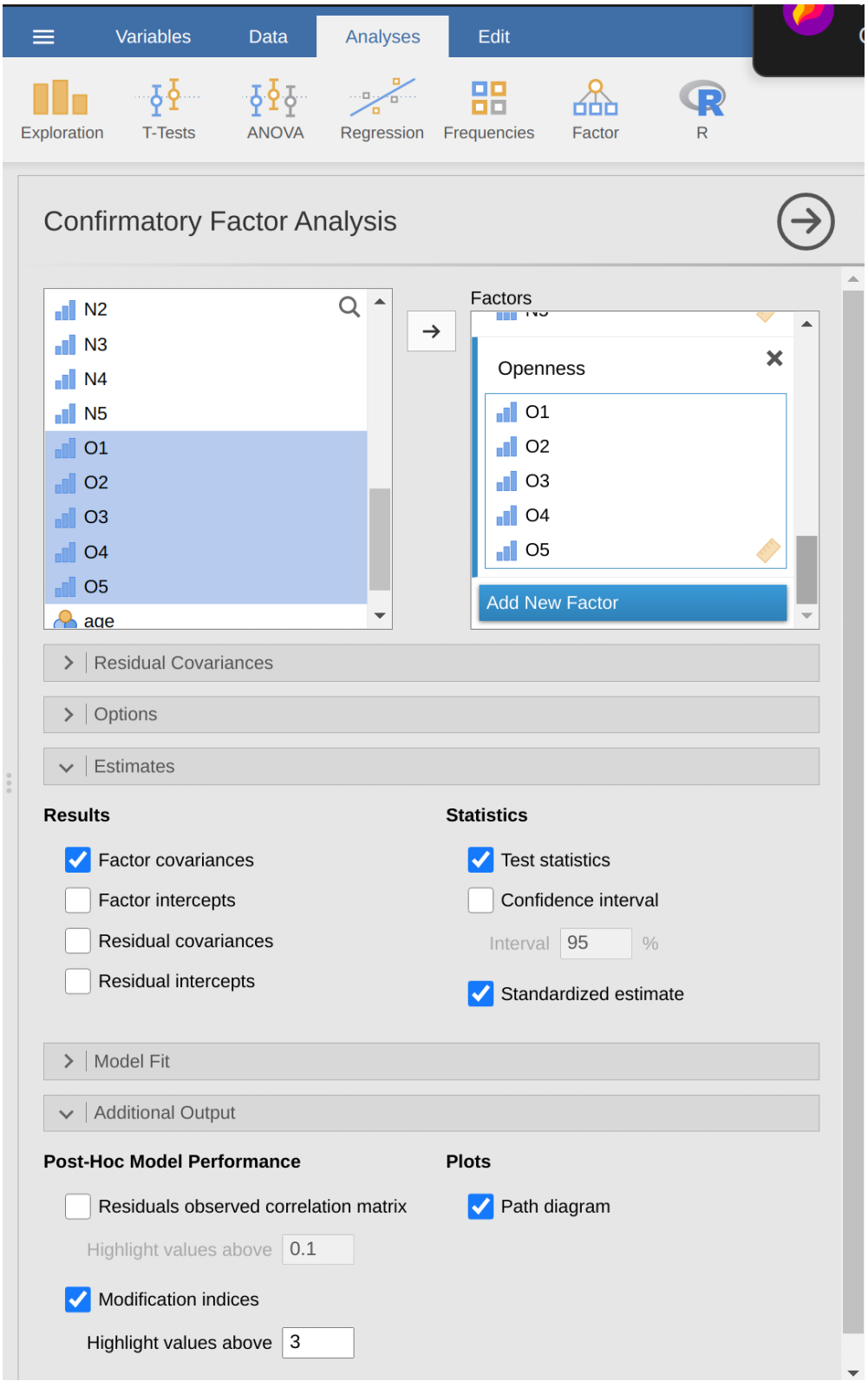


Figure 15.14: La ventana de análisis jamovi CFA

Model Fit

Test for Exact Fit

χ^2	df	p
739.73	265	< .00001

Fit Measures

			RMSEA 90% CI	
CFI	TLI	RMSEA	Lower	Upper
0.76	0.73	0.08	0.08	0.09

Figure 15.15: Los resultados de jamovi CFA Model Fit para nuestro modelo CFA

biarse. Otra opción es hacer algunos ajustes post hoc al modelo para mejorar el ajuste. Una forma de hacerlo es usar "índices de modificación" (Figure 15.18), especificados como una opción de "Salida adicional" en jamovi.

Lo que estamos buscando es el valor más alto del índice de modificación (MI). Luego juzgaríamos si tiene sentido agregar ese término adicional al modelo, usando una racionalización *post hoc*. Por ejemplo, podemos ver en Figure 15.18 que el MI más grande para las cargas factoriales que aún no están en el modelo es un valor de 28.786 para la carga de N4 ("A menudo se siente triste") en el factor latente Extraversión . Esto indica que si agregamos esta ruta al modelo, el valor de chi-cuadrado se reducirá aproximadamente en la misma cantidad.

Pero en nuestro modelo, podría decirse que agregar este camino realmente no tiene ningún sentido teórico o metodológico, por lo que no es una buena idea (a menos que pueda presentar un argumento persuasivo de que "A menudo me siento triste" mide tanto el neuroticismo como la extraversión). No puedo pensar en una buena razón. Pero, por el bien del argumento, supongamos que tiene algún sentido y agreguemos este camino al modelo. Vuelva a la ventana de análisis CFA (consulte Figure 15.14) y agregue N4 al factor de extraversión. Los resultados del CFA ahora cambiarán (no se muestra); el chi-cuadrado se ha reducido a alrededor de 709 (una caída de alrededor de 30, aproximadamente similar al tamaño del MI) y los otros índices de ajuste también han mejorado, aunque solo un poco. Pero no es suficiente: todavía no es un buen modelo de ajuste.

Confirmatory Factor Analysis

Factor Loadings

Factor	Indicator	Estimate	SE	Z	p	Stand. Estimate
Agreeableness	A1	0.47	0.10	4.73	< .00001	0.33
	A2	−0.77	0.08	−9.76	< .00001	−0.62
	A3	−1.06	0.08	−13.00	< .00001	−0.79
	A4	−0.75	0.10	−7.83	< .00001	−0.52
	A5	−0.94	0.08	−11.44	< .00001	−0.71
Conscientiousness	C1	0.80	0.08	10.53	< .00001	0.67
	C2	0.75	0.09	8.56	< .00001	0.57
	C3	0.54	0.09	6.19	< .00001	0.42
	C4	−1.07	0.09	−12.41	< .00001	−0.76
	C5	−1.12	0.10	−10.77	< .00001	−0.68
Extraversion	E1	0.90	0.11	8.51	< .00001	0.55
	E2	1.26	0.10	12.67	< .00001	0.76
	E3	−0.91	0.09	−9.89	< .00001	−0.63
	E4	−1.04	0.09	−11.28	< .00001	−0.69
	E5	−0.78	0.08	−9.37	< .00001	−0.60
Neuroticism	N1	1.34	0.09	14.17	< .00001	0.82
	N2	1.21	0.09	13.07	< .00001	0.76
	N3	1.22	0.10	12.49	< .00001	0.75
	N4	0.96	0.11	8.72	< .00001	0.57
	N5	0.84	0.11	7.63	< .00001	0.49
Openness	O1	0.67	0.08	8.12	< .00001	0.58
	O2	−0.65	0.12	−5.66	< .00001	−0.40
	O3	1.00	0.09	11.30	< .00001	0.83
	O4	0.23	0.09	2.73	0.00624	0.20
	O5	−0.51	0.09	−5.52	< .00001	−0.41

Figure 15.16: La tabla jamovi CFA Factor Loadings para nuestro modelo CFA

Factor Estimates

Factor Covariances

		Estimate	SE	Z	p	Stand. Estimate
Agreeableness	Agreeableness	1.00 [a]				
	Conscientiousness	−0.34	0.07	−4.51	< .00001	−0.34
	Extraversion	0.58	0.06	9.21	< .00001	0.58
	Neuroticism	0.16	0.08	2.14	0.03233	0.16
	Openness	−0.42	0.07	−5.95	< .00001	−0.42
Conscientiousness	Conscientiousness	1.00 [a]				
	Extraversion	−0.50	0.07	−7.76	< .00001	−0.50
	Neuroticism	−0.29	0.07	−3.92	0.00009	−0.29
	Openness	0.28	0.08	3.44	0.00058	0.28
Extraversion	Extraversion	1.00 [a]				
	Neuroticism	0.24	0.08	3.12	0.00179	0.24
	Openness	−0.53	0.07	−7.95	< .00001	−0.53
Neuroticism	Neuroticism	1.00 [a]				
	Openness	−0.18	0.08	−2.32	0.02033	−0.18
Openness	Openness	1.00 [a]				

[a] fixed parameter

Figure 15.17: La tabla de covarianzas del factor CFA jamovi para nuestro modelo CFA

Si se encuentra agregando nuevos parámetros a un modelo utilizando los valores de MI, siempre vuelva a verificar las tablas de MI después de cada nueva adición, ya que los MI se actualizan cada vez.

También hay una tabla de índices de modificación de covarianza residual producida por jamovi (Figure 15.19). En otras palabras, una tabla que muestre qué errores correlacionados, si se agregan al modelo, mejorarían más el ajuste del modelo. Es una buena idea mirar ambas tablas de MI al mismo tiempo, detectar el MI más grande, pensar si la adición del parámetro sugerido puede justificarse razonablemente y, si es posible, agregarlo al modelo. Y luego puede comenzar nuevamente a buscar el MI más grande en los resultados recalculados.

Puede seguir así todo el tiempo que desee, agregando parámetros al modelo en función del MI más grande y, finalmente, logrará un ajuste satisfactorio. ¡Pero también habrá una gran posibilidad de que al hacer esto hayas creado un monstruo! Un modelo feo y deforme que no tiene ningún sentido teórico ni pureza. En otras palabras, ¡ten mucho cuidado!

Hasta ahora hemos comprobado la estructura factorial obtenida en el AFE utilizando una segunda muestra y AFC. Desafortunadamente, no encontramos que la estructura factorial del EFA se confirmara en el CFA, por lo que está de vuelta en el tablero de dibujo en lo que respecta al desarrollo de esta escala de personalidad.

Aunque podríamos haber ajustado el CFA usando índices de modificación, realmente no había ninguna buena razón (que se me ocurriera) para incluir estas cargas factoriales adicionales sugeridas o covarianzas residuales. Sin embargo, a veces hay una buena

Modification Indices

Factor Loadings – Modification Indices

	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
A1		11.14	14.21	1.53	0.39
A2		1.20	1.02	3.65	0.53
A3		1.76	4.70	2.95	0.00
A4		6.99	4.15	0.06	2.30
A5		3.84	11.41	8.18	3.50
C1	4.22		1.96	0.55	0.15
C2	1.40		0.01	12.08	0.01
C3	0.85		1.54	12.77	0.86
C4	0.96		1.05	2.71	0.21
C5	0.23		1.03	4.75	0.21
E1	13.55	0.50		1.36	3.14
E2	5.36	1.67		19.01	2.33
E3	4.40	5.85		6.20	28.02
E4	22.83	1.51		0.30	9.37
E5	2.40	8.78		2.51	2.89
N1	1.15	0.59	1.08		0.15
N2	1.03	9.83	7.58		3.18
N3	0.21	0.01	0.01		0.01
N4	1.65	14.03	28.79		0.65
N5	1.13	0.03	0.84		2.90
O1	0.07	0.19	1.89	0.28	
O2	6.43	2.43	8.06	6.85	
O3	2.71	2.09	7.70	1.04	
O4	1.85	13.39	10.54	8.87	
O5	1.58	4.49	2.97	2.19	

Figure 15.18: Los índices de modificación de cargas factoriales CFA jamovi

Residual Covariances – Modification Indices

	A1	A2	A3	A4	A5	C1	C2	C3	C4	C5	E1	E2	E3	E4	E5	N1	N2	N3	N4	N5	O1	O2	O3	O4	O5
A1		7.93	4.45	0.55	3.05	1.33	1.99	1.01	3.24	5.87	0.04	0.10	7.92	0.06	6.91	1.28	11.77	1.24	1.63	2.13	6.01	0.85	0.80	8.11	1.05
A2			0.08	0.96	1.84	1.15	3.10	3.42	3.74	2.67	0.01	3.32	0.84	2.51	1.66	1.17	0.95	0.20	7.18	8.30	0.19	6.27	0.21	12.01	2.16
A3				0.01	0.12	0.41	0.64	0.09	3.26	1.08	5.67	7.55	4.53	3.07	0.82	0.27	0.58	1.07	0.65	0.98	0.03	0.50	0.42	1.25	0.14
A4					2.19	0.93	6.77	0.40	2.34	12.73	0.35	0.32	0.10	2.19	0.28	3.57	0.09	1.58	2.04	1.18	0.09	2.74	0.28	2.70	1.92
A5						1.36	0.00	0.47	0.16	1.78	0.25	3.90	0.43	1.66	2.28	1.93	0.15	0.39	0.19	0.03	0.04	2.98	0.25	0.06	0.17
C1							13.88	0.19	0.34	10.69	0.09	2.14	3.04	0.30	2.29	0.99	0.23	4.31	3.74	1.63	0.17	0.43	0.02	0.06	8.83
C2								6.14	9.74	0.60	2.04	3.18	0.14	0.56	5.96	0.15	1.89	0.04	2.60	6.32	0.25	9.81	0.69	4.28	0.09
C3									2.06	0.85	0.01	0.01	0.32	2.51	2.38	0.92	0.32	0.14	0.18	6.52	0.00	0.54	0.32	4.08	3.64
C4										3.62	0.12	3.05	0.25	1.07	0.07	7.97	1.73	6.53	1.52	4.20	0.01	11.85	5.32	5.08	6.37
C5											0.00	1.41	0.45	0.38	0.00	6.46	0.81	1.08	10.44	2.17	0.00	1.30	0.34	6.64	1.42
E1												10.71	0.97	0.04	0.36	2.78	1.63	1.40	4.29	0.06	2.56	0.50	0.03	0.93	0.04
E2													2.10	1.19	2.35	0.03	2.22	2.67	15.40	7.01	0.13	0.42	0.02	11.44	1.41
E3														0.93	1.23	1.50	0.05	1.21	1.55	2.72	3.34	2.67	16.90	2.86	0.60
E4															1.26	0.00	0.87	1.44	2.44	6.86	0.12	10.04	1.12	0.04	3.26
E5																1.36	6.54	1.54	1.10	0.00	5.97	0.90	0.42	0.33	0.23
N1																	45.00	4.12	25.70	0.25	1.42	0.09	0.21	0.75	0.11
N2																		4.37	3.48	6.25	3.23	2.53	0.82	0.04	4.68
N3																			19.26	0.91	1.19	0.06	0.08	0.02	6.97
N4																				8.57	0.01	0.25	0.74	10.89	0.65
N5																					2.31	3.26	0.13	1.92	2.61
O1																						0.04	4.69	0.33	0.95
O2																							0.06	3.04	12.95
O3																								2.93	3.24
O4																									0.17
O5																									

Figure 15.19: Índices de modificación de covarianza residual producidos por jamovi

razón para permitir que los residuos covaríen (o se correlacionen), y un buen ejemplo de esto se muestra en la siguiente sección sobre [CFA de múltiples características y múltiples métodos]. Antes de hacer eso, veamos cómo informar los resultados de un CFA.

15.3.2 Reportar un CFA

No existe una forma estándar formal de redactar un CFA, y los ejemplos tienden a variar según la disciplina y el investigador. Dicho esto, hay algunas piezas de información bastante estándar para incluir en su redacción:

1. Una justificación teórica y empírica del modelo hipotético.
2. Una descripción completa de cómo se especificó el modelo (p. ej., las variables indicadoras para cada factor latente, las covarianzas entre las variables latentes y cualquier correlación entre los términos de error). Sería bueno incluir un diagrama de ruta, como el de Figure 15.13.
3. Una descripción de la muestra (por ejemplo, información demográfica, tamaño de la muestra, método de muestreo).
4. Una descripción del tipo de datos utilizados (p. ej., nominales, continuos) y estadísticas descriptivas.
5. Pruebas de supuestos y método de estimación utilizado.
6. Una descripción de los datos faltantes y cómo se manejaron los datos faltantes.
7. El software y la versión utilizados para adaptarse al modelo.
8. Medidas y criterios utilizados para juzgar el ajuste del modelo.
9. Cualquier alteración realizada en el modelo original en función de los índices de ajuste o modificación del modelo.
10. Todas las estimaciones de parámetros (es decir, cargas, varianzas de error, (co)varianzas latentes) y sus errores estándar, probablemente en una tabla.

15.4 Múltiples Rasgos Múltiples Métodos CFA

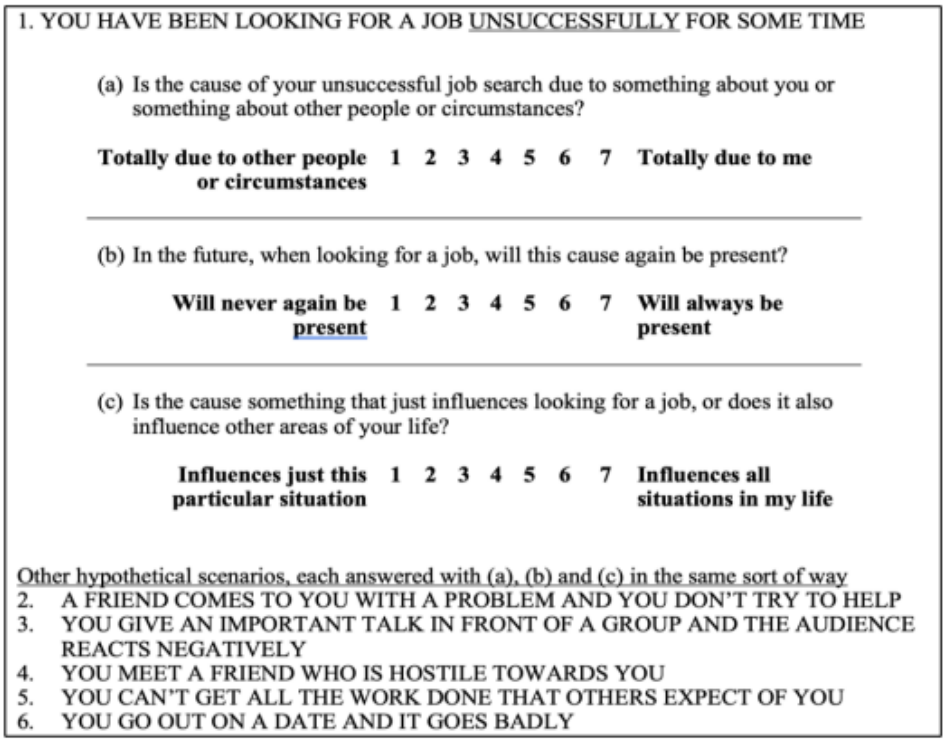
En esta sección, consideraremos cómo las diferentes técnicas o preguntas de medición pueden ser una fuente importante de variabilidad de datos, conocida como **varianza del método**. Para hacer esto, usaremos otro conjunto de datos psicológicos, uno que contiene datos sobre el "estilo atribucional".

Se utilizó el Cuestionario de Estilo Atribucional (ASQ) (Hewitt et al., 2004) para recopilar datos de bienestar psicológico de jóvenes en el Reino Unido y Nueva Zelanda. Midieron el estilo atribucional de los eventos negativos, que es la forma en que las personas suelen explicar la causa de las cosas malas que les suceden (Peterson & Seligman, 1984). El cuestionario de estilo atribucional (ASQ) mide tres aspectos del estilo atribucional:

- La internalidad es el grado en que una persona cree que la causa de un mal evento se debe a sus propias acciones.

- La estabilidad se refiere a la medida en que una persona cree habitualmente que la causa de un mal evento es estable a lo largo del tiempo.
- La globalidad se refiere al grado en que una persona cree habitualmente que la causa de un mal evento en un área afectará otras áreas de su vida.

Hay seis escenarios hipotéticos y para cada escenario los encuestados responden una pregunta dirigida a (a) la interioridad, (b) la estabilidad y (c) la globalidad. Así que hay $6 \times 3 = 18$ artículos en total. Consulte Figure 15.20 para obtener más detalles.



1. YOU HAVE BEEN LOOKING FOR A JOB UNSUCCESSFULLY FOR SOME TIME

(a) Is the cause of your unsuccessful job search due to something about you or something about other people or circumstances?

Totally due to other people or circumstances 1 2 3 4 5 6 7 **Totally due to me**

(b) In the future, when looking for a job, will this cause again be present?

Will never again be present 1 2 3 4 5 6 7 **Will always be present**

(c) Is the cause something that just influences looking for a job, or does it also influence other areas of your life?

Influences just this particular situation 1 2 3 4 5 6 7 **Influences all situations in my life**

Other hypothetical scenarios, each answered with (a), (b) and (c) in the same sort of way

2. A FRIEND COMES TO YOU WITH A PROBLEM AND YOU DON'T TRY TO HELP
3. YOU GIVE AN IMPORTANT TALK IN FRONT OF A GROUP AND THE AUDIENCE REACTS NEGATIVELY
4. YOU MEET A FRIEND WHO IS HOSTILE TOWARDS YOU
5. YOU CAN'T GET ALL THE WORK DONE THAT OTHERS EXPECT OF YOU
6. YOU GO OUT ON A DATE AND IT GOES BADLY

Figure 15.20: El Cuestionario de Estilo Atribucional (ASQ) para eventos negativos

Los investigadores están interesados en verificar sus datos para ver si hay algunos factores latentes subyacentes que las 18 variables observadas en el ASQ miden razonablemente bien.

Primero, prueban EFA con estas 18 variables (no se muestran), pero no importa cómo extraigan o roten, no pueden encontrar una buena solución factorial. Su intento de identificar los factores latentes subyacentes en el Cuestionario de Estilo Atribucional (ASQ) resultó infructuoso. Si obtiene resultados como este, entonces su teoría es incorrecta (no hay una estructura de factores latentes subyacente para el estilo atribucional, lo cual es posible), la muestra no es relevante (lo cual es poco probable dado el tamaño y las características de esta muestra de adultos jóvenes de el Reino Unido y Nueva Zelanda), o el análisis no era la herramienta adecuada para el trabajo. Vamos a ver esta tercera posibilidad.

Recuerde que había tres dimensiones medidas en el ASQ: Internalidad, Estabilidad y Globalidad, cada una medida por seis preguntas como se muestra en Figure 15.21.

¿Qué pasa si, en lugar de hacer un análisis en el que vemos cómo los datos van juntos en un sentido exploratorio, imponemos una estructura, como en Figure 15.21, sobre los datos y vemos qué tan bien se ajustan a nuestros datos preespecificados? estructura. En este sentido, estamos realizando un análisis confirmatorio, para ver qué tan bien los datos observados confirman un modelo preespecificado.

Por lo tanto, un análisis factorial confirmatorio (CFA) directo del ASQ especificaría tres factores latentes, como se muestra en las columnas de Figure 15.27, cada uno medido por seis variables observadas.

Internality	Stability	Globality
Q1a	Q1b	Q1c
Q2a	Q2b	Q2c
Q3a	Q3b	Q3c
Q4a	Q4b	Q4c
Q5a	Q5b	Q5c
Q6a	Q6b	Q6c

Figure 15.21: Seis preguntas sobre el ASQ para cada una de las dimensiones de Internalidad, Estabilidad y Globalidad

Podríamos representar esto como en el diagrama en Figure 15.22, que muestra que cada variable es una medida de un factor latente subyacente. Por ejemplo, INT1 se predice mediante el factor latente subyacente Internalidad. Y debido a que INT1 no es una medida perfecta del factor de internalidad, hay un término de error, e1, asociado con él. En otras palabras, e1 representa la varianza en INT1 que no se explica por el factor de Internalidad. Esto a veces se denomina "error de medición".

El siguiente paso es considerar si se debe permitir que los factores latentes se correlacionen en nuestro modelo. Como se mencionó anteriormente, en las ciencias psicológicas y del comportamiento, los constructos a menudo están relacionados entre sí, y también pensamos que la Internalidad, la Estabilidad y la Globalidad pueden estar correlacionadas entre sí, por lo que en nuestro modelo deberíamos permitir que estos factores latentes covaríen. , como se muestra en Figure 15.23.

Al mismo tiempo, debemos considerar si existe alguna buena razón sistemática para que algunos de los términos de error estén correlacionados entre sí. Volviendo a las preguntas del ASQ, había tres subpreguntas diferentes (a, byc) para cada pregunta principal (1-6). La P1 se refería a la búsqueda de empleo sin éxito y es plausible que esta pregunta tenga algunos aspectos metodológicos o artefactos distintivos además de las otras preguntas (2-5), quizás algo relacionado con la búsqueda de empleo. De manera similar, P2 se

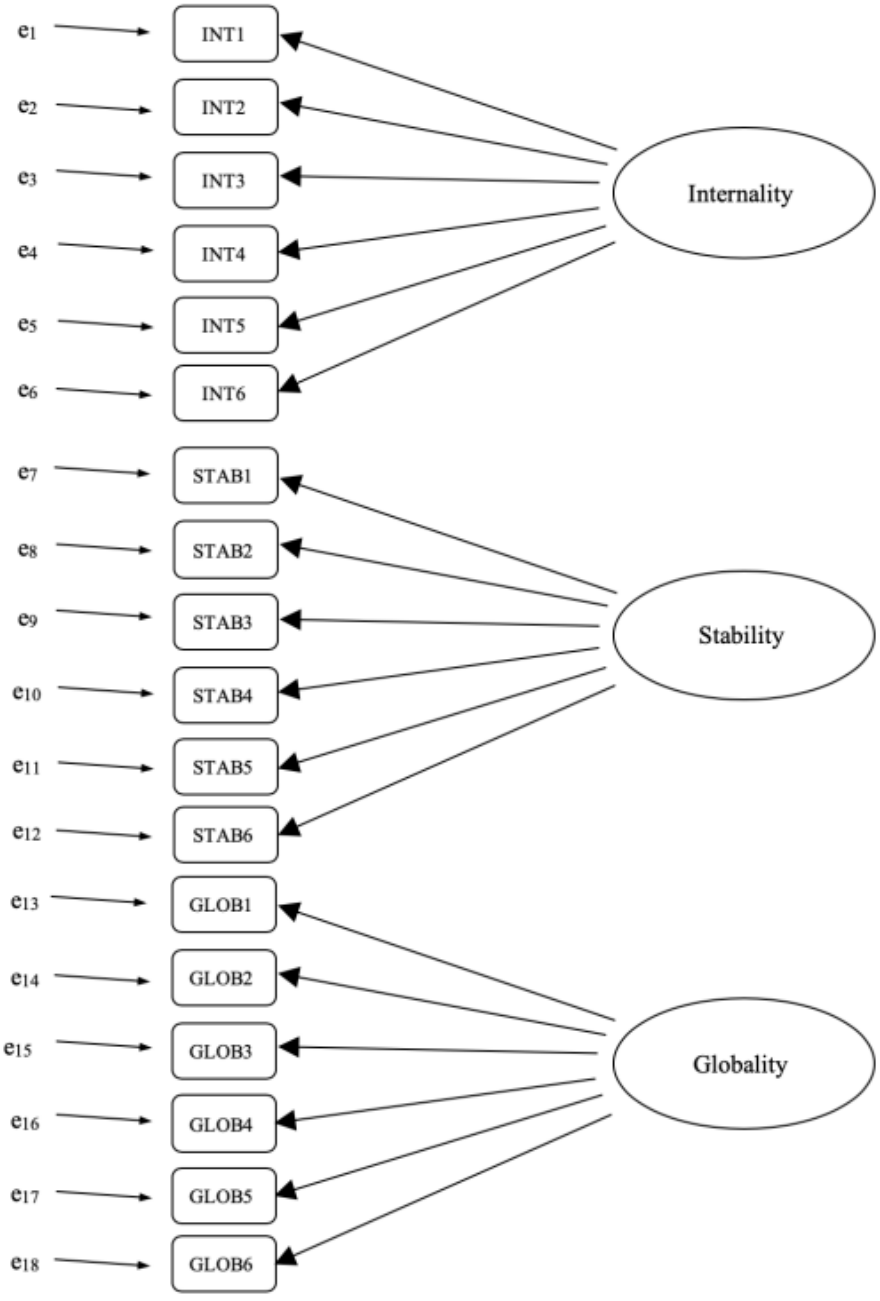


Figure 15.22: especificación previa inicial de la estructura del factor latente para el ASQ

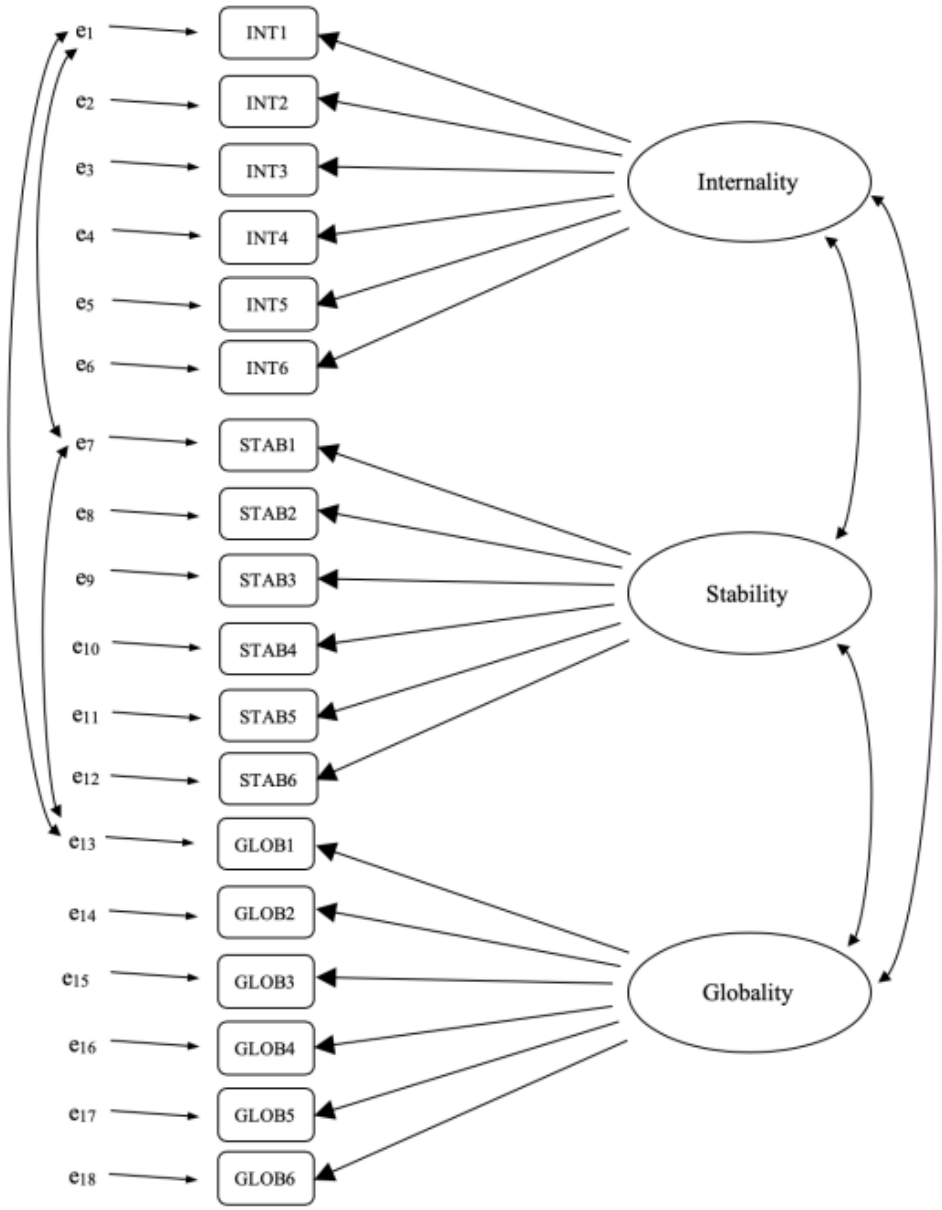


Figure 15.23: Especificación previa final de la estructura de factores latentes para el ASQ, incluidas las correlaciones de factores latentes y las correlaciones de términos de error de método compartido para la variable observada INT1, STAB1 y GLOB1, en un modelo CFA MTMM. Para mayor claridad, no se muestran otras correlaciones de términos de error preespecificados.

trataba de no ayudar a un amigo con un problema, y puede haber algunos aspectos metodológicos o de artefacto distintivos relacionados con no ayudar a un amigo que no están presentes en las otras preguntas (1 y 3-5).

Entonces, además de múltiples factores, también tenemos múltiples características metodológicas en el ASQ, donde cada una de las preguntas 1 a 6 tiene un "método" ligeramente diferente, pero cada "método" se comparte en las subpreguntas a, b y c. . Para incorporar estas diferentes características metodológicas en el modelo, podemos especificar que ciertos términos de error estén correlacionados entre sí. Por ejemplo, los errores asociados con INT1, STAB1 y GLOB1 deben correlacionarse entre sí para reflejar la varianza metodológica distinta y compartida de Q1a, Q1b y Q1c. Mirando Figure 15.21, esto significa que además de los factores latentes representados por las columnas, tendremos errores de medición correlacionados para las variables en cada fila de la tabla.

Si bien un modelo CFA básico como el que se muestra en Figure 15.22 podría probarse con nuestros datos observados, de hecho hemos creado un modelo más sofisticado, como se muestra en el diagrama en Figure 15.23. Este modelo CFA más sofisticado se conoce como modelo **Multi-Trait Multi-Method (MTMM)**, y es el que probaremos en jamovi.

15.4.1 MTMM CFA en Jamovi

Abra el archivo ASQ.csv y compruebe que las 18 variables (seis variables de "Internalidad", seis de "Estabilidad" y seis de "Globalidad") se especifican como variables continuas.

Para realizar MTMM CFA en jamovi:

- Seleccione Factor - Análisis factorial confirmatorio en la barra de botones principal de jamovi para abrir la ventana de análisis CFA (Figure 15.24).
- Seleccione las 6 variables INT y transfiéralas al cuadro 'Factores' y asígneles la etiqueta "Internalidad".
- Cree un nuevo Factor en el cuadro 'Factores' y etiquételo como "Estabilidad". Seleccione las 6 variables STAB y transfiéralas al cuadro 'Factores' debajo de la etiqueta "Estabilidad".
- Cree otro Factor nuevo en el cuadro 'Factores' y etiquételo como "Globalidad". Seleccione las 6 variables GLOB y transfiéralas al cuadro 'Factores' debajo de la etiqueta "Globalidad".
- Abra las opciones de Covarianzas residuales y, para cada una de nuestras correlaciones preespecificadas, mueva las variables asociadas al cuadro "Covarianzas residuales" de la derecha. Por ejemplo, resalte INT1 y STAB1 y luego haga clic en la flecha para moverlos. Ahora haga lo mismo para INT1 y GLOB1, para STAB1 y GLOB1, para INT2 y STAB2, para INT2 y GLOB2, para STAB2 y GLOB2, para INT3 y STAB3, y así sucesivamente.
- Verifique otras opciones apropiadas, los valores predeterminados están bien para este trabajo inicial, aunque es posible que desee verificar la opción "Diagrama de ruta" en "Gráficos" para ver que jamovi produce un diagrama (bastante) similar

a nuestro Figure 15.23 , e incluyendo todas las correlaciones de términos de error que hemos agregado anteriormente.

Una vez que hayamos configurado el análisis, podemos dirigir nuestra atención a la ventana de resultados de jamovi y ver qué es qué. Lo primero que debe observar es el "Ajuste del modelo", ya que esto nos dice qué tan bien se ajusta nuestro modelo a los datos observados (Figure 15.25). NB en nuestro modelo solo se estiman las covarianzas preespecificadas, todo lo demás se establece en cero, por lo que el ajuste del modelo prueba si los parámetros "libres" preespecificados no son cero y, por el contrario, si las otras relaciones en los datos – los que no hemos especificado en el modelo, se pueden mantener en cero.

Mirando Figure 15.25 podemos ver que el valor de chi-cuadrado es altamente significativo, lo cual no es una sorpresa dado el gran tamaño de la muestra (N = 2748). El CFI es 0,98 y el TLI también es 0,98, lo que indica un muy buen ajuste. El RMSEA es 0,02 con un intervalo de confianza del 90 % de 0,02 a 0,02, ¡muy ajustado!

En general, creo que podemos estar satisfechos de que nuestro modelo preespecificado se ajusta muy bien a los datos observados, lo que respalda nuestro modelo MTMM para el ASQ.

Ahora podemos pasar a ver las cargas factoriales y las estimaciones de la covarianza factorial, como en Figure 15.26. A menudo, las estimaciones estandarizadas son más fáciles de interpretar y se pueden especificar en la opción 'Estimaciones'. Estas tablas pueden incorporarse de manera útil en un informe escrito o artículo científico.

Puede ver en Figure 15.26 que todas nuestras cargas factoriales y covarianzas factoriales preespecificadas son significativamente diferentes de cero. En otras palabras, todos parecen estar haciendo una contribución útil al modelo.

Hemos tenido bastante suerte con este análisis, ¡obteniendo un muy buen ajuste en nuestro primer intento!

15.5 Análisis de confiabilidad de consistencia interna

Después de haber pasado por el proceso de desarrollo de escala inicial utilizando EFA y CFA, debería haber llegado a una etapa en la que la escala se sostiene bastante bien usando CFA con diferentes muestras. Una cosa que también le puede interesar en esta etapa es ver qué tan bien se miden los factores usando una escala que combina las variables observadas.

En psicometría, utilizamos el análisis de confiabilidad para proporcionar información sobre la consistencia con la que una escala mide una construcción psicológica (consulte la sección anterior sobre Section 2.3). **La consistencia interna** es lo que nos preocupa aquí, y se refiere a la consistencia entre todos los elementos individuales que componen una escala de medición. Entonces, si tenemos $V1, V2, V3, V4$ y $V5$ como variables de elementos observados, entonces podemos calcular una estadística que nos diga qué tan consistentes internamente son estos elementos para medir la construcción subyacente.

Una estadística popular utilizada para verificar la consistencia interna de una escala es el **alfa de Cronbach** (Chronbach, 1951). El alfa de Cronbach es una medida de equivalencia (si diferentes conjuntos de elementos de la escala darían los mismos resultados de medición). La equivalencia se prueba dividiendo los ítems de la escala en dos

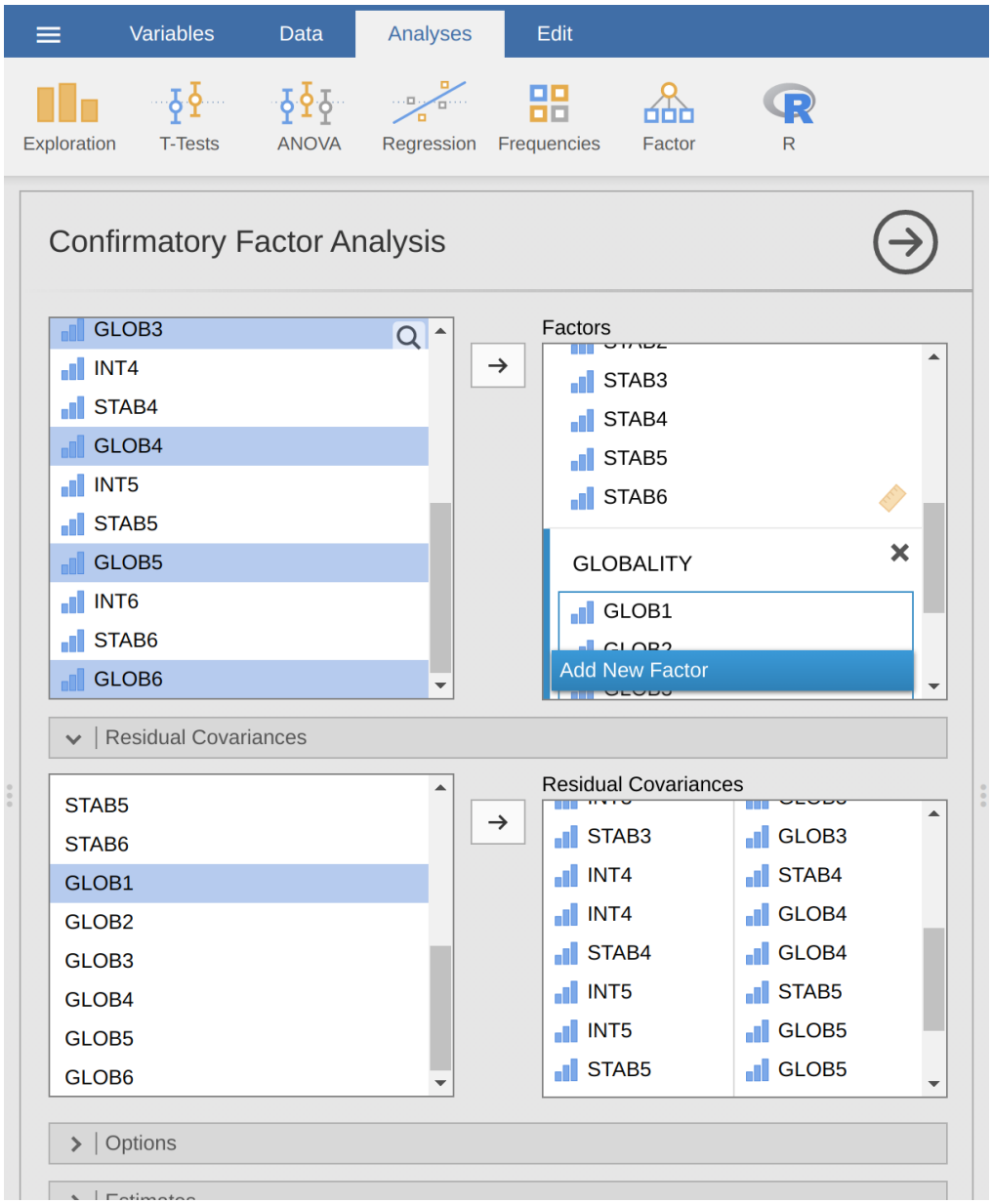
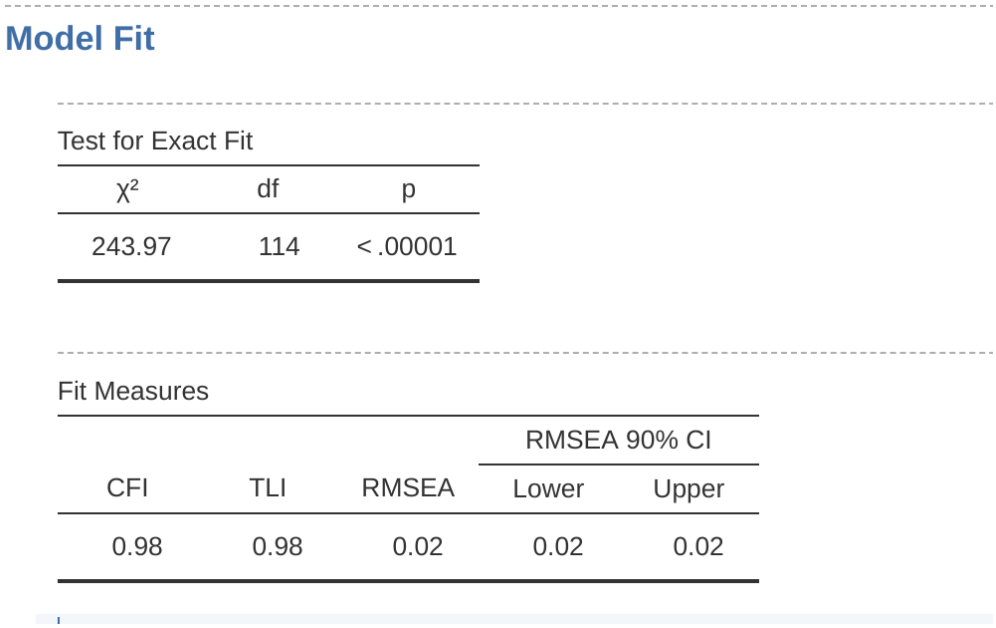


Figure 15.24: La ventana de análisis jamovi CFA



Model Fit

Test for Exact Fit

χ^2	df	p
243.97	114	< .00001

Fit Measures

			RMSEA 90% CI	
CFI	TLI	RMSEA	Lower	Upper
0.98	0.98	0.02	0.02	0.02

Figure 15.25: Los resultados de jamovi CFA Model Fit para nuestro modelo CFA MTMM

grupos (una "mitad dividida") y viendo si el análisis de las dos partes da resultados comparables. Por supuesto, hay muchas formas de dividir un conjunto de elementos, pero si se realizan todas las divisiones posibles, es posible generar una estadística que refleje el patrón general de los coeficientes de división por mitades. El alfa de Cronbach (α) es una estadística de este tipo: una función de todos los coeficientes divididos por la mitad de una escala. Si un conjunto de elementos que miden un constructo (por ejemplo, una escala de Extraversión) tiene un α de $0,80$, entonces la proporción de la varianza del error en la escala es de $0,20$. En otras palabras, una escala con α de 0.80 incluye aproximadamente un 20% de error.

PERO, (y ese es un GRAN "PERO"), el alfa de Cronbach no es una medida de unidimensionalidad (es decir, un indicador de que una escala mide un solo factor o construcción en lugar de múltiples construcciones relacionadas). Las escalas que son multidimensionales harán que se subestime alfa si no se evalúan por separado para cada dimensión, pero los valores altos de alfa no son necesariamente indicadores de unidimensionalidad. Por lo tanto, un α de 0,80 no significa que se tenga en cuenta el 80 % de una única construcción subyacente. Podría ser que el 80% provenga de más de una construcción subyacente. Es por eso que EFA y CFA son útiles para hacer primero.

Además, otra característica de α es que tiende a ser específico de la muestra: no es una característica de la escala, sino una característica de la muestra en la que se ha utilizado la escala. Una muestra sesgada, no representativa o pequeña podría producir un coeficiente α muy diferente al de una muestra grande y representativa. α incluso puede variar de una muestra grande a una muestra grande. Sin embargo, a pesar de estas limitaciones, el α de Cronbach ha sido popular en Psicología para estimar la

Confirmatory Factor Analysis

Factor Loadings

Factor	Indicator	Estimate	SE	Z	p	Stand. Estimate
INTERNALITY	INT1	0.55	0.05	12.28	< .00001	0.34
	INT2	0.50	0.05	10.52	< .00001	0.28
	INT3	0.61	0.04	13.95	< .00001	0.38
	INT4	0.64	0.05	13.45	< .00001	0.36
	INT5	0.54	0.04	12.52	< .00001	0.33
	INT6	0.66	0.04	16.50	< .00001	0.45
STABILITY	STAB1	0.53	0.04	14.97	< .00001	0.35
	STAB2	0.48	0.03	14.54	< .00001	0.34
	STAB3	0.69	0.03	20.20	< .00001	0.46
	STAB4	0.65	0.03	20.72	< .00001	0.47
	STAB5	0.67	0.03	21.47	< .00001	0.49
	STAB6	0.66	0.03	22.17	< .00001	0.51
GLOBALITY	GLOB1	0.71	0.04	18.16	< .00001	0.40
	GLOB2	0.73	0.04	20.32	< .00001	0.44
	GLOB3	0.93	0.04	25.10	< .00001	0.54
	GLOB4	0.83	0.03	24.99	< .00001	0.53
	GLOB5	0.76	0.03	22.71	< .00001	0.48
	GLOB6	0.96	0.03	27.66	< .00001	0.59

[3]

Factor Estimates

Factor Covariances

		Estimate	SE	Z	p	Stand. Estimate
INTERNALITY	INTERNALITY	1.00 [a]				
	STABILITY	0.52	0.03	17.10	< .00001	0.52
	GLOBALITY	0.45	0.03	14.96	< .00001	0.45
STABILITY	STABILITY	1.00 [a]				
	GLOBALITY	0.70	0.02	35.47	< .00001	0.70
GLOBALITY	GLOBALITY	1.00 [a]				

[a] fixed parameter

Figure 15.26: Las tablas jamovi CFA Factor Loadings and Covariances para nuestro modelo CFA MTMM

confiabilidad de la consistencia interna. Es bastante fácil de calcular, comprender e interpretar y, por lo tanto, puede ser una verificación inicial útil del rendimiento de la báscula cuando administra una báscula con una muestra diferente, de un entorno o población diferente, por ejemplo.

Una alternativa es **Omega de McDonald's** (ω), y jamovi también proporciona esta estadística. Mientras que α hace las siguientes suposiciones: (a) no hay correlaciones residuales, (b) los elementos tienen cargas idénticas y (c) la escala es unidimensional, ω no lo hace y, por lo tanto, es una estadística de confiabilidad más robusta. Si no se violan estas suposiciones, α y ω serán similares, pero si lo son, entonces se preferirá ω.

A veces se proporciona un umbral para α o ω, lo que sugiere un valor "suficientemente bueno". Esto podría ser algo así como αs de 0.70 o 0.80 que representan confiabilidad "aceptable" y "buena", respectivamente. Sin embargo, esto depende de lo que se supone que mida exactamente la báscula, por lo que los umbrales como este deben usarse con precaución. Podría ser mejor decir simplemente que un α o ω de $0,70$ está asociado con una varianza de error del 30 % en una escala, y un α o ω de $0,80$ está asociado con 20 %

¿Puede α ser demasiado alto? Probablemente: si obtiene un coeficiente α por encima de $0,95$, esto indica una alta intercorrelación entre los elementos y que podría haber demasiada especificidad demasiado redundante en la medición, con el riesgo de que el constructo que se mide sea quizás demasiado angosto.

15.5.1 Análisis de confiabilidad en jamovi

Tenemos una tercera muestra de datos de personalidad para usar para realizar análisis de confiabilidad: en el archivo bfi_sample3.csv. Una vez más, verifique que las 25 variables del ítem de personalidad estén codificadas como continuas. Para realizar un análisis de confiabilidad en jamovi:

- Seleccione Factor - Análisis de confiabilidad en la barra de botones principal de jamovi para abrir la ventana de análisis de confiabilidad (Figure 15.27).
- Seleccione las variables 5 A y transfiéralas al cuadro 'Artículos'.
- En la opción "Elementos de escala inversa", seleccione la variable A1 en el cuadro "Elementos de escala normal" y muévalo al cuadro "Elementos de escala inversa".
- Verifique otras opciones apropiadas, como en Figure 15.27.

Una vez hecho esto, mira la ventana de resultados de jamovi. Deberías ver algo como Figure 15.28. Esto nos dice que el coeficiente α de Cronbach para la escala de Amabilidad es 0,72. Esto significa que poco menos del 30 % de la puntuación de la escala de Amabilidad es una varianza del error. También se da ω de McDonald's, y esto es 0.74, no muy diferente de α.

También podemos comprobar cómo se puede mejorar α o ω si se elimina un elemento específico de la escala. Por ejemplo, α aumentaría a 0,74 y ω a 0,75 si eliminásemos el elemento A1. Este no es un gran aumento, por lo que probablemente no valga la pena hacerlo.

El proceso de cálculo y verificación de las estadísticas de la báscula (α y ω) es el mismo para todas las demás básculas, y todas tenían estimaciones de confiabilidad similares, excepto Openness. Para la Apertura, la cantidad de variación del error en la puntuación

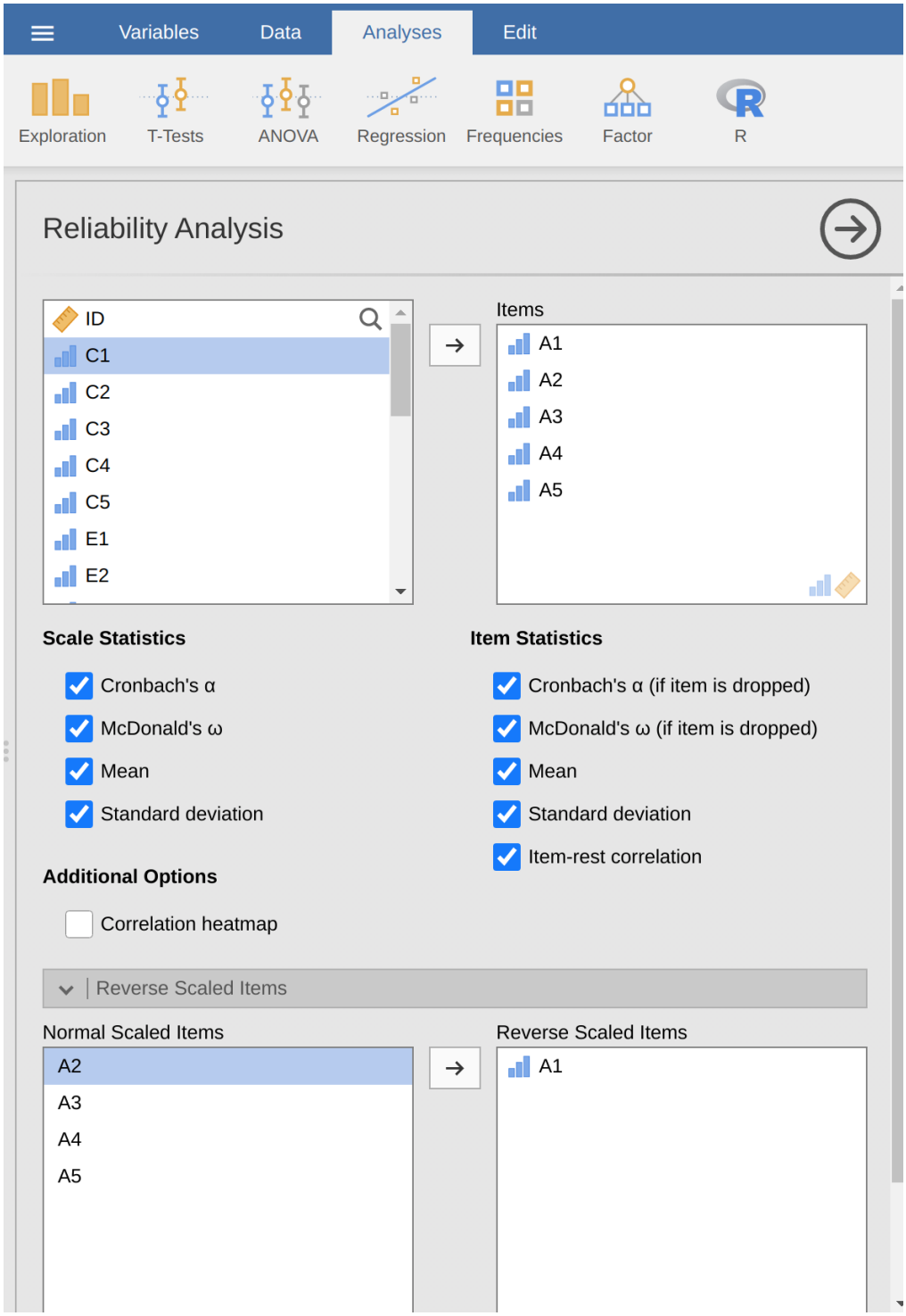


Figure 15.27: La ventana de análisis de confiabilidad jamovi

Reliability Analysis

Scale Reliability Statistics

	Mean	SD	Cronbach's α	McDonald's ω
scale	4.65	0.90	0.70	0.72

[3]

Item Reliability Statistics

				If item dropped	
	Mean	SD	Item-rest correlation	Cronbach's α	McDonald's ω
A1 [a]	4.63	1.37	0.29	0.72	0.73
A2	4.68	1.24	0.54	0.62	0.65
A3	4.64	1.30	0.61	0.59	0.61
A4	4.83	1.42	0.40	0.68	0.70
A5	4.48	1.33	0.47	0.64	0.66

[a] reverse scaled item

Figure 15.28: Los resultados del análisis de confiabilidad jamovi para el factor de amabilidad

de la escala es de alrededor del 40 %, lo cual es alto e indica que la Apertura es sustancialmente menos consistente como medida confiable de un atributo de personalidad que las otras escalas de personalidad.

15.6 Resumen

En este capítulo sobre análisis factorial y técnicas relacionadas, presentamos y demostramos análisis estadísticos que evalúan el patrón de relaciones en un conjunto de datos. Específicamente, hemos cubierto:

- Análisis Factorial Exploratorio (AFE). EFA es una técnica estadística para identificar factores latentes subyacentes en un conjunto de datos. Cada variable observada se conceptualiza como una representación del factor latente hasta cierto punto, indicado por una carga factorial. Los investigadores también utilizan EFA como una forma de reducción de datos, es decir, identificando variables observadas que pueden combinarse en nuevas variables de factores para análisis posteriores.
- Análisis de componentes principales (PCA) es una técnica de reducción de datos que, estrictamente hablando, no identifica factores latentes subyacentes. En cambio, PCA simplemente produce una combinación lineal de variables observadas.
- Análisis Factorial Confirmatorio (CFA). A diferencia de EFA, con CFA comienza con una idea, un modelo, de cómo las variables en sus datos se relacionan entre sí. Luego, prueba tu modelo con los datos observados y evalúa qué tan bien se ajusta el modelo a los datos.
- En [Multi-Trait Multi-Method CFA] (MTMM CFA), tanto el factor latente como la varianza del método se incluyen en el modelo en un enfoque que es útil cuando se utilizan diferentes enfoques metodológicos y, por lo tanto, la varianza del método es una consideración importante.
- [Análisis de fiabilidad de la consistencia interna]. Esta forma de análisis de confiabilidad prueba cuán consistentemente una escala mide una construcción de medición (psicológica).

Part VI

Finales, alternativas y perspectivas

Chapter 16

Estadística bayesianas

> *"En nuestros razonamientos relativos a los hechos, hay todos los grados imaginables de seguridad, desde la certeza más alta hasta la especie más baja de evidencia moral. Por lo tanto, un hombre sabio proporciona su creencia a la evidencia".*
> – David Hume [1]

Las ideas que le he presentado en este libro describen la estadística inferencial desde la perspectiva frecuentista. No estoy solo en hacer esto. De hecho, casi todos los libros de texto que se entregan a los estudiantes de psicología presentan las opiniones del estadístico frecuentista como *la* teoría de la estadística inferencial, la única forma verdadera de hacer las cosas. He enseñado de esta manera por razones prácticas. La visión frecuentista de la estadística dominó el campo académico de la estadística durante la mayor parte del siglo XX, y este dominio es aún más extremo entre los científicos aplicados. Era y es una práctica corriente entre los psicólogos utilizar métodos frecuentistas. Debido a que los métodos frecuentistas son omnipresentes en los artículos científicos, todos los estudiantes de estadística deben comprender esos métodos, de lo contrario, ¡no podrán entender lo que dicen esos artículos! Desafortunadamente, al menos en mi opinión, la práctica actual en psicología a menudo está equivocada y la dependencia de los métodos frecuentistas es en parte culpable. En este capítulo explico por qué pienso esto y ofrezco una introducción a la estadística bayesiana, un enfoque que creo que es generalmente superior al enfoque ortodoxo.

Este capítulo viene en dos partes. En las primeras tres secciones, hablo de qué se tratan las estadísticas bayesianas, cubriendo las reglas matemáticas básicas de cómo funciona, así como una explicación de por qué creo que el enfoque bayesiano es tan útil. Luego, proporciono una breve descripción general de cómo puede hacer pruebas t bayesianas.

[1] <a href="http://en.wikiquote.org/wiki/David%20Hume". target="_blank">http://en.wikiquote.org/wiki/David_Hume.

16.1 Razonamiento probabilístico por agentes racionales

Desde una perspectiva bayesiana, la inferencia estadística tiene que ver con la *revisión de creencias*. Comienzo con un conjunto de hipótesis candidatas h sobre el mundo. No sé cuál de estas hipótesis es verdadera, pero tengo algunas creencias sobre qué hipótesis son plausibles y cuáles no. Cuando observo los datos, d, tengo que revisar esas creencias. Si los datos son consistentes con una hipótesis, mi creencia en esa hipótesis se fortalece. Si los datos son inconsistentes con la hipótesis, mi creencia en esa hipótesis se debilita. ¡Eso es todo! Al final de esta sección, daré una descripción precisa de cómo funciona el razonamiento bayesiano, pero primero quiero trabajar con un ejemplo simple para presentar las ideas clave. Considere el siguiente problema de razonamiento.

Llevo un paraguas. ¿Crees que lloverá?

En este problema te he presentado un solo dato (d = llevo el paraguas) y te pido que me digas tu creencia o hipótesis sobre si está lloviendo. Tienes dos alternativas, h: o lloverá hoy o no lloverá. ¿Cómo deberías resolver este problema?

16.1.1 Prioridades: lo que creías antes

Lo primero que debes hacer es ignorar lo que te dije sobre el paraguas y escribir tus creencias preexistentes sobre la lluvia. Esto es importante. Si desea ser honesto acerca de cómo sus creencias han sido revisadas a la luz de nueva evidencia (datos), entonces debe decir algo sobre lo que creía antes de que aparecieran esos datos. Entonces, ¿qué podrías creer acerca de si lloverá hoy? Probablemente sepa que vivo en Australia y que gran parte de Australia es cálida y seca. La ciudad de Adelaide donde vivo tiene un clima mediterráneo, muy similar al sur de California, el sur de Europa o el norte de África. Estoy escribiendo esto en enero, así que puedes asumir que estamos en pleno verano. De hecho, es posible que haya decidido echar un vistazo rápido a Wikipedia[2] y haya descubierto que Adelaida recibe un promedio de 4,4 días de lluvia durante los 31 días de enero. Sin saber nada más, puede concluir que la probabilidad de lluvia en enero en Adelaide es de alrededor del 15 % y la probabilidad de un día seco es del 85 % (consulte Table 16.1). Si esto es realmente lo que crees sobre las lluvias en Adelaide (y ahora que te lo he dicho, apuesto a que esto realmente es lo que crees), entonces lo que he escrito aquí es tu **distribución anterior**, escrita $ P(h)$.

Table 16.1: ¿Qué tan probable es que llueva en Adelaide? Creencias preexistentes basadas en el conocimiento de la precipitación promedio de enero

Hypothesis	**Degree of Belief**
Rainy day	0.15
Dry day	0.85

16.1.2 Probabilidades: teorías sobre los datos

Para resolver el problema de razonamiento necesitas una teoría sobre mi comportamiento. ¿Cuándo lleva Dan un paraguas? Podrías adivinar que no soy un completo

[2] http://en.wikipedia.org/wiki/Climate_of_Adelaide

idiota, [3] y trato de llevar paraguas solo en días lluviosos. Por otro lado, también sabes que tengo niños pequeños, y no te sorprendería tanto saber que soy bastante olvidadizo con este tipo de cosas. Supongamos que en los días de lluvia recuerdo mi paraguas alrededor del 30% del tiempo (realmente soy terrible en esto). Pero digamos que en los días secos solo tengo un 5% de probabilidades de llevar un paraguas. Así que podría escribir esto como en Table 16.2.

Table 16.2: ¿Qué tan probable es que lleve un paraguas en días lluviosos y secos?

	Data	**Data**
Hypothesis	Umbrella	No umbrella
Rainy day	0.30	0.70
Dry day	0.05	0.95

Es importante recordar que cada celda de esta tabla describe sus creencias sobre qué datos d se observarán, *dada* la verdad de una hipótesis particular h. Esta "probabilidad condicional" se escribe $P(d|h)$, que se puede leer como "la probabilidad de d dada h". En las estadísticas bayesianas, esto se conoce como la **probabilidad** de los datos d dada la hipótesis h.[4]

16.1.3 La probabilidad conjunta de datos e hipótesis

En este punto todos los elementos están en su lugar. Habiendo anotado los antecedentes y la probabilidad, tiene toda la información que necesita para hacer un razonamiento bayesiano. La pregunta ahora es ¿cómo usamos esta información? Resulta que hay una ecuación muy simple que podemos usar aquí, pero es importante que comprenda por qué la usamos, así que intentaré desarrollarla a partir de ideas más básicas.

Comencemos con una de las reglas de la teoría de la probabilidad. Lo enumeré en Table 7.1, pero no le di mucha importancia en ese momento y probablemente lo ignoraste. La regla en cuestión es la que habla de la probabilidad de que dos cosas sean ciertas. En nuestro ejemplo, es posible que desee calcular la probabilidad de que hoy llueva (es decir, la hipótesis h es verdadera) y llevo un paraguas (es decir, se observan los datos d). La **probabilidad conjunta** de la hipótesis y los datos se escribe $P(d, h)$, y se puede calcular multiplicando la anterior $P(h)$ por la probabilidad $P(d|h)$. Matemáticamente, decimos que

$$P(d, h) = P(d|h)P(h)$$

[3] Es un acto de fe, lo sé, pero sigamos adelante, ¿de acuerdo?

[4] Um. Odio mencionar esto, pero algunos estadísticos se opondrían a que use la palabra "probabilidad" aquí. El problema es que la palabra "probabilidad" tiene un significado muy específico en las estadísticas frecuentistas, y no es lo mismo que lo que significa en las estadísticas bayesianas. Por lo que puedo decir, los bayesianos originalmente no tenían un nombre acordado para la probabilidad, por lo que se convirtió en una práctica común para las personas usar la terminología frecuentista. Esto no habría sido un problema excepto por el hecho de que la forma en que los bayesianos usan la palabra resulta ser bastante diferente a la forma en que lo hacen los frecuentistas. Este no es el lugar para otra larga lección de historia pero, para decirlo crudamente, cuando un bayesiano dice "una función de probabilidad" por lo general se refiere a una de las filas de la tabla. Cuando un frecuentador dice lo mismo, se refiere a la misma tabla, pero para ellos "una función de probabilidad" casi siempre se refiere a una de las columnas. Esta distinción es importante en algunos contextos, pero no es importante para nuestros propósitos.

Entonces, ¿cuál es la probabilidad de que hoy sea un día lluvioso y me acuerde de llevar un paraguas? Como comentamos anteriormente, el anterior nos dice que la probabilidad de un día lluvioso es del 15 %, y la probabilidad nos dice que la probabilidad de que me acuerde de mi paraguas en un día lluvioso es de 30%. Entonces, la probabilidad de que ambas cosas sean ciertas se calcula multiplicando las dos

$$\begin{aligned} P(lluvia, paraguas) &= P(paraguas|lluvia) \times P(lluvia) \\ &= 0.30 \times 0.15 \\ &= 0.045 \end{aligned}$$

En otras palabras, antes de que te digan nada de lo que realmente pasó, piensas que hay un 4,5% de probabilidad de que hoy sea un día lluvioso y que me acuerde de un paraguas. Sin embargo, por supuesto, hay cuatro cosas posibles que podrían suceder, ¿verdad? Así que repitamos el ejercicio para los cuatro. Si hacemos eso, terminamos con Table 16.3.

Table 16.3: Cuatro posibilidades combinando lluvia (o no) y paraguas (o no)

	Umbrella	**No-umbrella**
Rainy	0.045	0.105
Dry	0.0425	0.807

Esta tabla captura toda la información sobre cuál de las cuatro posibilidades es probable. Sin embargo, para obtener realmente una imagen completa, es útil sumar los totales de las filas y los totales de las columnas. Eso nos da Table 16.4.

Table 16.4: Cuatro posibilidades combinando lluvia (o no) y paraguas (o no), con totales de fila y columna

	Umbrella	**No-umbrella**	**Total**
Rainy	0.045	0.105	0.15
Dry	0.0425	0.807	0.85
Total	0.0875	0.912	1

Esta es una tabla muy útil, por lo que vale la pena tomarse un momento para pensar en lo que nos dicen todos estos números. Primero, observe que las sumas de las filas no nos dicen nada nuevo en absoluto. Por ejemplo, la primera fila nos dice que si ignoramos todo este asunto de los paraguas, la probabilidad de que hoy sea un día lluvioso es del 15 %. Eso no es sorprendente, por supuesto, ya que es nuestro anterior.[5] Lo importante no es el número en sí. Más bien, lo importante es que nos da cierta confianza en que nuestros cálculos son sensatos. Ahora eche un vistazo a las sumas de las columnas y observe que nos dicen algo que aún no hemos declarado explícitamente. De la misma manera que las sumas de las filas nos dicen la probabilidad de lluvia, las

[5]Para ser claros, la información "previa" es conocimiento o creencias preexistentes, antes de que recopilemos o usemos cualquier dato para mejorar esa información.

sumas de las columnas nos dicen la probabilidad de que lleve un paraguas. En concreto, la primera columna nos dice que de media (es decir, ignorando si es un día lluvioso o no) la probabilidad de que lleve paraguas es del 8,75%. Finalmente, observe que cuando sumamos los cuatro eventos lógicamente posibles, todo suma 1. En otras palabras, lo que hemos escrito es una distribución de probabilidad adecuada definida sobre todas las combinaciones posibles de datos e hipótesis.

Ahora, debido a que esta tabla es tan útil, quiero asegurarme de que comprenda a qué corresponden todos los elementos y cómo se escribieron (Table 16.5):

Table 16.5: Cuatro posibilidades que combinan lluvia (o no) y paraguas (o no), expresadas como probabilidades condicionales

	Umbrella	**No-umbrella**	
Rainy	P(Umbrella, Rainy)	P(No-umbrella, Rainy)	P(Rainy)
Dry	P(Umbrella, Dry)	P(No-umbrella, Dry)	P(Dry)
	P(Umbrella)	P(No-umbrella)	

Finalmente, usemos la notación estadística "adecuada". En el problema del día de lluvia, los datos corresponden a la observación de que tengo o no tengo paraguas. Así que dejaremos que d_1 se refiera a la posibilidad de que me observes con un paraguas, y d_2 se refiera a que me observes sin uno. De manera similar, h_1 es tu hipótesis de que hoy llueve, y h_2 es la hipótesis de que no. Usando esta notación, la tabla se parece a Table 16.6.

Table 16.6: Cuatro posibilidades que combinan lluvia (o no) y paraguas (o no), expresadas en términos hipotéticos como probabilidades condicionales

	d_1	d_2	
h_1	$P(h_1, d_1)$	$P(h_1, d_2)$	$P(h_1)$
h_2	$P(h_2, d_1)$	$P(h_2, d_2)$	$P(h_2)$
	$P(d_1)$	$P(d_2)$	

16.1.4 Actualización de creencias usando la regla de Bayes

La tabla que presentamos en la última sección es una herramienta muy poderosa para resolver el problema del día lluvioso, ya que considera las cuatro posibilidades lógicas y establece exactamente qué tan seguro está en cada una de ellas antes de recibir datos. Ahora es el momento de considerar qué sucede con nuestras creencias cuando en realidad se nos dan los datos. En el problema del día lluvioso, se le dice que realmente llevo un paraguas. Esto es algo así como un evento sorprendente. Según nuestra tabla, la probabilidad de que lleve un paraguas es solo del 8,75 %. Pero eso tiene sentido,

¿verdad? Una mujer que lleva un paraguas en un día de verano en una ciudad calurosa y seca es bastante inusual, por lo que realmente no esperabas eso. Sin embargo, los datos te dicen que es cierto. No importa cuán improbable pensara que era, ahora debe ajustar sus creencias para adaptarse al hecho de que ahora *sabe* que tengo un paraguas.[6] Para reflejar este nuevo conocimiento, nuestra tabla *revisada* debe tener los siguientes números. (ver Table 16.7).

Table 16.7: Revisión de creencias dados nuevos datos sobre llevar paraguas

	Umbrella	**No-umbrella**
Rainy		0
Dry		0
Total	1	0

En otras palabras, los hechos han eliminado cualquier posibilidad de "sin paraguas", por lo que tenemos que poner ceros en cualquier celda de la tabla que implique que no llevo paraguas. Además, sabes con certeza que llevo un paraguas, por lo que la suma de la columna de la izquierda debe ser 1 para describir correctamente el hecho de que $P(paraguas) = 1$.

¿Qué dos números debemos poner en las celdas vacías? Una vez más, no nos preocupemos por las matemáticas y, en su lugar, pensemos en nuestras intuiciones. Cuando escribimos nuestra tabla por primera vez, resultó que esas dos celdas tenían números casi idénticos, ¿verdad? Calculamos que la probabilidad conjunta de "lluvia y paraguas" era del 4,5 %, y la probabilidad conjunta de "seco y paraguas" era del 4,25 %. En otras palabras, antes de que te dijera que de hecho llevo un paraguas, habrías dicho que estos dos eventos eran casi idénticos en probabilidad, ¿no? Pero observe que ambas posibilidades son consistentes con el hecho de que en realidad llevo un paraguas. Desde la perspectiva de estas dos posibilidades, muy poco ha cambiado. Espero que esté de acuerdo en que sigue siendo cierto que estas dos posibilidades son igualmente plausibles. Entonces, lo que esperamos ver en nuestra tabla final son algunos números que preservan el hecho de que "lluvia y paraguas" es *ligeramente* más plausible que "seco y paraguas", al mismo tiempo que asegura que los números en la tabla suman. ¿Algo como Table 16.8, quizás?

Table 16.8: Revisión de probabilidades dados nuevos datos sobre llevar paraguas

	Umbrella	**No-umbrella**
Rainy	0.514	0
Dry	0.486	0
Total	1	0

Lo que esta tabla te está diciendo es que, después de que te digan que llevo un paraguas, crees que hay un 51,4 % de posibilidades de que hoy sea un día lluvioso y un 48,6 % de que no. ¡Esa es la respuesta a nuestro problema! La **probabilidad posterior** de que llueva $P(h\|d)$ dado que llevo paraguas es del 51,4%

[6] si fuéramos un poco más sofisticados, podríamos extender el ejemplo para acomodar la posibilidad de que esté mintiendo sobre el paraguas. Pero mantengamos las cosas simples, ¿de acuerdo?

¿Cómo calculé estos números? Probablemente puedas adivinar. Para calcular que había una probabilidad de 0.514 de "lluvia", todo lo que hice fue tomar la probabilidad de 0.045 de "lluvia y paraguas" y dividirla por la probabilidad de 0.0875 de "paraguas". Esto produce una tabla que satisface nuestra necesidad de que todo sume 1 y nuestra necesidad de no interferir con la verosimilitud relativa de los dos eventos que en realidad son consistentes con los datos. Para decir lo mismo usando la jerga estadística sofisticada, lo que he hecho aquí es dividir la probabilidad conjunta de la hipótesis y los datos $P(d, h)$ por la **probabilidad marginal** de los datos $P(d)$, y esto es lo que nos da la probabilidad posterior de la hipótesis dados los datos que se han observado. Para escribir esto como una ecuación: [7]

$$P(h|d) = \frac{P(d, h)}{P(d)}$$

Sin embargo, recuerda lo que dije al comienzo de la última sección, a saber, que la probabilidad conjunta $P(d, h)$ se calcula multiplicando el Pphq anterior por la probabilidad $P(d|h)$. En la vida real, las cosas que realmente sabemos escribir son los antecedentes y la probabilidad, así que volvamos a sustituirlos en la ecuación. Esto nos da la siguiente fórmula para la probabilidad posterior:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

Y esta fórmula, amigos, se conoce como **regla de Bayes**. Describe cómo un alumno comienza con creencias previas sobre la plausibilidad de diferentes hipótesis y le dice cómo se deben revisar esas creencias frente a los datos. En el paradigma bayesiano, todas las inferencias estadísticas fluyen de esta regla simple.

16.2 Pruebas de hipótesis bayesianas

En Chapter 9 describí el enfoque ortodoxo para la prueba de hipótesis. Tomó un capítulo entero para describirlo, porque la prueba de hipótesis nula es un artilugio muy elaborado que a la gente le resulta muy difícil entender. Por el contrario, el enfoque bayesiano para la prueba de hipótesis es increíblemente simple. Escojamos un escenario que sea muy parecido al escenario ortodoxo. Hay dos hipótesis que queremos comparar, una hipótesis nula h_0 y una hipótesis alternativa h_1. Antes de ejecutar el experimento, tenemos algunas creencias $P(h)$ sobre qué hipótesis son verdaderas. Realizamos un experimento y obtenemos datos d. A diferencia de la estadística frecuentista, la estadística bayesiana sí nos permite hablar de la probabilidad de que la hipótesis nula sea cierta. Mejor aún, nos permite calcular la **probabilidad posterior de la hipótesis nula**, usando la regla de Bayes:

$$P(h_0|d) = \frac{P(d|h_0)P(h_0)}{P(d)}$$

[7] puede notar que esta ecuación es en realidad una reafirmación de la misma regla básica que enumeré al comienzo de la última sección. Si multiplica ambos lados de la ecuación por $P(d)$, obtiene $P(d)P(h|d) = P(d, h)$, que es la regla para calcular las probabilidades conjuntas. Así que en realidad no estoy introduciendo ninguna regla "nueva" aquí, solo estoy usando la misma regla de una manera diferente.

Esta fórmula nos dice exactamente cuánta creencia debemos tener en la hipótesis nula después de haber observado los datos d. De manera similar, podemos calcular cuánta creencia colocar en la hipótesis alternativa usando esencialmente la misma ecuación. Todo lo que hacemos es cambiar el subíndice

$$P(h_1|d) = \frac{P(d|h_1)P(h_1)}{P(d)}$$

Es todo tan simple que me siento como un idiota incluso molestándome en escribir estas ecuaciones, ya que todo lo que estoy haciendo es copiar la regla de Bayes de la sección anterior.[8]

16.2.1 El factor de Bayes

En la práctica, la mayoría de los analistas de datos bayesianos tienden a no hablar en términos de probabilidades posteriores sin procesar $P(h_0|d)$ y $P(h_1|d)$. En cambio, tendemos a hablar en términos de la **razón de probabilidades posterior**. Piense en ello como apostar. Supongamos, por ejemplo, que la probabilidad posterior de la hipótesis nula es del 25 % y la probabilidad posterior de la alternativa es del 75 %. La hipótesis alternativa es tres veces más probable que la nula, por lo que decimos que las probabilidades son de 3:1 a favor de la alternativa. Matemáticamente, todo lo que tenemos que hacer para calcular las probabilidades posteriores es dividir una probabilidad posterior entre la otra

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{0.75}{0.25} = 3$$

O, para escribir lo mismo en términos de las ecuaciones anteriores

$$\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0)} \times \frac{P(h_1)}{P(h_0)}$$

En realidad, vale la pena ampliar esta ecuación. Aquí hay tres términos diferentes que debe conocer. En el lado izquierdo, tenemos las probabilidades posteriores, que te dicen lo que crees sobre la verosimilitud relativa de la hipótesis nula y la hipótesis alternativa después de ver los datos. En el lado derecho, tenemos las **cuotas previas**, que indican lo que pensabas antes de ver los datos. En el medio, tenemos el **factor de Bayes**, que describe la cantidad de evidencia proporcionada por los datos. (Table 16.9).

El factor de Bayes (a veces abreviado como BF) tiene un lugar especial en la prueba de hipótesis bayesiana, porque cumple una función similar al valor p en la prueba de hipótesis ortodoxa. El factor de Bayes cuantifica la fuerza de la evidencia proporcionada por los datos y, como tal, es el factor de Bayes que las personas tienden a informar cuando realizan una prueba de hipótesis bayesiana. La razón para informar los factores de Bayes en lugar de las probabilidades posteriores es que diferentes investigadores tendrán

[8] Obviamente, esta es una historia muy simplificada. Toda la complejidad de las pruebas de hipótesis bayesianas de la vida real se reduce a cómo se calcula la probabilidad $P(d\|h)$ cuando la hipótesis h es algo complejo y vago. No voy a hablar sobre esas complejidades en este libro, pero sí quiero resaltar que, aunque esta simple historia es cierta hasta donde llega, la vida real es más complicada de lo que puedo cubrir en un libro de texto de introducción a las estadísticas.

Table 16.9: Cuotas posteriores dado el factor Bsyes y cuotas previas

$\frac{P(h_1\|d)}{h_0\|d}$	=	$\frac{P(d\|h_1)}{d\|h_0}$	×	$\frac{P(h_1)}{h_0}$
⇑		⇑		⇑
Posterior odds		Bayes factor		Prior odds

antecedentes diferentes. Algunas personas pueden tener un fuerte sesgo para creer que la hipótesis nula es verdadera, otras pueden tener un fuerte sesgo para creer que es falsa. Debido a esto, lo cortés que debe hacer un investigador aplicado es informar el factor de Bayes. De esa manera, cualquier persona que lea el periódico puede multiplicar el factor de Bayes por sus propias probabilidades previas personales, y puede calcular por sí mismo cuáles serían las probabilidades posteriores. En cualquier caso, por convención nos gusta pretender que damos igual consideración tanto a la hipótesis nula como a la alternativa, en cuyo caso la probabilidad anterior es igual a 1, y la probabilidad posterior se vuelve igual al factor de Bayes.

16.2.2 Interpretación de los factores de Bayes

Una de las cosas realmente buenas del factor de Bayes es que los números son inherentemente significativos. Si ejecuta un experimento y calcula un factor de Bayes de 4, significa que la evidencia proporcionada por sus datos corresponde a probabilidades de apuestas de 4:1 a favor de la alternativa. Sin embargo, ha habido algunos intentos de cuantificar los estándares de evidencia que se considerarían significativos en un contexto científico. Los dos más utilizados son de Jeffreys (1961) y Kass & Raftery (1995). De los dos, tiendo a preferir la tabla Kass & Raftery (1995) porque es un poco más conservadora. Así que aquí está (Table 16.10).

Table 16.10: factores de Bayes y fuerza de la evidencia

Bayes factor	**Interpretation**
1 - 3	Negligible evidence
3 - 20	Positive evidence
20 - 150	Strong evidence
> 150	Very strong evidence

Y para ser perfectamente honesto, creo que incluso los estándares de Kass & Raftery (1995) están siendo un poco caritativos. Si fuera por mí, habría llamado a la categoría de "evidencia positiva" "evidencia débil". Para mí, cualquier cosa en el rango de 3:1 a 20:1 es evidencia "débil" o "modesta" en el mejor de los casos. Pero no hay reglas estrictas y rápidas aquí. Lo que cuenta como evidencia fuerte o débil depende completamente de qué tan conservador sea usted y de los estándares en los que insista su comunidad antes de estar dispuesta a etiquetar un hallazgo como "verdadero".

En cualquier caso, tenga en cuenta que todos los números enumerados anteriormente tienen sentido si el factor de Bayes es mayor que 1 (es decir, la evidencia favorece la hipótesis alternativa). Sin embargo, una gran ventaja práctica del enfoque bayesiano en relación con el enfoque ortodoxo es que también le permite cuantificar la evidencia

del nulo. Cuando eso suceda, el factor de Bayes será menor que 1. Puede optar por informar un factor de Bayes menor que 1, pero para ser honesto, lo encuentro confuso. Por ejemplo, suponga que la probabilidad de los datos bajo la hipótesis nula $P(d|h_0)$ es igual a 0,2, y la probabilidad correspondiente $P(d|h_1)$ bajo la hipótesis alternativa es 0,1. Usando las ecuaciones dadas arriba, el factor de Bayes aquí sería

$$BF = \frac{P(d|h_1)}{P(d|h_0)} = \frac{0.1}{0.2} = 0.5$$

Leído literalmente, este resultado dice que la evidencia a favor de la alternativa es de 0.5 a 1. Encuentro esto difícil de entender. Para mí, tiene mucho más sentido poner la ecuación "al revés" e informar la cantidad de evidencia a favor del valor nulo. En otras palabras, lo que calculamos es esto

$$BF' = \frac{P(d|h_0)}{P(d|h_1)} = \frac{0.2}{0.1} = 2$$

Y lo que reportaríamos es un factor de Bayes de 2:1 a favor del nulo. Mucho más fácil de entender, y puede interpretar esto usando la tabla de arriba.

16.3 ¿Por qué ser bayesiano?

Hasta este punto me he centrado exclusivamente en la lógica que sustenta las estadísticas bayesianas. Hemos hablado sobre la idea de "probabilidad como un grado de creencia" y lo que implica sobre cómo un agente racional debería razonar sobre el mundo. La pregunta que tienes que responderte a ti mismo es esta: ¿cómo quieres hacer tus estadísticas? ¿Quiere ser un estadístico ortodoxo y basarse en distribuciones muestrales y valores p para guiar sus decisiones? ¿O quiere ser bayesiano, confiando en cosas como creencias previas, factores de Bayes y las reglas para la revisión de creencias racionales? Y para ser perfectamente honesto, no puedo responder esta pregunta por ti. En última instancia, depende de lo que creas que es correcto. Es su llamada y solo su llamada. Dicho esto, puedo hablar un poco sobre por qué prefiero el enfoque bayesiano.

16.3.1 Estadísticas que significan lo que crees que significan

> *Sigues usando esa palabra. No creo que signifique lo que crees que significa*
> – Íñigo Montoya, La princesa prometida [9]

Para mí, una de las mayores ventajas del enfoque bayesiano es que responde a las preguntas correctas. Dentro del marco bayesiano, es perfectamente sensato y permisible referirse a "la probabilidad de que una hipótesis sea verdadera". Incluso puedes intentar calcular esta probabilidad. En última instancia, ¿no es eso lo que quiere que le digan sus pruebas estadísticas? Para un ser humano real, esto parecería ser el objetivo principal de hacer estadísticas, es decir, determinar qué es verdad y qué no lo es. Cada vez que no esté exactamente seguro de cuál es la verdad, debe usar el lenguaje de la teoría de

[9] http://www.imdb.com/title/tt0093779/quotes . Debo señalar de paso que no soy la primera persona que usa esta cita para quejarse de los métodos frecuentadores. Rich Morey y sus colegas tuvieron la idea primero. Lo estoy robando descaradamente porque es una cita increíble para usar en este contexto y me niego a perder cualquier oportunidad de citar *La princesa prometida*.

la probabilidad para decir cosas como "hay un 80 % de posibilidades de que la teoría A sea cierta, pero un 20 % de posibilidades de que la teoría B sea cierta". en cambio".

Esto parece tan obvio para un ser humano, pero está explícitamente prohibido dentro del marco ortodoxo. Para un frecuentador, tales declaraciones son una tontería porque "la teoría es verdadera" no es un evento repetible. Una teoría es verdadera o no lo es, y no se permiten declaraciones probabilísticas, sin importar cuánto quieras hacerlas. Hay una razón por la cual, en la Sección 9.5, le advertí repetidamente que no interpretara el valor p como la probabilidad de que la hipótesis nula sea verdadera. Hay una razón por la que casi todos los libros de texto sobre estadísticas se ven obligados a repetir esa advertencia. Es porque la gente quiere desesperadamente que esa sea la interpretación correcta. A pesar del dogma frecuentista, una vida de experiencia enseñando a estudiantes universitarios y haciendo análisis de datos a diario me sugiere que la mayoría de los humanos reales piensan que "la probabilidad de que la hipótesis sea cierta" no solo es significativa, es lo que más nos importa. . Es una idea tan atractiva que incluso los estadísticos capacitados caen presa del error de tratar de interpretar un valor p de esta manera. Por ejemplo, aquí hay una cita de un informe oficial de Newspoll en 2013, que explica cómo interpretar su análisis de datos (frecuentista): [10]

> *A lo largo del informe, en su caso, se han observado cambios estadísticamente significativos. Todas las pruebas de significación se han basado en el nivel de confianza del 95 por ciento.* ***Esto significa que si se observa que un cambio es estadísticamente significativo, existe un 95 por ciento de probabilidad de que haya ocurrido un cambio real****, y no se debe simplemente a una variación aleatoria. (énfasis añadido)*

¡No! Eso no es lo que significa $p < .05$. Eso no es lo que significa un 95% de confianza para un estadístico frecuentista. La sección en negrita es simplemente incorrecta. Los métodos ortodoxos no pueden decirle que "hay un 95% de posibilidades de que haya ocurrido un cambio real", porque este no es el tipo de evento al que se pueden asignar probabilidades frecuentistas. Para un frecuentador ideológico, esta frase no debería tener sentido. Incluso si eres un frecuentador más pragmático, sigue siendo la definición incorrecta de un valor p. Simplemente no está permitido o es correcto decirlo si desea confiar en las herramientas estadísticas ortodoxas.

Por otro lado, supongamos que eres bayesiano. Aunque el pasaje en negrita es la definición incorrecta de un valor p, es más o menos exactamente lo que quiere decir un bayesiano cuando dice que la probabilidad posterior de la hipótesis alternativa es superior al 95%. Y aquí está la cosa. Si el posterior bayesiano es en realidad lo que desea informar, ¿por qué está tratando de usar métodos ortodoxos? Si desea hacer afirmaciones bayesianas, todo lo que tiene que hacer es ser bayesiano y usar herramientas bayesianas.

Hablando por mí mismo, descubrí que esto es lo más liberador de cambiar a la vista bayesiana. Una vez que haya dado el salto, ya no tendrá que envolver su cabeza en definiciones contrarias a la intuición de los valores p. No tiene que molestarse en recordar por qué no puede decir que está 95% seguro de que la verdadera media se encuentra dentro de algún intervalo. Todo lo que tiene que hacer es ser honesto acerca de lo que creía antes de realizar el estudio y luego informar lo que aprendió al hacerlo. Suena bien, ¿no? Para mí, esta es la gran promesa del enfoque bayesiano. Usted hace el análisis que

[10] http ://about.abc.net.au/reports-publications/appreciation-survey-summary-report-2013/

realmente quiere hacer y expresa lo que realmente cree que le están diciendo los datos.

16.3.2 Estándares probatorios en los que puede creer

> *Si p* está por debajo de .02, esto indica claramente que la hipótesis *nula* no explica la totalidad de los hechos. No nos equivocaremos a menudo si trazamos una línea convencional en .05 y consideramos que los valores más pequeños de *p* indican una discrepancia real.
> – Sir Ronald Fisher (Fisher, 1925)

Considere la cita anterior de Sir Ronald Fisher, uno de los fundadores de lo que se ha convertido en el enfoque ortodoxo de las estadísticas. Si alguien alguna vez ha tenido derecho a expresar una opinión sobre la función prevista de los valores p, es Fisher. En este pasaje, tomado de su guía clásica Métodos estadísticos para trabajadores de la investigación, deja bastante claro lo que significa rechazar una hipótesis nula en $p < .05$. En su opinión, si consideramos que $p < 0{,}05$ significa que hay "un efecto real", entonces "no nos equivocaremos a menudo". Esta vista no es inusual. En mi experiencia, la mayoría de los practicantes expresan puntos de vista muy similares a los de Fisher. En esencia, se supone que la convención $p < .05$ representa un estándar probatorio bastante estricto.

Bueno, ¿qué tan cierto es eso? Una forma de abordar esta pregunta es tratar de convertir los valores p en factores de Bayes y ver cómo se comparan los dos. No es algo fácil de hacer porque un valor p es un tipo de cálculo fundamentalmente diferente a un factor de Bayes, y no miden lo mismo. Sin embargo, ha habido algunos intentos de resolver la relación entre los dos, y es algo sorprendente. Por ejemplo, Johnson (2013) presenta un caso bastante convincente de que (al menos para las pruebas t) el umbral $p < .05$ corresponde aproximadamente a un factor de Bayes de entre 3:1 y 5:1 a favor de la alternativa. Si eso es correcto, entonces la afirmación de Fisher es un poco exagerada. Supongamos que la hipótesis nula es cierta aproximadamente la mitad de las veces (es decir, la probabilidad previa de H_0 es 0,5), y usamos esos números para calcular la probabilidad posterior de la hipótesis nula dado que ha sido rechazada en $p < .05$. Utilizando los datos de Johnson (2013), vemos que si rechaza el valor nulo en p ă .05, estará en lo correcto aproximadamente el 80 % de las veces. No sé usted, pero, en mi opinión, un estándar probatorio que le asegure que se equivocará en el 20 % de sus decisiones no es suficiente. El hecho es que, contrariamente a la afirmación de Fisher, si rechaza en $p < 0{,}05$, muy a menudo se equivocará. No es un umbral probatorio muy estricto en absoluto.

16.3.3 El valor p es una mentira.

> *El pastel es mentira.*
> *El pastel es mentira.*
> *El pastel es mentira.*
> *El pastel es mentira.*
> – Portal[11]

Bien, en este punto podrías estar pensando que el verdadero problema no es con las estadísticas ortodoxas, sino con el estándar $p < .05$. En cierto sentido, eso es cierto. La recomendación que da Johnson (2013) no es que "todos deben ser bayesianos ahora". En

[11] http://knowyourmeme.com/memes/the-cake-is-a-lie .

cambio, la sugerencia es que sería más inteligente cambiar el estándar convencional a algo así como un nivel de $p < .01$. Esa no es una opinión irrazonable, pero en mi opinión, el problema es un poco más grave que eso. En mi opinión, hay un problema bastante grande en la forma en que se construyen la mayoría (pero no todas) las pruebas de hipótesis ortodoxas. Son groseramente ingenuos acerca de cómo los humanos realmente investigan y, debido a esto, la mayoría de los valores de p están equivocados.

Suena como una afirmación absurda, ¿verdad? Bueno, considere el siguiente escenario. Se le ocurrió una hipótesis de investigación realmente emocionante y diseñó un estudio para probarla. Eres muy diligente, así que ejecutas un análisis de potencia para determinar cuál debería ser el tamaño de la muestra y ejecutas el estudio. Ejecutas tu prueba de hipótesis y obtienes un valor p de 0.072. Realmente jodidamente molesto, ¿verdad?

¿Qué debes hacer? Aquí hay algunas posibilidades:

1. Concluyes que no hay efecto e intentas publicarlo como resultado nulo
2. Supone que podría haber un efecto e intenta publicarlo como un resultado "en el límite significativo"
3. Te rindes e intentas un nuevo estudio
4. Reúne algunos datos más para ver si el valor p sube o (¡preferiblemente!) cae por debajo del criterio "mágico" de $p < .05$

¿Cuál escogerías? Antes de seguir leyendo, le insto a que se tome un tiempo para pensarlo. Se honesto contigo mismo. Pero no te preocupes demasiado por eso, porque estás jodido sin importar lo que elijas. Basado en mis propias experiencias como autor, revisor y editor, así como en las historias que escuché de otros, esto es lo que sucederá en cada caso:

- Comencemos con la opción 1. Si intenta publicarlo como un resultado nulo, el artículo tendrá dificultades para publicarse. Algunos revisores pensarán que $p = .072$ no es realmente un resultado nulo. Argumentarán que está en el límite significativo. Otros revisores estarán de acuerdo en que es un resultado nulo, pero afirmarán que, aunque algunos resultados nulos son publicables, el suyo no lo es. Uno o dos revisores podrían incluso estar de su lado, pero tendrá que luchar una batalla cuesta arriba para lograrlo.

- Bien, pensemos en la opción número 2. Supongamos que intenta publicarlo como un resultado límite significativo. Algunos revisores afirmarán que es un resultado nulo y que no debería publicarse. Otros afirmarán que la evidencia es ambigua y que debe recopilar más datos hasta que obtenga un resultado claro y significativo. Una vez más, el proceso de publicación no le favorece.

- Dadas las dificultades para publicar un resultado "ambiguo" como $p = .072$, la opción número 3 puede parecer tentadora: rendirse y hacer otra cosa. Pero esa es una receta para el suicidio profesional. Si te rindes y pruebas un nuevo proyecto cada vez que te enfrentas a la ambigüedad, tu trabajo nunca se publicará. Y si estás en la academia sin un registro de publicación puedes perder tu trabajo. Así que esa opción está descartada.

- Parece que está atascado con la opción 4. No tiene resultados concluyentes, por lo que decide recopilar más datos y volver a ejecutar el análisis. Parece sensato, pero desafortunadamente para usted, si hace esto, todos sus valores p ahora son incorrectos. Todos ellos. No solo los valores p que calculó para este estudio.

> Todos ellos. Todos los valores p que calculó en el pasado y todos los valores p que calculará en el futuro. Afortunadamente, nadie se dará cuenta. Te publicarán y habrás mentido.

¿Esperar lo? ¿Cómo puede ser cierta esa última parte? Quiero decir, suena como una estrategia perfectamente razonable, ¿no? Recolectó algunos datos, los resultados no fueron concluyentes, por lo que ahora lo que desea hacer es recopilar más datos hasta que los resultados sean concluyentes. ¿Qué está mal con eso?

Honestamente, no hay nada de malo en ello. Es algo razonable, sensato y racional. En la vida real, esto es exactamente lo que hace todo investigador. Desafortunadamente, la teoría de nula Prueba de hipótesis como la describí en un capítulo anterior le prohíbe hacer esto.[12] La razón es que la teoría asume que el experimento ha terminado y todos los datos están in. Y debido a que asume que el experimento ha terminado, solo considera dos decisiones posibles. Si usa el umbral convencional $p < .05$, esas decisiones se muestran en Table 16.11.

Table 16.11: prueba de significación de hipótesis nula convencional (NHST) con $p < 0{,}05$)

Outcome	Action
p less than .05	Reject the null
p greater than .05	Retain the null

Lo que estás haciendo es agregar una tercera acción posible al problema de toma de decisiones. Específicamente, lo que estás haciendo es usar el valor p como una razón para justificar continuar con el experimento. Y como consecuencia, ha transformado el procedimiento de toma de decisiones en uno que se parece más a Table 16.12.

Table 16.12: llevar a cabo la recopilación de datos en función de los valores p obtenidos en las pruebas preliminares

Outcome	Action
p less than .05	Stop the experiment and reject the null
p between .05 and .1	Continue the experiment
p greater than .1	Stop the experiment and retain the null

La teoría "básica" de nula prueba de hipótesis no está construida para manejar este tipo de cosas, no en la forma que describí en ese capítulo anterior. Si usted es el tipo

[12] Para ser completamente honesto, debo reconocer que no todas las pruebas estadísticas ortodoxas se basan en esta suposición tonta. Hay una serie de herramientas de análisis secuencial que a veces se utilizan en ensayos clínicos y similares. Estos métodos se basan en el supuesto de que los datos se analizan a medida que llegan, y estas pruebas no se rompen terriblemente en la forma en que me quejo aquí. Sin embargo, los métodos de análisis secuencial se construyen de una manera muy diferente a la versión "estándar" de la prueba de hipótesis nula. No se incluyen en ningún libro de texto introductorio y no se utilizan mucho en la literatura psicológica. La preocupación que planteo aquí es válida para todas las pruebas ortodoxas que he presentado hasta ahora y para casi todas las pruebas que he visto reportadas en los artículos que leí.

de persona que elegiría "recolectar más datos" en la vida real, eso implica que no está tomando decisiones de acuerdo con las reglas de la prueba de hipótesis nula. Incluso si llega a la misma decisión que la prueba de hipótesis, no está siguiendo el proceso de decisión que implica, y es esta falla en seguir el proceso lo que está causando el problema.[13] Su p -Los valores son una mentira.

Peor aún, son una mentira de una manera peligrosa, porque todos son *demasiado pequeños*. Para darle una idea de lo malo que puede ser, considere el siguiente escenario (en el peor de los casos). Imagina que eres un investigador súper entusiasta con un presupuesto ajustado que no prestó atención a mis advertencias anteriores. Usted diseña un estudio comparando dos grupos. Desea desesperadamente ver un resultado significativo en el nivel $p < .05$, pero realmente no desea recopilar más datos de los necesarios (porque es costoso). Para reducir los costos, comienza a recopilar datos, pero cada vez que llega un conjunto de observaciones, ejecuta una prueba t en sus datos. Si la prueba t dice $p < .05$, detiene el experimento e informa un resultado significativo. Si no, sigue recopilando datos. Siga haciendo esto hasta que alcance su límite de gasto predefinido para este experimento. Digamos que el límite se activa en $N = 1000$ observaciones. Como resultado, la verdad del asunto es que no se puede encontrar ningún efecto real: la hipótesis nula es verdadera. Entonces, ¿cuál es la probabilidad de que llegues al final del experimento y (correctamente) concluyas que no hay efecto? En un mundo ideal, la respuesta aquí debería ser 95%. Después de todo, el punto central del criterio $p < .05$ es controlar la tasa de error Tipo I al 5 %, por lo que lo que esperamos es que solo haya un 5 % de posibilidades de rechazar falsamente la hipótesis nula en esta situación. . Sin embargo, no hay garantía de que eso sea cierto. Estás rompiendo las reglas. Debido a que está ejecutando pruebas repetidamente, "echando un vistazo" a sus datos para ver si ha obtenido un resultado significativo, todas las apuestas están canceladas.

Entonces, ¿qué tan malo es? La respuesta se muestra como una línea sólida en Figure 16.1, y es asombrosamente mala. Si echa un vistazo a sus datos después de cada observación, hay un 53% de posibilidades de que cometa un error de tipo I. Eso es, um, un poco más grande que el 5% que se supone que es. Y no mejora mucho con un vistazo menos frecuente: si solo miras cada 10 o cada 50 observaciones. entonces las tasas de error de Tipo I siguen siendo demasiado altas: 38% y 29%, respectivamente. A modo de comparación, imagine que ha utilizado la siguiente estrategia. Comience a recopilar datos. Cada vez que llegue una observación, ejecute pruebas t bayesianas y mire el factor de Bayes. Asumiré que Johnson (2013) tiene razón, y trataré un factor de Bayes de 3:1 como aproximadamente equivalente a un valor p de .05.[14] Esta vez, nuestro investigador de gatillo feliz utiliza el siguiente procedimiento. Si el factor de Bayes es 3:1 o más a favor del nulo, detenga el experimento y conserve el nulo. Si es 3:1 o más a favor de la alternativa, detenga el experimento y rechace el nulo. De lo contrario, continúe probando. Ahora, como la última vez, supongamos que la hipótesis nula es verdadera. ¿Lo que sucede? Da la casualidad de que también ejecuté las simulaciones para este escenario y los resultados se muestran como la línea discontinua en @ fig-fig16-1. Resulta que la tasa de error de Tipo I para echar un vistazo cada vez que llega una nueva observación es del 24%, mucho más baja que la tasa del 53% que obtuvimos al usar la prueba t ortodoxa. Y para asomarse cada 10 o 50 observaciones

[13] un problema relacionado: http://xkcd.com/1478/ .

[14] Algunos lectores podrían preguntarse por qué elegí 3:1 en lugar de 5:1, dado que Johnson (2013) sugiere que $p = 0,05$ se encuentra en algún lugar de ese rango. Lo hice para ser caritativo con el valor p. Si hubiera elegido un factor Bayesiano de 5:1, los resultados se verían incluso mejor para el enfoque bayesiano.

las tasas son del 11% y 8%, respectivamente.

En cierto modo, esto es notable. Todo el *punto* de la prueba de hipótesis nula ortodoxa es controlar la tasa de error Tipo I. Los métodos bayesianos en realidad no están diseñados para hacer esto en absoluto. Sin embargo, resulta que cuando se enfrenta a un investigador de "gatillo feliz" que continúa realizando pruebas de hipótesis a medida que ingresan los datos, el enfoque bayesiano es mucho más efectivo. Incluso el estándar 3:1, que la mayoría de los bayesianos consideraría inaceptablemente laxo, es mucho más seguro que la regla p < 0,05.

16.3.4 ¿Es realmente tan malo?

El ejemplo que di en la sección anterior es una situación bastante extrema. En la vida real, las personas no realizan pruebas de hipótesis cada vez que llega una nueva observación. Por lo tanto, no es justo decir que el umbral p < .05 "realmente" corresponde a una tasa de error Tipo I del 53% (es decir, p = 0.53). Pero el hecho es que si desea que sus valores p sean honestos, debe cambiar a una forma completamente diferente de hacer pruebas de hipótesis o aplicar una regla estricta de no mirar a escondidas. No se le permite utilizar los datos para decidir cuándo finalizar el experimento. No se le permite mirar un valor p "límite" y decidir recopilar más datos. Ni siquiera se le permite cambiar su estrategia de análisis de datos después de mirar los datos. Está estrictamente obligado a seguir estas reglas, de lo contrario, los valores p que calcule no tendrán sentido.

Y sí, estas reglas son sorprendentemente estrictas. Como ejercicio de clase hace un par de años, les pedí a los estudiantes que pensaran en este escenario. Suponga que comenzó a realizar su estudio con la intención de reunir $N = 80$ personas. Cuando comienza el estudio, sigues las reglas y te niegas a mirar los datos o realizar cualquier prueba. Pero cuando llegas a $N = 50$ tu fuerza de voluntad cede… y echas un vistazo. ¿Adivina qué? ¡Tienes un resultado significativo! Ahora, claro, sabes que dijiste que seguirías realizando el estudio con un tamaño de muestra de $N = 80$, pero parece un poco inútil ahora, ¿verdad? El resultado es significativo con un tamaño de muestra de $N = 50$, entonces, ¿no sería un desperdicio e ineficiente seguir recopilando datos? ¿No estás tentado a parar? ¿Solo un poco? Bueno, tenga en cuenta que si lo hace, su tasa de error de Tipo I en $p < .05$ simplemente se disparó al 8%. Cuando reporta $p < .05$ en su trabajo, lo que realmente está diciendo es $p < .08$. Así de malas pueden ser las consecuencias de "solo un vistazo".

Ahora considera esto. La literatura científica está llena de pruebas t, ANOVA, regresiones y pruebas de chi-cuadrado. Cuando escribí este libro no elegí estas pruebas arbitrariamente. La razón por la que estas cuatro herramientas aparecen en la mayoría de los textos de introducción a la estadística es que son las herramientas básicas de la ciencia. Ninguna de estas herramientas incluye una corrección para lidiar con el "vistazo de datos": todas asumen que no lo estás haciendo. Pero, ¿qué tan realista es esa suposición? En la vida real, ¿cuántas personas cree que "miraron" sus datos antes de que terminara el experimento y adaptaron su comportamiento posterior después de ver cómo se veían los datos? Excepto cuando el procedimiento de muestreo está fijado por una restricción externa, supongo que la respuesta es "la mayoría de la gente lo ha hecho". Si eso ha sucedido, puede inferir que los valores p informados son incorrectos. Peor aún, debido a que no sabemos qué proceso de decisión siguieron en realidad, no tenemos forma de saber cuáles deberían haber sido los valores p. No puede calcular un valor p

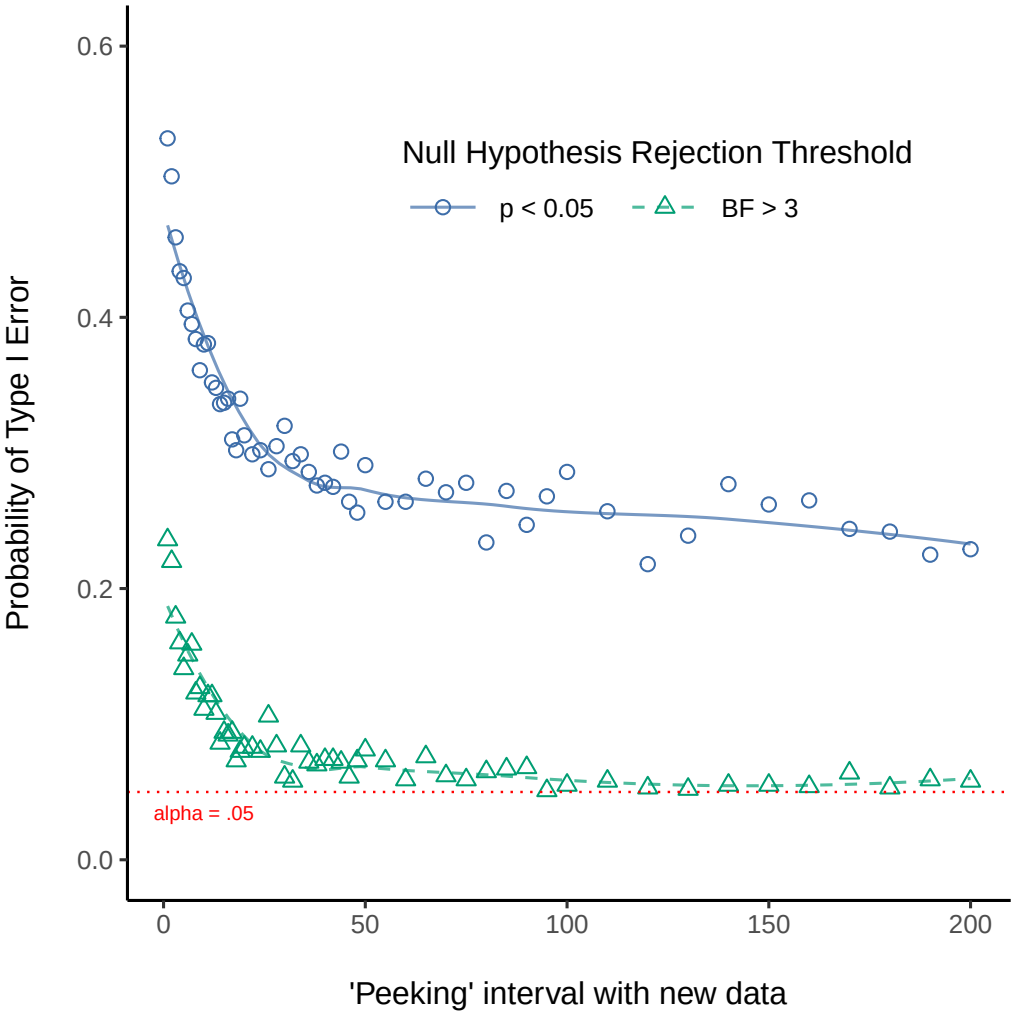


Figure 16.1: Probabilidad de error de tipo I en un experimento con N objetivo de 1000 por grupo y **mirar** en diferentes intervalos:- las cosas pueden salir muy mal si **mira** los datos y vuelve a ejecutar las pruebas como datos nuevos llega Si eres frecuentador, esto es *muy incorrecto* (círculos azules y línea continua). Si eres bayesiano, no está tan mal (triángulos verdes y línea discontinua). El nivel alfa se fijó en 0,05 (línea de puntos roja) en esta simulación.

cuando no conoce el procedimiento de toma de decisiones que utilizó el investigador. Y así, el valor p informado sigue siendo una mentira.

Teniendo en cuenta todo lo anterior, ¿cuál es el mensaje para llevar a casa? No es que los métodos bayesianos sean infalibles. Si un investigador está decidido a hacer trampa, siempre puede hacerlo. La regla de Bayes no puede impedir que la gente mienta, ni puede impedir que manipulen un experimento. Ese no es mi punto aquí. Mi punto es el mismo que planteé al comienzo del libro en la Sección 1.1: la razón por la que realizamos pruebas estadísticas es para protegernos de nosotros mismos. Y la razón por la que "mirar a escondidas los datos" es tan preocupante es que es muy tentador, incluso para los investigadores honestos. Una teoría para la inferencia estadística tiene que reconocer esto. Sí, podría tratar de defender los valores de p diciendo que es culpa del investigador por no usarlos correctamente, pero en mi opinión, eso no entiende el punto. Una teoría de la inferencia estadística que es tan completamente ingenua acerca de los humanos que ni siquiera considera la posibilidad de que el investigador pueda ver sus propios datos no es una teoría que valga la pena tener. En esencia, mi punto es este:

> *Las buenas leyes tienen su origen en la mala moral.*
> – Ambrosius Macrobius [15]

Las buenas reglas para las pruebas estadísticas deben reconocer la fragilidad humana. Ninguno de nosotros está libre de pecado. Ninguno de nosotros está más allá de la tentación. Un buen sistema de inferencia estadística debería funcionar incluso cuando lo utilizan seres humanos reales. La prueba de hipótesis nula ortodoxa no lo hace.[16]

16.4 Pruebas t bayesianas

Un tipo importante de problema de inferencia estadística discutido en este libro es [Comparación de dos medias], discutido con cierto detalle en el capítulo sobre pruebas t. Si puede recordar ese momento, recordará que hay varias versiones de la prueba t. Hablaré un poco sobre las versiones bayesianas de las pruebas t de muestras independientes y la prueba t de muestras pareadas en esta sección.

16.4.1 Prueba t de muestras independientes

El tipo más común de prueba t es la prueba t de muestras independientes, y surge cuando tiene datos como en el conjunto de datos harpo.csv que usamos en Chapter 11 en pruebas t. En este conjunto de datos, tenemos dos grupos de estudiantes, los que recibieron lecciones de Anastasia y los que tomaron sus clases con Bernadette. La pregunta que queremos responder es si hay alguna diferencia en las calificaciones que reciben estos

[15] http://www.quotationspage.com/quotes/Ambrosius%20Macrobius/%3C/a

[16] De acuerdo, solo sé que algunos frecuentadores informados leerán esto y comenzarán a quejarse de esta sección. Mira, no soy tonto. Sé absolutamente que si adopta una perspectiva de análisis secuencial puede evitar estos errores dentro del marco ortodoxo. También sé que puede diseñar estudios explícitamente con análisis intermedios en mente. Así que sí, en cierto sentido estoy atacando una versión de "hombre de paja" de los métodos ortodoxos. Sin embargo, el hombre de paja que estoy atacando es el que *usan casi todos los practicantes*. Si alguna vez llega al punto en que los métodos secuenciales se convierten en la norma entre los psicólogos experimentales y ya no estoy obligado a leer 20 ANOVA extremadamente dudosos al día, prometo que reescribiré esta sección y reduciré el vitriolo. Pero hasta que llegue ese día, mantendré mi afirmación de que los métodos predeterminados del factor de Bayes son mucho más sólidos frente a las prácticas de análisis de datos que existen en el mundo real. Los métodos ortodoxos *predeterminados* apestan, y todos lo sabemos.

dos grupos de estudiantes. En ese [capítulo] (Comparación de dos medias) sugerí que podría analizar este tipo de datos utilizando la prueba t de muestras independientes en jamovi, que nos dio los resultados en Figure 16.2. Como obtenemos un p-valor inferior a 0,05, rechazamos la hipótesis nula.

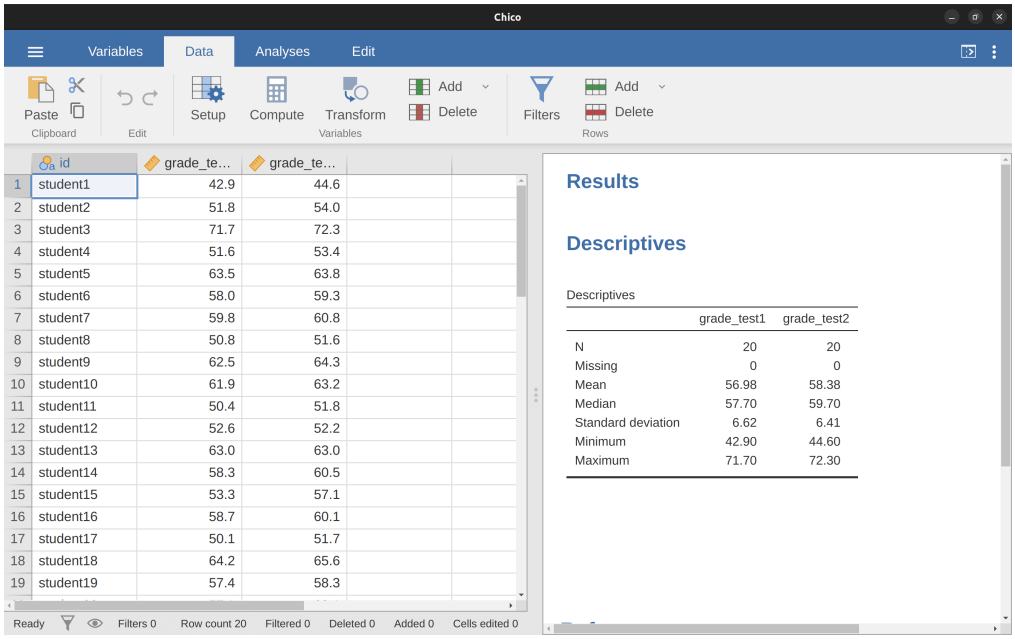


Figure 16.2: resultado de la prueba t de muestras independientes en jamovi

¿Cómo es la versión bayesiana de la prueba t? Podemos obtener el análisis del factor de Bayes seleccionando la casilla de verificación 'Factor de Bayes' en la opción 'Pruebas' y aceptando el valor predeterminado sugerido para el 'Previo'. Esto da los resultados que se muestran en la tabla en Figure 16.3. Lo que obtenemos en esta tabla es un factor estadístico de Bayes de 1,75, lo que significa que la evidencia proporcionada por estos datos es de aproximadamente 1,8:1 a favor de la hipótesis alternativa.

Antes de continuar, vale la pena resaltar la diferencia entre los resultados de la prueba ortodoxa y la bayesiana. Según la prueba ortodoxa, obtuvimos un resultado significativo, aunque apenas. Sin embargo, mucha gente aceptaría felizmente $p = .043$ como evidencia razonablemente sólida de un efecto. Por el contrario, tenga en cuenta que la prueba bayesiana ni siquiera alcanza una probabilidad de 2:1 a favor de un efecto y, en el mejor de los casos, se consideraría una evidencia muy débil. En mi experiencia, ese es un resultado bastante típico. Los métodos bayesianos generalmente requieren más evidencia antes de rechazar el valor nulo.

16.4.2 Prueba t de muestras pareadas

En la Sección 11.5, analicé el conjunto de datos chico.csv en el que las calificaciones de los estudiantes se midieron en dos pruebas, y estábamos interesados en saber si las calificaciones aumentaron de la prueba 1 a la prueba 2. Debido a que cada estudiante hizo ambas pruebas, la herramienta que usamos utilizado para analizar los datos fue una

Independent Samples T-Test

		Statistic	±%	df	p	Mean difference	SE difference		Effect Size
grade	Student's t	2.12		31.00	0.04253	5.48	2.59	Cohen's d	0.74
	Bayes factor $_{10}$	1.75	0.00						
	Welch's t	2.03		23.02	0.05361	5.48	2.69	Cohen's d	0.72

Figure 16.3: análisis de factores de Bayes junto con la prueba t de muestras independientes

prueba t de muestras pareadas. Figure 16.4 muestra la tabla de resultados jamovi para la prueba t pareada convencional junto con el análisis factorial de Bayes. En este punto, espero que puedas leer este resultado sin ninguna dificultad. Los datos proporcionan evidencia de alrededor de 6000:1 a favor de la alternativa. ¡Probablemente podríamos rechazar el nulo con cierta confianza!

Paired Samples T-Test

								95% Confidence Interval			
			Statistic	±%	df	p	Mean difference	SE difference	Lower	Upper		Effect Size
grade_test2	grade_test1	Student's t	6.48		19.00	< .00001	1.40	0.22	0.95	1.86	Cohen's d	1.45
		Bayes factor $_{10}$	5991.58	0.00								

Figure 16.4: Muestras pareadas T-Test y Bayes Factor dan como resultado jamovi

16.5 Resumen

La primera mitad de este capítulo se centró principalmente en los fundamentos teóricos de las estadísticas bayesianas. Presenté las matemáticas de cómo funciona la inferencia bayesiana en la sección sobre Razonamiento probabilístico por agentes racionales, y brindé una descripción general muy básica de las pruebas de hipótesis bayesianas]. Finalmente, dediqué algo de espacio a hablar sobre por qué creo que [vale la pena usar los métodos bayesianos] (¿Por qué ser bayesiano?).

Luego di un ejemplo práctico, con pruebas t bayesianas. Si está interesado en aprender más sobre el enfoque bayesiano, hay muchos buenos libros que podría consultar. El libro de John Kruschke *Doing Bayesian Data Analysis* es un muy buen lugar para comenzar (Kruschke, 2011) y es una buena combinación de teoría y práctica. Su enfoque es un poco diferente al enfoque del "factor de Bayes" que he discutido aquí, por lo que no cubrirá el mismo terreno. Si es psicólogo cognitivo, puede consultar Lee & Wagenmakers (2014). Elegí estos dos porque creo que son especialmente útiles para las personas en mi disciplina, pero hay muchos libros buenos, ¡así que mira a tu alrededor!

Epílogo

> *"Empieza por el principio", dijo el Rey muy gravemente, "y continúa hasta que llegues al final: luego detente"* – Lewis Carroll

Se siente algo extraño estar escribiendo este capítulo, y más que un poco inapropiado. Un epílogo es lo que escribes cuando terminas un libro, y este libro realmente no está terminado. Aún faltan muchas cosas en este libro. Todavía no tiene un índice. Faltan *muchas* referencias. No hay ejercicios de "hágalo usted mismo". Y en general, siento que hay muchas cosas que están mal con la presentación, organización y contenido de este libro. Dado todo eso, no quiero tratar de escribir un epílogo "adecuado". Todavía no he terminado de escribir el contenido sustantivo, por lo que no tiene sentido tratar de reunirlo todo. Pero esta versión del libro se pondrá en línea para que la usen los estudiantes, y es posible que también compre una copia impresa, por lo que quiero darle al menos una apariencia de cierre. Así que vamos a darle una oportunidad, ¿de acuerdo?

Las estadísticas no descubiertas

Primero, voy a hablar un poco sobre algunos de los contenidos que desearía haber tenido la oportunidad de incluir en esta versión del libro, solo para que puedan tener una idea de qué otras ideas existen en el mundo de las estadísticas. Creo que esto sería importante incluso si este libro se estuviera acercando a un producto final. Una cosa que los estudiantes a menudo no se dan cuenta es que sus clases de introducción a la estadística son solo eso, una introducción. Si desea salir al mundo más amplio y realizar análisis de datos reales, debe aprender muchas herramientas nuevas que amplían el contenido de sus conferencias de pregrado en todo tipo de formas diferentes. No asuma que algo no se puede hacer solo porque no se cubrió en la licenciatura. No asuma que algo es lo correcto solo porque se cubrió en una clase de pregrado. Para evitar que seas víctima de esa trampa, creo que es útil ofrecerte una descripción general de algunas de las otras ideas que existen.

Omisiones dentro de los temas tratados

Incluso dentro de los temas que he cubierto en el libro, hay muchas omisiones que me gustaría corregir en una versión futura del libro. Solo apegado a cosas que son puramente estadísticas (en lugar de cosas asociadas con jamovi), la siguiente es una lista representativa pero no exhaustiva de temas que me gustaría ampliar en algún momento:

- **Otros tipos de correlaciones.** En [Correlación y regresión] hablé de dos tipos de

correlación: Pearson y Spearman. Ambos métodos de evaluación de la correlación son aplicables al caso en el que tiene dos variables continuas y desea evaluar la relación entre ellas. ¿Qué pasa con el caso en que sus variables son ambas de escala nominal? ¿O cuando uno es de escala nominal y el otro es continuo? En realidad, existen métodos para calcular las correlaciones en tales casos (p. ej., la correlación policórica), y sería bueno que se incluyeran.

- **Más detalles sobre los tamaños de los efectos.** En general, creo que el tratamiento de los tamaños de los efectos a lo largo del libro es un poco más superficial de lo que debería ser. En casi todos los casos, he tendido a elegir solo una medida del tamaño del efecto (generalmente la más popular) y describirla. Sin embargo, para casi todas las pruebas y modelos hay múltiples formas de pensar sobre el tamaño del efecto, y me gustaría entrar en más detalles en el futuro.
- **Tratar con suposiciones violadas.** En varias partes del libro he hablado sobre algunas cosas que puede hacer cuando descubre que se violan las suposiciones de su prueba (o modelo), pero creo que debería para decir más sobre esto. En particular, creo que hubiera sido bueno hablar con mucho más detalle sobre cómo puedes transformar variables para solucionar problemas. Hablé un poco sobre esto [asuntos pragmáticos, pero creo que la discusión no es lo suficientemente detallada.
- **Términos de interacción para regresión.** En ANOVA factorial hablé sobre el hecho de que puedes tener términos de interacción en un ANOVA, y también señalé que ANOVA puede interpretarse como una especie de modelo de regresión lineal. Sin embargo, cuando hablé de regresión en [Correlación y regresión] no mencioné interacciones en absoluto. Sin embargo, nada le impide incluir términos de interacción en un modelo de regresión. Es un poco más complicado averiguar qué significa realmente una "interacción" cuando se habla de la interacción entre dos predictores continuos, y se puede hacer de más de una manera. Aun así, me hubiera gustado hablar un poco sobre esto.
- **Método de comparación planificada.** Como mencioné en ANOVA factorial, no siempre es apropiado usar una corrección post hoc como el HSD de Tukey cuando se hace un ANOVA, especialmente cuando tenía una idea muy clara (y limitada) conjunto de comparaciones que le preocupaban de antemano. Me gustaría hablar más sobre esto en el futuro.
- **Métodos de comparación múltiple.** Incluso dentro del contexto de hablar sobre pruebas post hoc y comparaciones múltiples, me hubiera gustado hablar sobre los métodos con más detalle y hablar sobre qué otros métodos existen además de las pocas opciones que mencioné.

Faltan modelos estadísticos en el libro

La estadística es un campo enorme. Las herramientas principales que he descrito en este libro (pruebas de chi-cuadrado, pruebas t, regresión y ANOVA) son herramientas básicas que se usan ampliamente en el análisis de datos cotidianos y forman el núcleo de la mayoría de los libros de introducción a las estadísticas. Sin embargo, hay muchas otras herramientas por ahí. Hay tantas situaciones de análisis de datos que estas herramientas no cubren, y sería genial darle una idea de cuánto más hay, por ejemplo:

- **Regresión no lineal.** Cuando discutimos la regresión en el Capítulo **12**, vimos que la regresión asume que la relación entre los predictores y los resultados es

lineal. Por otro lado, cuando hablamos sobre el problema más simple de la correlación en el Capítulo **4**, vimos que existen herramientas (p. ej., correlaciones de Spearman) que pueden evaluar relaciones no lineales entre variables. Hay una serie de herramientas en estadística que se pueden usar para hacer una regresión no lineal. Por ejemplo, algunos modelos de regresión no lineal suponen que la relación entre predictores y resultados es monotónica (p. ej., regresión isotónica), mientras que otros suponen que es suave pero no necesariamente monótona (p. ej., regresión de Lowess), mientras que otros suponen que la relación es de una forma conocida que pasa a ser no lineal (por ejemplo, regresión polinomial).

- **Regresión logística.** Otra variación de la regresión ocurre cuando la variable de resultado es binaria, pero los predictores son continuos. Por ejemplo, suponga que está investigando las redes sociales y quiere saber si es posible predecir si alguien está en Twitter o no en función de sus ingresos, su edad y una variedad de otras variables. Este es básicamente un modelo de regresión, pero no puede usar la regresión lineal regular porque la variable de resultado es binaria (o está en Twitter o no lo está). Debido a que la variable de resultado es binaria, no hay forma de que los residuos puedan distribuirse normalmente. Hay una serie de herramientas que los estadísticos pueden aplicar a esta situación, la más destacada de las cuales es la regresión logística.

- **El modelo lineal general (GLM).** El GLM es en realidad una familia de modelos que incluye regresión logística, regresión lineal, (algunas) regresiones no lineales, ANOVA y muchos otros. La idea básica en el GLM es esencialmente la misma idea que sustenta los modelos lineales, pero permite la idea de que sus datos podrían no estar distribuidos normalmente y permite relaciones no lineales entre los predictores y los resultados. Hay muchos análisis muy útiles que puede ejecutar que se encuentran dentro del GLM, por lo que es muy útil conocerlos.

- **Análisis de supervivencia.** En Una breve introducción al diseño de investigación hablé sobre la "deserción diferencial", la tendencia de las personas a abandonar el estudio de manera no aleatoria. En aquel entonces, estaba hablando de ello como una posible preocupación metodológica, pero hay muchas situaciones en las que el desgaste diferencial es realmente lo que le interesa. Suponga, por ejemplo, que le interesa saber cuánto tiempo la gente jugar diferentes tipos de juegos de computadora en una sola sesión. ¿La gente tiende a jugar juegos RTS (estrategia en tiempo real) durante períodos más largos que juegos FPS (disparos en primera persona)? Puede diseñar su estudio de esta manera. Las personas entran al laboratorio y pueden jugar durante el tiempo que deseen. Una vez que terminan, registras el tiempo que pasaron jugando. Sin embargo, debido a restricciones éticas, supongamos que no puedes permitir que sigan jugando más de dos horas. Muchas personas dejarán de jugar antes del límite de dos horas, por lo que sabrás exactamente cuánto tiempo jugaron. Pero algunas personas se toparán con el límite de dos horas, por lo que no sabes cuánto tiempo habrían seguido jugando si hubieras podido continuar con el estudio. Como consecuencia, sus datos se censuran sistemáticamente: se pierde todos los tiempos muy largos. ¿Cómo analiza estos datos con sensatez? Este es el problema que resuelve el análisis de supervivencia. Está diseñado específicamente para manejar esta situación, en la que se pierde sistemáticamente un "lado" de los datos porque el estudio finalizó. Se usa mucho en la investigación de la salud y, en ese contexto, a menudo se usa literalmente para analizar la supervivencia. Por ejemplo, puede estar rastreando a

personas con un tipo particular de cáncer, algunas que han recibido el tratamiento A y otras que han recibido el tratamiento B, pero solo tiene fondos para rastrearlas durante 5 años. Al final del período de estudio, algunas personas están vivas, otras no. En este contexto, el análisis de supervivencia es útil para determinar qué tratamiento es más efectivo e informarle sobre el riesgo de muerte que enfrentan las personas con el tiempo.

- **Modelos mixtos.** El ANOVA de medidas repetidas se usa a menudo en situaciones en las que tiene observaciones agrupadas dentro de unidades experimentales. Un buen ejemplo de esto es cuando realiza un seguimiento de personas individuales en múltiples puntos de tiempo. Digamos que estás rastreando la felicidad a lo largo del tiempo, para dos personas. La felicidad de Aaron comienza en 10, luego baja a 8 y luego a 6. La felicidad de Belinda comienza en 6, luego sube a 8 y luego a 10. Estas dos personas tienen el mismo nivel "general" de felicidad (el promedio en todo el grupo). tres puntos de tiempo es 8), por lo que un análisis ANOVA de medidas repetidas trataría a Aaron y Belinda de la misma manera. Pero eso está claramente mal. La felicidad de Aaron disminuye, mientras que la de Belinda aumenta. Si desea analizar de manera óptima los datos de un experimento en el que las personas pueden cambiar con el tiempo, entonces necesita una herramienta más poderosa que ANOVA de medidas repetidas. Las herramientas que la gente usa para resolver este problema se denominan modelos "mixtos", porque están diseñados para aprender sobre unidades experimentales individuales (por ejemplo, la felicidad de personas individuales a lo largo del tiempo), así como efectos generales (por ejemplo, el efecto del dinero en la felicidad a lo largo del tiempo).). ANOVA de medidas repetidas es quizás el ejemplo más simple de un modelo mixto, pero hay mucho que puede hacer con modelos mixtos que no puede hacer con ANOVA de medidas repetidas.

- **Escalamiento multidimensional.** El análisis factorial es un ejemplo de un modelo de "aprendizaje no supervisado". Lo que esto significa es que, a diferencia de la mayoría de las herramientas de "aprendizaje supervisado" que he mencionado, no puede dividir sus variables en predictores y resultados. La regresión es aprendizaje supervisado, mientras que el análisis factorial es aprendizaje no supervisado. Sin embargo, no es el único tipo de modelo de aprendizaje no supervisado. Por ejemplo, en el análisis factorial uno se ocupa del análisis de correlaciones entre variables. Sin embargo, hay muchas situaciones en las que realmente te interesa analizar las similitudes o diferencias entre objetos, elementos o personas. Hay una serie de herramientas que puede utilizar en esta situación, la más conocida de las cuales es el escalado multidimensional (MDS). En MDS, la idea es encontrar una representación "geométrica" de sus elementos. Cada elemento se "traza" como un punto en algún espacio, y la distancia entre dos puntos es una medida de cuán diferentes son esos elementos.

- **Clustering.** Otro ejemplo de un modelo de aprendizaje no supervisado es el agrupamiento (también conocido como clasificación), en el que desea organizar todos sus elementos en grupos significativos, de modo que los elementos similares se asignen a los mismos grupos. Gran parte de la agrupación no está supervisada, lo que significa que no sabe nada sobre cuáles son los grupos, solo tiene que adivinar. Existen otras situaciones de "agrupamiento supervisado" en las que es necesario predecir la pertenencia a grupos en función de otras variables, y esas pertenencias a grupos son en realidad observables. La regresión logística es un

buen ejemplo de una herramienta que funciona de esta manera. Sin embargo, cuando en realidad no conoce las membresías del grupo, debe usar diferentes herramientas (p. ej., agrupación en clústeres k-means). Incluso hay situaciones en las que desea hacer algo llamado "agrupamiento en clústeres semisupervisado", en el que conoce la membresía del grupo para algunos elementos pero no para otros. Como probablemente pueda adivinar, la agrupación en clústeres es un tema bastante amplio y algo bastante útil para conocer.

- **Modelos causales.** Una cosa de la que no he hablado mucho en este libro es cómo puede usar modelos estadísticos para aprender sobre las relaciones causales entre variables. Por ejemplo, considere las siguientes tres variables que podrían ser de interés al pensar en cómo murió alguien en un pelotón de fusilamiento. Podríamos querer medir si se dio o no una orden de ejecución (variable A), si un tirador disparó o no su arma (variable B) y si la persona recibió o no una bala (variable C). Estas tres variables están todas correlacionadas entre sí (por ejemplo, existe una correlación entre las armas que se disparan y las personas que reciben balas), pero en realidad queremos hacer afirmaciones más sólidas sobre ellas que simplemente hablar de correlaciones. Queremos hablar de causalidad. Queremos poder decir que la orden de ejecución (A) hace que el tirador dispare (B) lo que hace que alguien reciba un disparo (C). Podemos expresar esto mediante una notación de flecha dirigida: lo escribimos como $A \rightarrow B \rightarrow C$. Esta "cadena causal" es una explicación fundamentalmente diferente para los eventos que aquella en la que el tirador dispara primero, lo que provoca el disparo $B \rightarrow C$, y luego hace que el verdugo emita "retroactivamente" la orden de ejecución, $B \rightarrow$. Este modelo de "efecto común" dice que tanto A como C son causados por B. Puede ver por qué son diferentes. En el primer modelo causal, si hubiésemos conseguido que el verdugo no diera la orden (interviniendo para cambiar A), entonces no se habría producido ningún disparo. En el segundo modelo, el tiro habría ocurrido de cualquier manera porque el tirador no estaba siguiendo la orden de ejecución. Existe una gran literatura en estadística sobre cómo tratar de comprender las relaciones causales entre las variables, y existen varias herramientas diferentes para ayudarlo a probar diferentes historias causales sobre sus datos. La más utilizada de estas herramientas (al menos en psicología) es el modelo de ecuaciones estructurales (SEM), y en algún momento me gustaría ampliar el libro para hablar de ello.

Por supuesto, incluso esta lista está incompleta. No he mencionado el análisis de series de tiempo, la teoría de la respuesta al ítem, el análisis de la cesta de la compra, los árboles de clasificación y regresión, o cualquiera de una amplia gama de otros temas. Sin embargo, la lista que he dado anteriormente es esencialmente mi lista de deseos para este libro. Claro, duplicaría la longitud del libro, pero significaría que el alcance se ha vuelto lo suficientemente amplio como para cubrir la mayoría de las cosas que los investigadores de psicología aplicada necesitarían usar.

Otras formas de hacer inferencias

Un sentido diferente en el que este libro está incompleto es que se centra bastante en una visión muy estrecha y anticuada de cómo se debe hacer la estadística inferencial. En Estimación de cantidades desconocidas de una muestra hablé un poco sobre la idea de estimadores imparciales, distribuciones de muestreo, etc. En Prueba de hipótesis hablé sobre la teoría de la prueba de significancia de la hipótesis nula y los valores p. Estas

ideas existen desde principios del siglo XX, y las herramientas de las que he hablado en el libro se basan en gran medida en las ideas teóricas de esa época. Me he sentido obligado a ceñirme a esos temas porque la gran mayoría del análisis de datos en la ciencia también depende de esas ideas. Sin embargo, la teoría de la estadística no se limita a esos temas y, aunque todo el mundo debería conocerlos debido a su importancia práctica, en muchos aspectos esas ideas no representan las mejores prácticas para el análisis de datos contemporáneo. Una de las cosas con las que estoy especialmente contento es que he podido ir un poco más allá. [Estadísticas bayesianas] ahora presenta la perspectiva bayesiana con una cantidad razonable de detalles, pero el libro en general todavía está muy inclinado hacia la ortodoxia frecuentista. Además, hay una serie de otros enfoques de la inferencia que vale la pena mencionar:

- Arranque. A lo largo del libro, cada vez que introduje una prueba de hipótesis, tuve una fuerte tendencia a hacer afirmaciones como "la distribución de muestreo para BLAH es una distribución t" o algo así. En algunos casos, en realidad he intentado justificar esta afirmación. Por ejemplo, cuando hablé de las pruebas χ^2 en Análisis de datos categóricos hice referencia a la relación conocida entre las distribuciones normales y las distribuciones χ^2 (ver [Introducción a la probabilidad) para explicar cómo terminamos suponiendo que la distribución muestral del estadístico de bondad de ajuste es χ^2 . Sin embargo, también es cierto que muchas de estas distribuciones de muestreo son, bueno, incorrectas. La prueba χ^2 es un buen ejemplo. Se basa en una suposición sobre la distribución de sus datos, ¡una suposición que se sabe que es incorrecta para tamaños de muestra pequeños! A principios del siglo XX, no había mucho que pudieras hacer sobre esta situación. Los estadísticos habían desarrollado resultados matemáticos que decían que "bajo suposiciones BLAH sobre los datos, la distribución de muestreo es aproximadamente BLAH", y eso era lo mejor que podía hacer. Muchas veces ni siquiera tenían eso. Hay muchas situaciones de análisis de datos para las que nadie ha encontrado una solución matemática para las distribuciones de muestreo que necesita. Y así hasta finales del siglo XX, las pruebas correspondientes no existían o no funcionaban. Sin embargo, las computadoras han cambiado todo eso ahora. Hay muchos trucos sofisticados y algunos no tan sofisticados que puedes usar para evitarlo. El más simple de estos es el arranque, y en su forma más simple es increíblemente simple. Lo que hace es simular los resultados de su experimento muchas veces, bajo las suposiciones gemelas de que (a) la hipótesis nula es verdadera y (b) la distribución de la población desconocida en realidad se ve bastante similar a sus datos sin procesar. En otras palabras, en lugar de suponer que los datos están (por ejemplo) distribuidos normalmente, simplemente suponga que la población tiene el mismo aspecto que su muestra y luego use computadoras para simular la distribución de muestreo para su estadística de prueba si esa suposición se cumple. A pesar de basarse en una suposición un tanto dudosa (es decir, ¡la distribución de la población es la misma que la muestra!), el bootstrapping es un método rápido y fácil que funciona notablemente bien en la práctica para muchos problemas de análisis de datos.

- Validación cruzada. Una pregunta que aparece en mis clases de estadística de vez en cuando, generalmente por parte de un estudiante que intenta ser provocativo, es "¿Por qué nos preocupamos por las estadísticas inferenciales? ¿Por qué no simplemente describir su muestra?" La respuesta a la pregunta suele ser algo como esto: "Debido a que nuestro verdadero interés como científicos no es la muestra

específica que hemos observado en el pasado, queremos hacer predicciones sobre los datos que podríamos observar en el futuro". Muchos de los problemas en la inferencia estadística surgen debido al hecho de que siempre esperamos que el futuro sea similar pero un poco diferente al pasado. O, de manera más general, los datos nuevos no serán exactamente iguales a los datos antiguos. Lo que hacemos, en muchas situaciones, es tratar de derivar reglas matemáticas que nos ayuden a sacar las inferencias que tienen más probabilidades de ser correctas para los datos nuevos, en lugar de elegir las declaraciones que mejor describen los datos antiguos. Por ejemplo, dados dos modelos A y B, y un conjunto de datos X que recopilaste hoy, trata de elegir el modelo que describa mejor un nuevo conjunto de datos Y que recopilarás mañana. A veces conviene simular el proceso, y eso es lo que hace la validación cruzada. Lo que hace es dividir su conjunto de datos en dos subconjuntos, $X1$ y $X2$. Utilice el subconjunto $X1$ para entrenar el modelo (por ejemplo, estime los coeficientes de regresión, digamos), pero luego evalúe el rendimiento del modelo en el otro $X2$. Esto le da una medida de qué tan bien se generaliza el modelo de un conjunto de datos antiguo a uno nuevo y, a menudo, es una mejor medida de qué tan bueno es su modelo que si simplemente lo ajusta al conjunto de datos completo X.

- Estadísticas robustas. La vida es desordenada y nada funciona realmente como se supone que debe hacerlo. Esto es tan cierto para las estadísticas como para cualquier otra cosa, y cuando tratamos de analizar datos, a menudo nos encontramos con todo tipo de problemas en los que los datos son más confusos de lo que se supone que deben ser. Las variables que se supone que se distribuyen normalmente no se distribuyen normalmente, las relaciones que se supone que son lineales no son realmente lineales, y algunas de las observaciones en su conjunto de datos son casi con seguridad basura (es decir, no miden lo que se supone que deben medir).). Todo este desorden se ignora en la mayor parte de la teoría estadística que desarrollé en este libro. Sin embargo, ignorar un problema no siempre lo resuelve. A veces, está bien ignorar el desorden, porque algunos tipos de herramientas estadísticas son "robustas", es decir, si los datos no satisfacen sus suposiciones teóricas, aún así funcionan bastante bien. Otros tipos de herramientas estadísticas no son sólidas, e incluso pequeñas desviaciones de los supuestos teóricos hacen que se rompan. Las estadísticas robustas son una rama de las estadísticas que se ocupa de esta pregunta y hablan de cosas como el "punto de ruptura" de una estadística. Es decir, ¿qué tan desordenados deben ser sus datos antes de que no se pueda confiar en la estadística? Toqué esto en algunos lugares. La media no es un estimador robusto de la tendencia central de una variable, pero la mediana sí lo es. Por ejemplo, supón que te dijera que las edades de mis cinco mejores amigos son 34, 39, 31, 43 y 4003 años. ¿Qué edad crees que tienen en promedio? Es decir, ¿qué significa aquí la verdadera población? Si usa la media muestral como estimador de la media poblacional, obtiene una respuesta de 830 años. Si usa la mediana muestral como estimador de la media poblacional, obtiene una respuesta de 39 años. Tenga en cuenta que, a pesar de que "técnicamente" está haciendo lo incorrecto en el segundo caso (¡usando la mediana para estimar la media!), en realidad está obteniendo una mejor respuesta. El problema aquí es que una de las observaciones es claramente, obviamente, una mentira. No tengo un amigo de 4003 años. Probablemente sea un error tipográfico, probablemente quise escribir 43. Pero, ¿y si hubiera escrito 53 en lugar de 43 o 34 en lugar de 43? ¿Podría estar seguro de si esto fue un error tipográfico o no? A veces, los

errores en los datos son sutiles, por lo que no puede detectarlos simplemente observando la muestra, pero siguen siendo errores que contaminan sus datos y aún afectan sus conclusiones. Las estadísticas sólidas se ocupan de cómo puede hacer inferencias seguras, incluso cuando se enfrenta a una contaminación de la que no tiene conocimiento. Es algo muy bueno.

Temas varios

- Suponga que está realizando una encuesta y le interesa el ejercicio y el peso. Envías datos a cuatro personas. Adam dice que hace mucho ejercicio y no tiene sobrepeso. Briony dice que hace mucho ejercicio y no tiene sobrepeso. Carol dice que no hace ejercicio y tiene sobrepeso. Tim dice que no hace ejercicio y se niega a responder la pregunta sobre su peso. Elaine no devuelve la encuesta. Ahora tiene un problema de falta de datos. Falta una encuesta completa y falta una pregunta de otra, ¿Qué haces al respecto? Ignorar los datos que faltan no es, en general, algo seguro. Pensemos en la encuesta de Tim aquí. En primer lugar, observe que, sobre la base de sus otras respuestas, parece ser más similar a Carol (ninguno de nosotros hace ejercicio) que a Adam o Briony. Entonces, si te vieras obligado a adivinar su peso, dirías que está más cerca de ella que de ellos. Tal vez haría alguna corrección por el hecho de que Adam y Tim son hombres y Briony y Carol son mujeres. El nombre estadístico para este tipo de adivinanzas es "imputación". Hacer la imputación de manera segura es difícil, pero es importante, especialmente cuando los datos que faltan se pierden de manera sistemática. Debido al hecho de que a las personas con sobrepeso a menudo se las presiona para que se sientan mal por su peso (a menudo gracias a campañas de salud pública), en realidad tenemos motivos para sospechar que las personas que no responden tienen más probabilidades de tener sobrepeso que las personas que sí lo hacen. responde Imputar un peso a Tim significa que el número de personas con sobrepeso en la muestra probablemente aumentará de 1 de 3 (si ignoramos a Tim) a 2 de 4 (si imputamos el peso de Tim). Claramente esto importa. Pero hacerlo con sensatez es más complicado de lo que parece. Anteriormente, sugerí que deberías tratar a Tim como Carol, ya que dieron la misma respuesta a la pregunta del ejercicio. Pero eso no es del todo correcto. Hay una diferencia sistemática entre ellos. Ella respondió la pregunta y Tim no. Dadas las presiones sociales que enfrentan las personas con sobrepeso, ¿no es probable que Tim tenga *más* sobrepeso que Carol? Y, por supuesto, esto sigue ignorando el hecho de que no es sensato imputar un peso *único* a Tim, como si realmente supieras su peso. En cambio, lo que debe hacer es imputar un rango de conjeturas plausibles (lo que se conoce como imputación múltiple), para capturar el hecho de que está más inseguro sobre el peso de Tim que sobre el de Carol. Y no comencemos con el problema planteado por el hecho de que Elaine no envió la encuesta. Como probablemente pueda adivinar, lidiar con los datos faltantes es un tema cada vez más importante. De hecho, me han dicho que muchas revistas en algunos campos no aceptarán estudios en los que falten datos a menos que se siga algún tipo de esquema de imputación múltiple sensato.

- Análisis de potencia. En Prueba de hipótesis hablé del concepto de potencia (es decir, qué tan probable es que pueda detectar un efecto si realmente existe) y me refiero al análisis de potencia, una colección de herramientas que son útiles para evaluar la potencia de su estudio. posee. El análisis de potencia puede ser útil para planificar un estudio (p. ej., averiguar qué tamaño de muestra es probable

que necesite), pero también cumple una función útil en el análisis de datos que ya recopiló. Por ejemplo, suponga que obtiene un resultado significativo y tiene una estimación del tamaño del efecto. Puede usar esta información para estimar cuánta potencia tenía realmente su estudio. Esto es bastante útil, especialmente si el tamaño de su efecto no es grande. Por ejemplo, suponga que rechaza la hipótesis nula en $p < .05$, pero usa el análisis de potencia para determinar que su potencia estimada fue solo .08. El resultado significativo significa que, si la hipótesis nula fuera cierta, había un 5% de posibilidades de obtener datos como este. Pero la potencia baja significa que, incluso si la hipótesis nula es falsa y el tamaño del efecto es realmente tan pequeño como parece, solo hay un 8 % de posibilidades de obtener datos como los que obtuvo usted. ¡Esto sugiere que debe ser bastante cauteloso, porque la suerte parece haber jugado un papel importante en sus resultados, de una forma u otra!

- Análisis de datos utilizando modelos inspirados en la teoría. En algunas partes de este libro he mencionado los datos de tiempo de respuesta (RT), donde se registra cuánto tiempo le toma a alguien hacer algo (por ejemplo, tomar una decisión simple). He mencionado que los datos de RT son casi invariablemente no normales y positivamente sesgados. Además, existe una cosa conocida como compensación entre velocidad y precisión: si intenta tomar decisiones demasiado rápido (RT bajo), es probable que tome decisiones más pobres (menor precisión). Entonces, si mide tanto la precisión de las decisiones de un participante como su RT, probablemente encontrará que la velocidad y la precisión están relacionadas. Hay más en la historia que esto, por supuesto, porque algunas personas toman mejores decisiones que otras, independientemente de lo rápido que vayan. Además, la velocidad depende tanto de los procesos cognitivos (es decir, el tiempo dedicado a pensar) como de los fisiológicos (p. ej., qué tan rápido puede mover los músculos). Está empezando a parecer que analizar estos datos será un proceso complicado. Y de hecho lo es, pero una de las cosas que encuentras cuando profundizas en la literatura psicológica es que ya existen modelos matemáticos (llamados "modelos de muestreo secuencial") que describen cómo las personas toman decisiones simples, y estos modelos toman en cuenta un muchos de los factores que mencioné anteriormente. No encontrará ninguno de estos modelos inspirados teóricamente en un libro de texto estándar de estadística. Los libros de texto de estadísticas estándar describen herramientas estándar, herramientas que podrían aplicarse significativamente en muchas disciplinas diferentes, no solo en psicología. ANOVA es un ejemplo de una herramienta estándar que es tan aplicable a la psicología como a la farmacología. Los modelos de muestreo secuencial no lo son, son más o menos específicos de la psicología. Esto no los convierte en herramientas menos poderosas. De hecho, si está analizando datos en los que las personas tienen que tomar decisiones rápidamente, debería usar modelos de muestreo secuencial para analizar los datos. Usar ANOVA o regresión o lo que sea no funcionará tan bien, porque los supuestos teóricos que los sustentan no coinciden bien con sus datos. Por el contrario, los modelos de muestreo secuencial se diseñaron explícitamente para analizar este tipo específico de datos, y sus suposiciones teóricas se ajustan muy bien a los datos.

Aprendiendo los conceptos básicos y aprendiéndolos en jamovi

Bueno, esa era una lista larga. E incluso esa lista está enormemente incompleta. Realmente hay muchas grandes ideas en estadística que no he cubierto en este libro. Puede parecer bastante deprimente terminar un libro de texto de casi 500 páginas solo para que te digan que esto es solo el comienzo, especialmente cuando comienzas a sospechar que la mitad de lo que te han enseñado está mal. Por ejemplo, hay mucha gente en el campo que argumentaría fuertemente en contra del uso del modelo ANOVA clásico, ¡pero le he dedicado dos capítulos completos! El ANOVA estándar puede ser atacado desde una perspectiva bayesiana, o desde una perspectiva estadística robusta, o incluso desde una perspectiva de "simplemente está mal" (la gente usa con mucha frecuencia ANOVA cuando en realidad debería estar usando modelos mixtos). Entonces, ¿por qué aprenderlo en absoluto?

Como yo lo veo, hay dos argumentos clave. En primer lugar, está el argumento del pragmatismo puro. Correcta o incorrectamente, ANOVA es ampliamente utilizado. Si desea comprender la literatura científica, debe comprender ANOVA. Y en segundo lugar, está el argumento del "conocimiento incremental". De la misma manera que fue útil haber visto ANOVA unidireccional antes de intentar aprender ANOVA factorial, comprender ANOVA es útil para comprender herramientas más avanzadas, porque muchas de esas herramientas amplían o modifican la configuración básica de ANOVA de alguna manera. Por ejemplo, aunque los modelos mixtos son mucho más útiles que ANOVA y la regresión, nunca he oído hablar de nadie que haya aprendido cómo funcionan los modelos mixtos sin haber trabajado primero con ANOVA y la regresión. Tienes que aprender a gatear antes de poder escalar una montaña.

En realidad, quiero llevar este punto un poco más lejos. Una cosa que he hecho mucho en este libro es hablar sobre los fundamentos. Pasé mucho tiempo en la teoría de la probabilidad. Hablé sobre la teoría de la estimación y las pruebas de hipótesis con más detalle del necesario. ¿Por qué hice todo esto? Mirando hacia atrás, podría preguntarse si realmente necesitaba pasar todo ese tiempo hablando sobre qué es una distribución de probabilidad, o por qué había incluso una sección sobre densidad de probabilidad. Si el objetivo del libro era enseñarle cómo ejecutar una prueba t o un ANOVA, ¿era todo eso realmente necesario? ¿Fue todo esto una gran pérdida de tiempo para todos?

La respuesta, espero que esté de acuerdo, es no. El objetivo de una estadística introductoria no es enseñar ANOVA. No es para enseñar pruebas t, regresiones, histogramas o valores p. El objetivo es iniciarlo en el camino para convertirse en un analista de datos calificado. Y para que usted se convierta en un analista de datos capacitado, debe poder hacer más que ANOVA, más que pruebas t, regresiones e histogramas. Tienes que ser capaz de pensar correctamente acerca de los datos. Debe poder aprender los modelos estadísticos más avanzados de los que hablé en la última sección y comprender la teoría en la que se basan. Y necesita tener acceso a un software que le permita usar esas herramientas avanzadas. Y aquí es donde, al menos en mi opinión, todo el tiempo extra que he dedicado a los fundamentos vale la pena. Si comprende la teoría de la probabilidad, le resultará mucho más fácil pasar de los análisis frecuentistas a los bayesianos.

En resumen, creo que la gran recompensa por aprender estadística de esta manera es la extensibilidad. Para un libro que solo cubre los conceptos básicos del análisis de datos, este libro tiene una sobrecarga enorme en términos de aprendizaje de la teoría de la

probabilidad, etc. Hay muchas otras cosas que te empujan a aprender además de los análisis específicos que cubre el libro. Entonces, si su objetivo había sido aprender a ejecutar un ANOVA en el mínimo tiempo posible, este libro no era una buena opción. Pero como digo, no creo que ese sea tu objetivo. Creo que quieres aprender a hacer análisis de datos. Y si ese es realmente su objetivo, querrá asegurarse de que las habilidades que aprenda en su clase introductoria de estadísticas sean extensibles de forma natural y limpia a los modelos más complicados que necesita en el análisis de datos del mundo real. Quiere asegurarse de aprender a usar las mismas herramientas que usan los analistas de datos reales, para que pueda aprender a hacer lo que ellos hacen. Y sí, está bien, eres un principiante en este momento (o lo eras cuando comenzaste este libro), pero eso no significa que debas contarte una historia tonta, una historia en la que no te cuente sobre densidad de probabilidad, o una historia donde no les cuento sobre la pesadilla que es el ANOVA factorial con diseños desbalanceados. Y eso no significa que deban darle juguetes para bebés en lugar de herramientas de análisis de datos adecuadas. Los principiantes no son tontos, simplemente les falta conocimiento. Lo que necesita es que no se le oculten las complejidades del análisis de datos del mundo real. Lo que necesita son las habilidades y herramientas que le permitirán manejar esas complejidades cuando inevitablemente lo embosquen en el mundo real.

Y lo que espero es que este libro, o el libro terminado en el que se convertirá algún día, pueda ayudarlo con eso.

Nota del autor: lo mencioné antes, pero lo mencionaré rápidamente de nuevo. La lista de referencias del libro es terriblemente incompleta. Por favor, no asuma que estas son las únicas fuentes en las que he confiado. La versión final de este libro tendrá muchas más referencias. Y si ve algo que suena inteligente en este libro que no parece tener una referencia, puedo prometerle absolutamente que la idea fue de otra persona. Este es un libro de texto introductorio: ninguna de las ideas es original. Me haré responsable de todos los errores, pero no puedo atribuirme nada de lo bueno. Todo lo inteligente de este libro provino de otra persona, y todos merecen la atribución adecuada por su excelente trabajo. Todavía no he tenido la oportunidad de dárselo.

Referencias

Adair, G. (1984). The hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69*, 334–345.

Agresti, A. (1996). *An introduction to categorical data analysis*. Wiley.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, *27*, 17–21.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 398–404.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, *40*, 318–335.

Box, J. F. (1987). Guinness, gosset, fisher, and small samples. *Statistical Science*, *2*, 45–52.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association*, *69*, 364–367.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Chronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16(3)*, 297–334.

Cochran, W. G. (1954). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, *23*, 315–345.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cramer, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52–64.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions (3rd ed)*. Wiley.

Everitt, B. S. (1996). *Making sense of statistics in psychology. A second-level course*. Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Fisher, R. A. (1922a). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, *84*, 87–94.

Fisher, R. A. (1922b). On the mathematical foundation of theoretical statistics. *Philo-*

sophical Transactions of the Royal Society A, *222*, 309–368.
Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.
Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*, 328–331.
Geschwind, N. (1972). Language and the brain. *Scientific American*, *226(4)*, 76–83.
Hays, W. L. (1994). *Statistics* (5th ed.). Harcourt Brace.
Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
Hewitt, A. K., Foxcroft, D. R., & MacDonald, J. (2004). Multitrait-multimethod confirmatory factor analysis of the attributional style questionnaire. *Personality and Individual Differences*, *37(7)*, 1483–1491.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.
Hróbjartsson, A., & Gøtzsche, P. (2010). Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews*, *1*. https://doi.org//10.1002/14651858.CD003974.pub3
Hsu, J. C. (1996). *Multiple comparisons: Theory and methods*. Chapman; Hall.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), 697–701.
Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford.
Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *48*, 19313–19317.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.
Keynes, J. M. (1923). *A tract on monetary reform*. Macmillan; Company.
Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Academic Press.
Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583–621.
Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *Public Library of Science One*, *9*, 1–8.
Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, *73*, 253–263.
Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. Springer.
Levene, H. (1960). Robust tests for equality of variances. In I. O. et al (Ed.), *Contributions to probability and statistics: Essays in honor of harold hotelling* (pp. 278–292). Stanford University Press.
Meehl, P. H. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.
Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably

supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157–175.

Peterson, C., & Seligman, M. (1984). Causal explanations as a risk factor for depression: Theory and evidence. *Psychological Review*, *91*, 347–374.

Pfungst, O. (1911). *Clever hans (the horse of mr. Von osten): A contribution to experimental animal and human psychology* (C. L. Rahn, Trans.). Henry Holt.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. Appleton.

Sahai, H., & Ageel, M. I. (2000). *The analysis of variance: Fixed, random and mixed models*. Birkhauser.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584.

Sokal, R. R., & Rohlf, F. J. (1994). *Biometry: The principles and practice of statistics in biological research* (3rd ed.). Freeman.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stigler, S. M. (1986). *The history of statistics*. Harvard University Press.

Student, A. (1908). The probable error of a mean. *Biometrika*, *6*, 1–2.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.

Wilkinson, L., Wills, D., Rope, D., Norton, A., & Dubbs, R. (2006). *The grammar of graphics*. Springer.

Adair, G. (1984). The hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, *69*, 334–345.

Agresti, A. (1996). *An introduction to categorical data analysis*. Wiley.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, *27*, 17–21.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*, 398–404.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, *40*, 318–335.

Box, J. F. (1987). Guinness, gosset, fisher, and small samples. *Statistical Science*, *2*, 45–52.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for equality of variances. *Journal of the American Statistical Association*, *69*, 364–367.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Chronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16(3)*, 297–334.

Cochran, W. G. (1954). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, *23*, 315–345.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Cramer, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52–64.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions (3rd ed)*. Wiley.

Everitt, B. S. (1996). *Making sense of statistics in psychology. A second-level course*. Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299.

Fisher, R. A. (1922a). On the interpretation of χ^2 from contingency tables, and the calculation of *p*. *Journal of the Royal Statistical Society*, *84*, 87–94.

Fisher, R. A. (1922b). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society A*, *222*, 309–368.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*, 328–331.

Geschwind, N. (1972). Language and the brain. *Scientific American*, *226(4)*, 76–83.

Hays, W. L. (1994). *Statistics* (5th ed.). Harcourt Brace.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hewitt, A. K., Foxcroft, D. R., & MacDonald, J. (2004). Multitrait-multimethod confirmatory factor analysis of the attributional style questionnaire. *Personality and Individual Differences*, *37(7)*, 1483–1491.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Hróbjartsson, A., & Gøtzsche, P. (2010). Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews*, *1*. https://doi.org//10.1002/14651858.CD003974.pub3

Hsu, J. C. (1996). *Multiple comparisons: Theory and methods*. Chapman; Hall.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), 697–701.

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *48*, 19313–19317.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Keynes, J. M. (1923). *A tract on monetary reform*. Macmillan; Company.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Academic Press.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583–621.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *Public Library of Science One*, *9*, 1–8.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association*, *73*, 253–263.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. Springer.

Levene, H. (1960). Robust tests for equality of variances. In I. O. et al (Ed.), *Contributions to probability and statistics: Essays in honor of harold hotelling* (pp. 278–292). Stanford University Press.

Meehl, P. H. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157–175.

Peterson, C., & Seligman, M. (1984). Causal explanations as a risk factor for depression: Theory and evidence. *Psychological Review*, *91*, 347–374.

Pfungst, O. (1911). *Clever hans (the horse of mr. Von osten): A contribution to experimental animal and human psychology* (C. L. Rahn, Trans.). Henry Holt.

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. Appleton.

Sahai, H., & Ageel, M. I. (2000). *The analysis of variance: Fixed, random and mixed models*. Birkhauser.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584.

Sokal, R. R., & Rohlf, F. J. (1994). *Biometry: The principles and practice of statistics in biological research* (3rd ed.). Freeman.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stigler, S. M. (1986). *The history of statistics*. Harvard University Press.

Student, A. (1908). The probable error of a mean. *Biometrika*, *6*, 1–2.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.

Wilkinson, L., Wills, D., Rope, D., Norton, A., & Dubbs, R. (2006). *The grammar of graphics*. Springer.

jamovi Stats. Open. Now.

This textbook covers the contents of an introductory statistics class, as typically taught to undergraduate psychology, health or social science students. The book covers how to get started in jamovi as well as giving an introduction to data manipulation. From a statistical perspective, the book discusses descriptive statistics and graphing first, followed by chapters on probability theory, sampling and estimation, and null hypothesis testing. After introducing the theory, the book covers the analysis of contingency tables, correlation, *t*-tests, regression, ANOVA and factor analysis. Bayesian statistics are touched on at the end of the book.

Citation: Navarro DJ and Foxcroft DR (2022). learning statistics with jamovi: a tutorial for psychology students and other beginners. (Version 0.75). [DOI: 10.24384/hgc3-7p15](https://dx.doi.org/10.24384/hgc3-7p15)