2.ª EDICIÓN

MANUAL PARA EL ANÁLISIS POLÍTICO CUANTITATIVO

ADRIÁN PIGNATARO

EDITORIAL UCR
2024

Acerca de los libros digitales

- Las opciones de visualización y funcionalidades de un libro digital dependen de las capacidades de la aplicación que se utiliza para la lectura de libros digitales.
- La aplicación utilizada para lectura de libros electrónicos en formato ePub puede alterar la integridad física de poemas.
- La Editorial UCR ha hecho lo posible por asegurar que los URLs de sitios externos a los que se hace referencia en este libro sean correctos y estén activos en el momento de la publicación de este libro digital. Sin embargo, no es responsable de estos sitios web, por lo que no puede garantizar que seguirán estando activos o manteniendo contenido apropiado.
- Consulte las respuestas a preguntas frecuentes sobre libros digitales en nuestro sitio web.

CC.SIBDI.UCR - CIP/4086

Nombres: Pignataro, Adrián, autor.
Título: Manual para el análisis político cuantitativo / Adrián Pignataro.
Descripción: Segunda edición digital. | San José, Costa Rica : Editorial UCR, 2024.

Identificadores: **ISBN: 978-9968-02-142-5** (PDF)

Materias: LEMB: Estadística política. | Ciencia política – Metodología. | Estadística matemática
Clasificación: CDD 320.072.7 –ed. 23

Edición aprobada por la Comisión Editorial de la Universidad de Costa Rica.
Segunda edición digital (PDF): 2024

Ciudad Universitaria Rodrigo Facio. San José, Costa Rica.
Apdo.: 11501-2060 • Tel.: 2511 5310 • Fax: 2511 5257
administracion.siedin@ucr.ac.cr
www.editorial.ucr.ac.cr

Contenido

Prefacio

Este libro ofrece una introducción a los métodos y los modelos fundamentales de la estadística para aplicarse en la investigación empírica. Está destinado a personas estudiantes, investigadoras y docentes de ciencia política, pero personas interesadas en disciplinas relacionadas también podrían encontrar utilidad en la obra. Esta segunda edición, aunque producto de una intensa revisión, mantiene el mismo objetivo de la primera: a partir de un nivel de conocimiento previo básico en matemática y estadística, la meta es exponer de forma clara los conceptos, con un número reducido de fórmulas, sin demostraciones formales y mediante el uso de ejemplos propios de la disciplina de la ciencia política, con el propósito de que esto haga interesante lo que desde otro enfoque podría resultar intrincado y tedioso. A diferencia de su predecesora, esta edición utiliza el programa gratuito de código abierto R (https://www.r-project.org/) como la herramienta de análisis estadístico –y con este mismo programa se escribió y diagramó el libro–.

La vertiente cuantitativa es amplia. En consecuencia, este texto se limita a los métodos estadísticos clásicos o frecuentistas (no se aborda la perspectiva bayesiana) y excluye otras metodologías, también cuantitativas, como la teoría de juegos y las técnicas de análisis de datos masivos (*big data*). Por lo tanto, el libro se centra en el análisis estadístico aplicado en la ciencia política, para lo cual se organiza en nueve capítulos.

El capítulo 1 expone los principios de la investigación con orientación pospositivista y cuantitativa, así como sus alcances y limitaciones. El capítulo 2 brinda los elementos esenciales para comprender la teoría clásica de la inferencia estadística: parámetros, estimadores, errores, intervalos de confianza y pruebas de hipótesis.

Los métodos bivariados se introducen en el capítulo 3, con la comparación de dos medias (prueba t) y varias medias (análisis de variancia o Anova). Luego, el capítulo 4

proporciona algunas medidas de asociación, tanto para dos variables categóricas (prueba *chi* cuadrado y coeficiente V de Cramer), como para dos variables métricas (coeficiente r de Pearson).

Los siguientes tres capítulos se centran en modelos de regresión lineal para variables dependientes métricas (capítulos 5 y 6) y regresión logística para variables dependientes categóricas (capítulo 7). Los capítulos finales comprenden métodos tradicionalmente denominados multivariados. El capítulo 8 presenta el análisis de conglomerados, específicamente el método aglomerante jerárquico, pensado como una herramienta práctica para describir pocos casos, aplicable en política comparada, por ejemplo. Por último, el capítulo 9 introduce el análisis de factores de tipo exploratorio.

Se cierra el libro con tres apéndices. Para personas sin conocimientos sobre R, el apéndice A ofrece una concisa introducción a este lenguaje, aunque para un abordaje más comprensivo, que considere las múltiples posibilidades de R para tratar y manipular bases de datos, debe recurrirse a fuentes adicionales. El apéndice B sugiere bibliografía complementaria para profundizar temas que se estudiaron y para ahondar en otros que no se abarcaron del todo, por ejemplo, el análisis de sobrevivencia, el análisis de datos de panel y los modelos multinivel. El apéndice C hace referencia a los modelos lineales generalizados como un marco teórico que unifica la mayoría de los contenidos estudiados.

En el texto se utilizan las siguientes bases de datos en ejemplos y ejercicios:

- "CIEPnoviembre2020.dta" (capítulos 3, 4, 6, 7 y 9);
- "eleccionesCR2018.xlsx" (capítulo 4);
- "eleccionesCentroamerica.xlsx" (capítulo 8);
- "democraciasLijphart.xlsx" (capítulo 9);
- "JohanSkytte.xlsx" (apéndice A).

Estas bases se pueden descargar en la dirección: https://hdl.handle.net/10669/89296.

Antes de cerrar este prefacio, quiero agradecer a las distintas personas que apoyaron este proyecto, empezando por las generaciones de estudiantes de la Escuela de Ciencias Políticas de la Universidad de Costa Rica que pacientemente estudiaron con la primera edición de este libro y con los borradores de la segunda versión, así como a las y los colegas docentes que han encontrado útil este material para sus cursos, en particular, Allan Abarca, Carolina Zamora, Fátima Ruiz, Jesús Guzmán, Katherine Barquero, Orlando Vega, Ronald Alfaro y Steffan Gómez. Sin la positiva recepción de profesores,

profesoras y estudiantes, esta segunda versión posiblemente no habría visto la luz. Michelle Taylor-Robinson y Nehemia Geva brindaron sugerencias valiosas en torno a la enseñanza de experimentos en ciencia política. Ronald Alfaro y Jesús Guzmán me facilitaron los datos de la encuesta de noviembre de 2020 del Centro de Investigación y Estudios Políticos (CIEP) de la Universidad de Costa Rica, como material de ejemplos y ejercicios. Shu Wei Chou contribuyó para mejorar algunas explicaciones en el capítulo sobre inferencia estadística. Agradezco además a Tim Henrichsen, por motivarme a reeditar este libro con R, y a Maria Giovanna Sessa, por inspirarme a escribir una breve introducción a R para las personas no iniciadas, la cual se convirtió en el apéndice A.

Por último, reconozco el apoyo de las compañeras y los compañeros de la Editorial UCR, en particular, Alexander Jiménez, director, Jessica López y Pamela Bolaños, editoras, y Aída Cascante, jefa de la sección de diseño, por haber impulsado la publicación de esta segunda edición.

Capítulo 1

Metodología de la investigación cuantitativa

1.1 Una historia de dos disciplinas

Los métodos de investigación en ciencia política son variados: cuantitativos, cualitativos, comparativos, históricos, experimentales y etnográficos, entre otros (Box-Steffensmeier *et al.*, 2008). Los métodos estadísticos están en una posición especial, ya que, desde una perspectiva histórica, es posible distinguir una estrecha relación entre la estadística como disciplina y la ciencia política moderna. La estadística, entendida como la ciencia para aprender de los datos y medir, controlar y comunicar la incertidumbre (Davidian y Louis, 2012), ha tenido un desarrollo paralelo a las ciencias sociales modernas, en una dinámica de reforzamiento mutuo.

El vínculo entre estadística y ciencias sociales se remonta al siglo XVII, cuando se desarrolló en Inglaterra la aritmética política que recopilaba y organizaba datos políticos, sociales, demográficos y económicos. A su vez, en el continente europeo, se cultivaba la *Statistik* alemana, referida al estudio de los Estados, la cual gradualmente incorporó datos cuantitativos y dotó de nombre a la disciplina. Estas tradiciones investigativas, concentradas en la recolección de información, pero carentes de métodos analíticos, convergen con la teoría de la probabilidad francesa –más abstracta y matemática– para dar origen a la ciencia estadística (Piovani, 2007).

A finales del siglo XIX y principios del XX, la teoría estadística y sus métodos de análisis avanzaron gracias a aplicaciones en diversos campos, no solo políticos, demográficos y sociales, sino también en agronomía, genética y epidemiología, donde destacan figuras como Adolphe Quetelet, Florence Nightingale, Francis Galton, Karl Pearson y Ronald Fisher (Salsburg, 2001). No en vano, el historiador Stigler (2010) denominó a esta era "La Ilustración Estadística".

El impacto de la estadística no es solo instrumental, al suministrar métodos y técnicas a otras disciplinas, sino que también influye en la forma de pensar los problemas; se pasa de un razonamiento determinista a uno probabilístico. La ciencia política no escapó de este cambio de paradigma. A mediados del siglo XX, Barrington Moore (1966) declaraba de forma determinística "no burguesía, no democracia" (p. 418; traducción propia), para referirse a la transición de una sociedad hacia un régimen democrático, entendiendo que la presencia de la burguesía conducía inevitablemente al establecimiento de la democracia. Más adelante, cerrando el siglo, en la misma línea de investigación, pero con una argumentación probabilística, Adam Przeworski y sus coautores (2000) sostenían:

> La probabilidad de que una dictadura muera y se establezca una democracia es en gran medida aleatoria en relación con los ingresos per cápita [...]. Pero la probabilidad de que, una vez establecida, una democracia sobreviva se incrementa abrupta y monotónicamente conforme el ingreso per cápita es mayor (p. 273, traducción propia).

La estadística introduce, en consecuencia, una nueva forma de pensar los procesos políticos, como muestran los estudios citados de democratización, aunque el contraste entre miradas deterministas y probabilísticas no significa que se haya dado una conversión total hacia la segunda. Los razonamientos dependen más de las premisas epistemológicas y de los enfoques metodológicos. Los métodos cuantitativos son conceptualmente probabilísticos, mientras que, en la vertiente cualitativa, el análisis de condiciones necesarias y suficientes y el rastreo del proceso (*process tracing*) se aproximan más a la visión determinística (Beach y Pedersen, 2013). Por lo tanto, puede sostenerse que ambas perspectivas conviven en la investigación politológica actual (Mahoney y Goertz, 2006).

Distintos campos de la ciencia política han incorporado la estadística y han crecido gracias a ella. Quizás el más conocido es el estudio de la opinión pública y el comportamiento político. El análisis de la cultura política, un tema clave de la revolución conductual, surgió, en buena medida, gracias a la entonces novedosa "metodología y tecnología de

la investigación mediante encuestas" (Almond, 1999, p. 201), es decir, el desarrollo del muestreo, de las técnicas de medición y de los métodos de inferencia estadísticos.

El inicio de las encuestas por muestreo se remonta a los años del *New Deal* en Estados Unidos, cuando era necesario tener cifras sobre el desempleo y la actividad económica –otro ejemplo de complicidad entre estadística y ciencias sociales–. Poco después, George Gallup y Louis Bean iniciaron los sondeos políticos (Salsburg, 2001, pp. 172-175). En Costa Rica, las encuestas nacieron a mediados de la década de 1960 como actividades privadas de partidos políticos. Luego, en 1974, la Oficina de Información del Ministerio de la Presidencia empieza a realizar estudios periódicos de la opinión pública. Posteriormente, se establecen las primeras casas encuestadoras privadas: CID-Gallup en 1977 y UNIMER en 1986 (Hernández Rodríguez, 2004). Desde entonces, las encuestas por muestreo y el análisis estadístico constituyen –en el mundo y en Costa Rica– el estándar metodológico para estudiar el voto, la participación política y la formación de actitudes y opiniones.

Sin embargo, el aporte de los métodos cuantitativos no se limita al análisis de la opinión pública y el comportamiento político. En el campo de la política comparada, el análisis de datos agregados emergió prácticamente al mismo tiempo que el análisis de encuestas (Schmitter, 2009). En este tipo de investigación se relacionan variables que se miden en unidades políticas nacionales (países) y subnacionales (regiones, provincias, cantones), tales como indicadores económicos (producto interno bruto, inflación, tasa de desempleo), componentes institucionales (régimen político, sistema electoral) y configuraciones de actores (sistemas de partidos, pluralismo/corporativismo). Los estudios de Lijphart (1999) sobre variedades de la democracia, de Persson y Tabellini (2003) sobre los efectos económicos de las instituciones y de Przeworski *et al.* (2000) sobre condiciones económicas para la democratización ilustran el uso de métodos estadísticos en problemas de política comparada. También se analizan estadísticamente proyectos de ley (Anderson *et al.*, 2003; Muñoz-Portillo, 2021), nominaciones en cargos ejecutivos (McCarty y Razaghian, 1999), fallos judiciales (Hendershot *et al.*, 2013), vetos presidenciales (Cameron, 2000), incidentes terroristas (Enders *et al.*, 2011), discursos políticos (Jenne *et al.*, 2021), programas de partidos políticos (Chavarría-Mora y Angell, 2023), publicaciones en redes sociales (Aruguete y Calvo, 2018; Barberá *et al.*, 2019) e incluso el movimiento físico de congresistas en el plenario (Dietrich, 2021).

En el campo de las relaciones internacionales, cuyo abordaje tradicional ha sido cualitativo e histórico, hay también aplicaciones de métodos cuantitativos. Con técnicas estadísticas se han examinado las causas de los conflictos internacionales (De Mesquita y Lalman, 1988), el cumplimiento de acuerdos (McLaughlin Mitchell y Hensel, 2007) y la proliferación nuclear (Fuhrmann, 2009). La polémica tesis del choque de civilizaciones (Huntington, 1996) ha sido contundentemente desacreditada por la evidencia estadística (Chiozza, 2002; Russett *et al.*, 2000). Otros han examinado en detalle la teoría de la democracia-paz, según la cual las democracias no luchan entre sí (Lektzian y Souva, 2009). Existen, además, investigaciones que enlazan el estudio de la opinión pública con la política exterior (Berinsky, 2009; Pignataro y Cascante Segura, 2017).

En síntesis, los métodos estadísticos forman parte de la generación de conocimiento en la ciencia política a nivel global. Asimismo, el razonamiento probabilístico ha permeado sus múltiples áreas de estudio. Sin embargo, la estadística no está solamente presente en la investigación aplicada. Desde la ciencia política existe una vigorosa línea de trabajo enfocada en la metodología cuantitativa. Revistas académicas, como *Political Analysis* y *Political Science Research and Methods*, publican periódicamente artículos centrados en el desarrollo de métodos y en la prescripción de los modelos más aptos para distintos problemas según la naturaleza de los datos.

1.2 El diseño metodológico cuantitativo

Antes de abordar los métodos y los modelos estadísticos, es importante enmarcar el estilo de investigación en el cual nos ubicamos. Al adoptarse un diseño metodológico cuantitativo se parte de una perspectiva denominada pospositivista. Ontológicamente, el positivismo asumía un mundo objetivo fuera de la mente de la persona observadora. Sin embargo, a diferencia del positivismo clásico, el pospositivismo reconoce, primero, que la realidad solo se captura de manera parcial, segundo, que el conocimiento está influenciado por la persona que observa y, tercero, que el conocimiento se genera en forma de leyes probabilísticas con un grado de incertidumbre (Porta y Keating, 2008). La probabilidad, por ende, se integra no solo en el método, sino también en la visión epistemológica, como se señaló antes.

Cuando se asume un diseño metodológico cuantitativo, es necesario contar con un estado del conocimiento sobre el tema y de teorías que proporcionen las hipótesis que se probarían. Las hipótesis relacionan variables, es decir, características de las observaciones

o casos de estudio. Por ejemplo, la teoría del voto económico sostiene, en términos generales, que el desempeño económico influye en los resultados electorales (Lewis-Beck y Stegmaier, 2013). Por lo tanto, una hipótesis contrastable es: *A mayor crecimiento económico, mayor apoyo al partido en el gobierno.* En esta hipótesis, el crecimiento económico es la variable independiente, mientras que el apoyo al partido en el gobierno es la variable dependiente.

Para probar las relaciones teóricas entre variables, es decir, las hipótesis, se debe contar con datos. ¿Cómo pasar de la abstracción teórica (variables) a los datos empíricos? A través de la operacionalización. Puesto que las variables son conceptos teóricos, se debe decidir cómo medirlas en la realidad. Algunas variables son relativamente simples de operacionalizar. Para la hipótesis mencionada del voto económico, el crecimiento económico se puede medir como el cambio interanual del producto interno bruto (PIB); y el apoyo al partido en el gobierno como el porcentaje de votos que recibe en la siguiente elección el partido que ocupa la presidencia (en un sistema presidencial) o que preside el gabinete (en un sistema parlamentario).

Otras variables son más difíciles de medir. Piénsese en conceptos como apoyo al sistema político, clase social, poder presidencial e influencia internacional. En ocasiones, se deben seleccionar varios indicadores o mediciones empíricas para aproximarse a los conceptos, porque uno solo no basta. Por ejemplo, la legitimidad política, según Seligson (2002), se mide al considerar la creencia en un sistema judicial justo, en el respeto a las instituciones políticas, en la protección de los derechos básicos, en el orgullo de vivir en el sistema político y en el grado de apoyo al sistema político. Son necesarias varias preguntas en una encuesta de opinión para medir un concepto abstracto y complejo como legitimidad.

Finalmente, con base en los indicadores escogidos para las variables y los datos medidos a través de los indicadores, se prueban las relaciones entre las variables, es decir, las hipótesis teóricamente generadas. Este diseño de investigación, que inicia con teoría e hipótesis y finaliza con variables e indicadores, sigue el modelo llamado deductivo-hipotético (Bunge, 1999). Una versión estilizada de este diseño de investigación se presenta en la figura 1.1.

El modelo deductivo-hipotético (como cualquier otro modelo) representa solo una guía sobre cómo debería ser la investigación. Por un lado, en la práctica es posible –y hasta necesario– cambiar la secuencia. Puede ser que, aunque las hipótesis hayan

sido claramente esbozadas, simplemente no existan datos disponibles para probarlas. Schmitter (2008) denomina "serendipia" (*serendipity*) a este proceso en el que se aprende y se regresa a etapas anteriores de la investigación para mejorar las decisiones. En determinadas circunstancias, habría que redefinir la pregunta o repensar el alcance de la hipótesis en términos de la factibilidad.

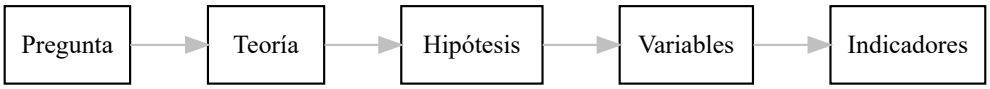


Figura 1.1 Esquema del diseño cuantitativo de investigación

Pese a la serendipia, hay acciones que desde el punto de vista ético se deben evitar. Reformular una hipótesis luego de obtener los resultados, para que los hallazgos respondan positivamente la pregunta inicial, es una práctica deshonesta. Asimismo, escoger intencionalmente métodos que arrojen resultados estadísticamente significativos (ver capítulo 2), a sabiendas de que bajo distintas condiciones no lo son, reduce el avance científico (Gerber y Malhotra, 2008; Head *et al.*, 2015).

1.3 Alcances y limitaciones

La investigación cuantitativa goza de ventajas respecto a otros métodos, pero también de limitaciones. Algunos alcances de la investigación cuantitativa son:

- *Permite estudiar una cantidad amplia de observaciones.* Los métodos estadísticos analizan grandes números y resumen vasta información que de otra manera sería materialmente intratable. Por ejemplo, los estudios de opinión por encuestas recopilan información de cientos y miles de personas, en contraste con las entrevistas cualitativas e historias de vida que se aplican en un número pequeño de casos.
- *Generaliza.* La inferencia estadística permite obtener conclusiones para una población extensa con base en una muestra seleccionada aleatoriamente que es pequeña en relación con esta población. Otros métodos de muestreo no aleatorio –por ejemplo, las entrevistas cualitativas seleccionadas por conveniencia– exploran poblaciones, pero no permiten generalizar. El capítulo 2 abarca los principios para generalizar desde una muestra.

- *Analiza muchas variables*. Los métodos clásicos de comparación infieren conclusiones utilizando pocos casos que sean diferentes en todo, excepto por una circunstancia común (método de la semejanza); o semejantes en todo, excepto en una variable (método de la diferencia). Bajo esta lógica, se seleccionan pocas variables, de forma que el diseño no se indetermine, pues si hay más variables que casos no se puede concluir cuál es el factor explicativo (la causa) con el método comparado (Lijphart, 1971). El análisis cualitativo comparado, originalmente propuesto por Charles Ragin (1987), puede incorporar más variables que los métodos clásicos de semejanza y diferencia, pero tampoco escapa del problema de muchas variables/pocos casos. En cambio, con los modelos estadísticos se pueden incluir muchas variables, ya que hay más casos y, por consiguiente, más grados de libertad. Los capítulos 6, 7, 8 y 9 ilustran la capacidad de los métodos para hacer análisis multivariable.
- *Prueba teorías explicativas*. Aunque existen estudios cuantitativos con orientación descriptiva, el fin último de una investigación en el marco pospositivista es examinar hipótesis explicativas que relacionan una variable dependiente con variables independientes (King *et al.*, 1994). Los modelos de regresión (capítulos 5, 6 y 7) son herramientas especialmente aptas para probar hipótesis explicativas.
- *Calcula un poder explicativo y un error*. En todo trabajo científico es posible equivocarse. Los métodos estadísticos permiten medir la precisión o la capacidad explicativa de un modelo y el error de una inferencia.

Ahora bien, los estudios cuantitativos conllevan las siguientes limitaciones:

- *Determinan efectos de causas, no causas de efectos*. Siguiendo a Mahoney y Goertz (2006), con los métodos estadísticos se prueba la validez de los factores explicativos, pero no se pueden establecer de antemano cuáles son estas causas o explicaciones; son las teorías previamente desarrolladas –no los métodos– las que indican cuáles variables explicativas se deben considerar. De hecho, los métodos estadísticos no están orientados hacia la generación de teoría. Las teorías e hipótesis que se prueban deben desarrollarse por otros métodos como los estudios de caso, la teoría fundamentada, la modelización formal u otros. Los métodos estadísticos pueden confirmar o refutar la teoría, pero no construirla.
- *Generalizan, pero no detallan casos particulares*. Los métodos estadísticos trabajan con efectos promedio. En cambio, explicar puntualmente por qué ocurre un fenómeno en determinado contexto requiere de herramientas cualitativas, como

el análisis histórico y el rastreo del proceso (*process tracing*), que develen los mecanismos causales del fenómeno (Beach y Pedersen, 2013).

- *El número de variables debe ser menor al número de observaciones.* Mientras que algunos métodos cualitativos, como la etnografía, son ricos en la diversidad de características y propiedades que describen, los métodos estadísticos usualmente requieren que el número de observaciones sea mayor a la cantidad de variables. Matemáticamente, los métodos estadísticos necesitan grados de libertad, es decir, más datos que incógnitas.
- *Se encuentran limitados por el desarrollo teórico y computacional.* Los modelos, métodos y técnicas que se aplican son aquellos que la teoría estadística ha formulado y que los paquetes computacionales permiten implementar. De hecho, muchos cálculos serían prácticamente intratables de no ser por las computadoras. Afortunadamente, los métodos estadísticos están en continuo desarrollo y los paquetes abiertos como R se actualizan constantemente con nuevas rutinas y funciones.

Tanto por sus limitaciones metodológicas como por su tendencia hegemónica en la ciencia política (especialmente en la academia estadounidense), la investigación con orientación cuantitativa ha recibido fuertes críticas. Un ejemplo de la reacción es el movimiento Perestroika. Este nació con un correo electrónico, firmado bajo el seudónimo Mr. Perestroika, que se diseminó masivamente. El mensaje atacaba el énfasis cuantitativo, conductista y racionalista imperante en la ciencia política anglosajona. Se decía que la estadística alcanzó niveles técnicos elevados que oscurecen la importancia sustantiva de los fenómenos e ignoran aspectos básicos sobre la definición de los conceptos y la calidad de los datos (Monroe, 2007). Posteriormente, textos como *Rethinking Social Inquiry* (Brady y Collier, 2010) intentaron debatir con la hegemonía cuantitativa y reivindicar los métodos cualitativos en la investigación politológica.

Una crítica persistente que se hace a los métodos estadísticos es que la correlación no implica causalidad, en otras palabras, que un patrón de datos no refleja necesariamente el mecanismo causal. Por ejemplo, el consumo de cigarrillos ha disminuido al mismo tiempo que la venta de DVD. Sin embargo, sería absurdo aducir que el tabaco motivaba la compra de películas en dicho formato: simplemente coinciden los cambios en el comportamiento de consumo. Es decir, hay asociación, pero no una relación causa-efecto.

Aunque "correlación no es causalidad" se ha convertido en un mantra popular, es más difícil definir qué es causalidad. Por un lado, la discusión filosófica y estadística sobre

la causalidad es amplia y compleja (Brady, 2008; Goldthorpe, 2001; Holland, 1986). Por el otro, se sostiene que la teoría estadística clásica parte de bases débiles y escasa comprensión sobre causalidad (Pearl y Mackenzie, 2018). Por ello, se han trabajado nuevos métodos estadísticos con la intención de robustecer las inferencias causales, por ejemplo, con técnicas como *matching*, modelos de efectos fijos para datos de panel, modelos de diferencias en diferencias y el uso de variables instrumentales.

Una forma de sobrellevar las limitaciones de la estadística en la ciencia política, incluida la crítica de correlación sin causalidad, consiste en combinar métodos cualitativos y cuantitativos mediante diseños o métodos mixtos (Creswell, 2009). Los diseños mixtos pueden adoptar varias formas. Aquellos de tipo concurrente aplican simultáneamente métodos cualitativos y cuantitativos de recolección y análisis de datos, estrategia también conocida como triangulación (Creswell, 2009, p. 213). Un ejemplo del diseño concurrente es la etnoencuesta (*ethnosurvey*), la cual combina la aplicación de cuestionarios semi-estructurados, aptos para el análisis estadístico, con técnicas de la etnografía como la observación y las historias de vida. La etnoencuesta se ha utilizado fructíferamente en el estudio de procesos migratorios (Massey, 1987).

Otro tipo de diseño mixto es más bien secuencial. Este diseño puede partir del análisis de datos cualitativo y continuar con el análisis de datos cuantitativo, o al revés. En el primer caso, una investigación mixta puede iniciar con un análisis cualitativo, como un grupo focal que permita construir un cuestionario para aplicarse en una encuesta con muestreo. En el segundo caso, puede realizarse un análisis cuantitativo inicial de muchas observaciones para luego profundizar en algunos casos que se desvíen del patrón (los casos desviantes). En la política comparada esta estrategia se ha denominado análisis anidado (*nested analysis*) (Lieberman, 2005). En este diseño, se parte del estudio cuantitativo de múltiples países para luego profundizar cualitativamente en algunos casos que mejor ejemplifican la teoría o intentar explicar casos que más bien se desvían de ella. Por ejemplo, Mainwaring y Pérez-Liñán (2013) desarrollan una teoría sobre sobrevivencia y caída de las democracias, la cual, en una primera fase, prueban en 20 países latinoamericanos entre 1945 y 2005. En una segunda etapa, analizan dos casos específicos –Argentina y El Salvador– con evidencia histórica para profundizar en los mecanismos causales. Este diseño mixto, por lo tanto, combina las fortalezas de análisis cuantitativo y cualitativo.

1.4 Algunos conceptos claves

Antes de proceder con los modelos y las técnicas del análisis cuantitativo, conviene repasar algunos conceptos sobre tipos de datos y de variables. Desconocer la estructura de los datos o interpretar erróneamente los niveles de medición puede llevarnos a resultados problemáticos, no importa qué tan sofisticado sea el método.

1.4.1 Tipos de datos

Primero, los datos se distinguen según su estructura en transversales, series de tiempo y longitudinales o de panel. Los *datos transversales* se obtienen para múltiples observaciones ($N > 1$) con una única medición en el tiempo ($T = 1$). Una encuesta en la que las personas son entrevistadas una vez (aunque los días de la entrevista pueden variar) es un conjunto de datos transversales. Otro ejemplo es la comparación de los niveles de participación electoral entre las ocho circunscripciones electorales en Costa Rica para la elección presidencial de 2022 (datos de Tribunal Supremo de Elecciones, 2022). La participación, en este caso, se mide una vez en el tiempo y, por lo tanto, corresponde a datos transversales (figura 1.2). Por el contrario, si examinamos los porcentajes de participación en todas las elecciones presidenciales de Costa Rica, desde 1953 hasta 2022 (figura 1.3), tenemos un ejemplo de *serie de tiempo* o *serie cronológica*: varias mediciones temporales ($T > 1$) para una misma unidad de análisis ($N = 1$). Las series de tiempo son comunes en economía (desempleo, inflación, porcentaje de pobreza); en ciencia política pueden encontrarse, por ejemplo, en los porcentajes de aprobación de las personas gobernantes de un país en el tiempo o en el número de leyes aprobadas anualmente por un congreso.

Cuando existe una combinación de múltiples mediciones en el tiempo ($T > 1$) para varias observaciones ($N > 1$) se tienen *datos de panel* o *longitudinales*. El término panel, utilizado para describir datos, nació en el contexto de encuestas, específicamente en los estudios de opinión pública conducidos por el sociólogo Paul Lazarsfeld y sus colegas, bajo la idea de encuestar a las mismas personas varias veces para determinar los cambios individuales en el tiempo (Lazarsfeld y Fiske, 1938). Los datos de panel formados por individuos se llaman micropaneles. También existe el macropanel, cuando se repiten mediciones temporales a conjuntos de países u otras unidades geográficas. La participación histórica en las provincias costarricenses, desde 1953 hasta 2022, ejemplifica un macropanel, ilustrado en la figura 1.4.

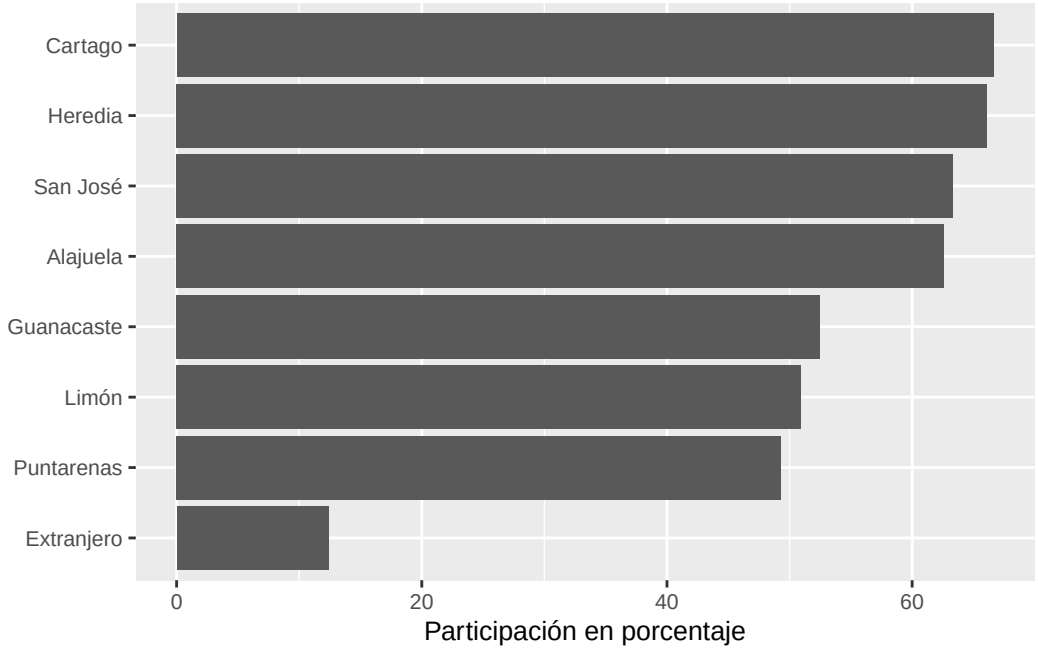


Figura 1.2 Ejemplo de datos transversales: participación electoral en la elección presidencial de Costa Rica 2022 (primera vuelta); datos del Tribunal Supremo de Elecciones (2022)

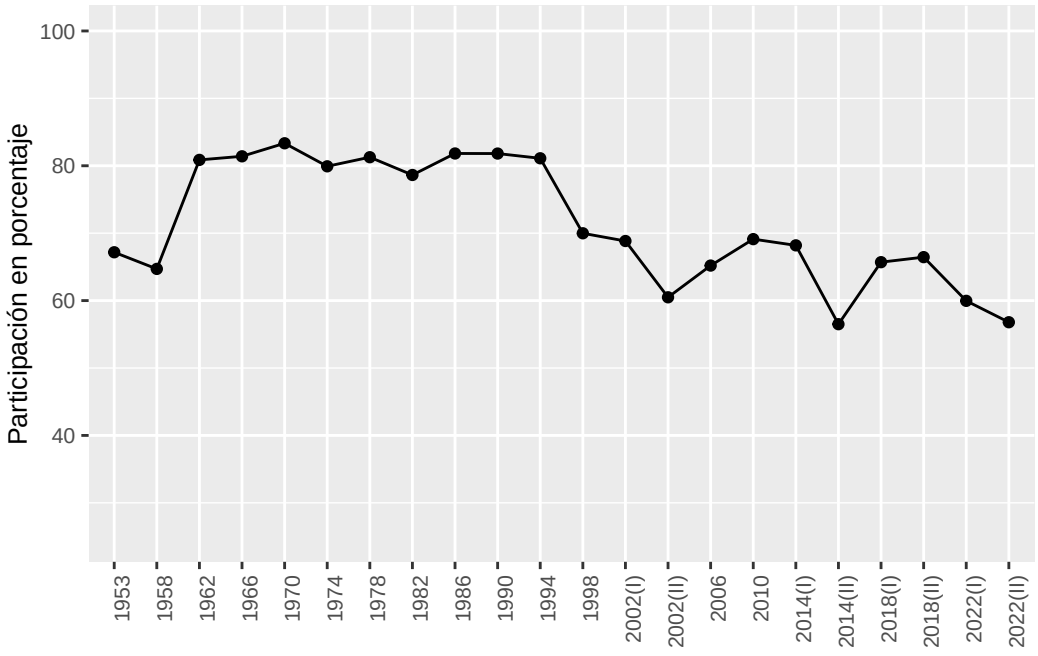


Figura 1.3 Ejemplo de serie de tiempo: participación electoral en Costa Rica, periodo 1953-2022; datos del Tribunal Supremo de Elecciones (2022)

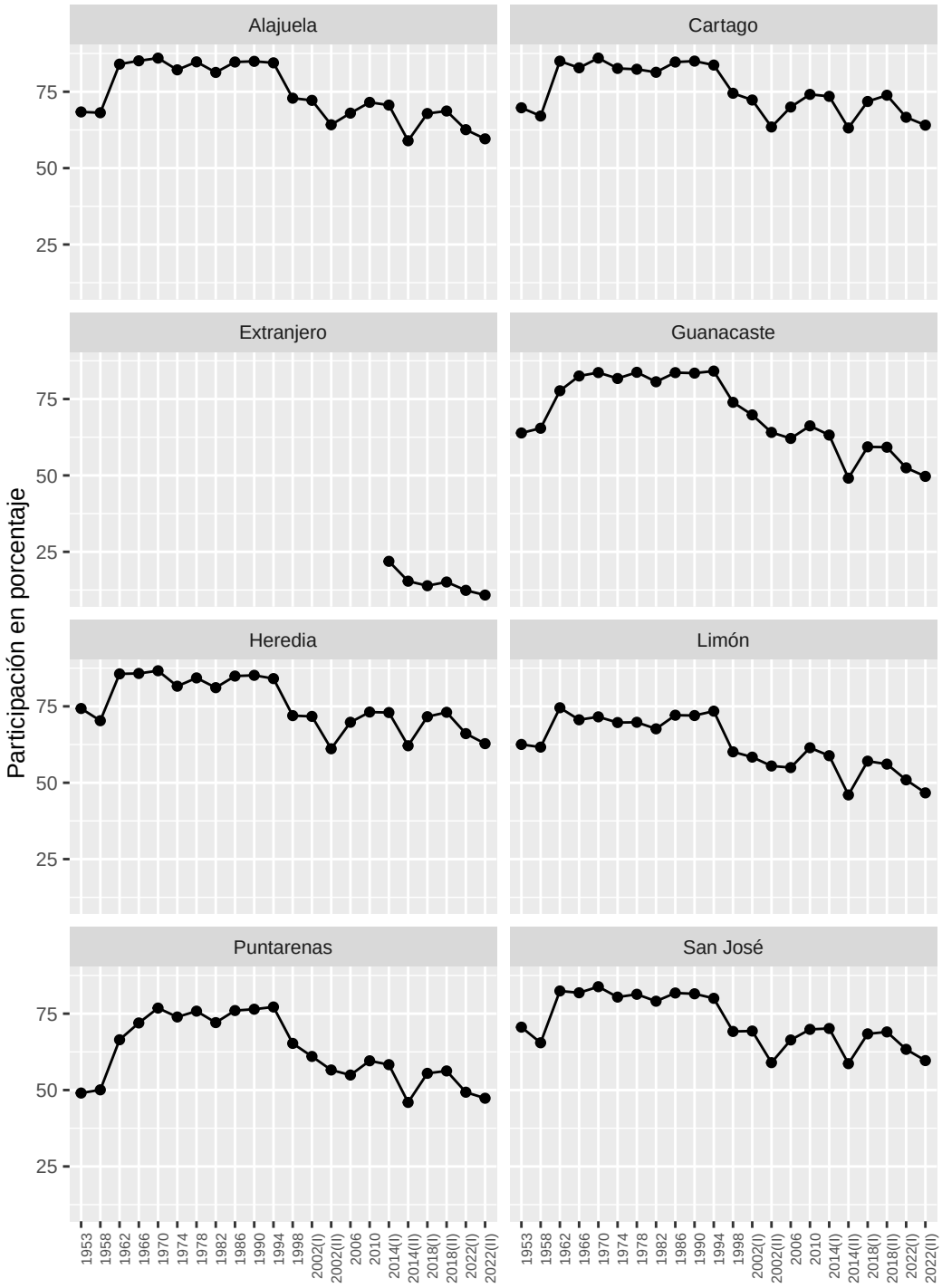


Figura 1.4 Ejemplo de datos de panel: participación electoral en Costa Rica según circunscripciones, periodo 1953-2022; datos del Tribunal Supremo de Elecciones (2022)

Por otra parte, los datos se clasifican según su proceso de generación como experimentales y observacionales. En los *experimentos* la variación de los datos es una consecuencia parcial del diseño de la investigación, generalmente la asignación (o no) de un tratamiento a las unidades de observación (Morton y Williams, 2008). Un experimento se entiende fácilmente cuando se piensa en forma de ensayo clínico para probar un medicamento: dos grupos de personas reciben aleatoriamente (y sin saber a cuál grupo pertenecen) dos medicamentos, uno viejo y uno nuevo. Puesto que el tratamiento se distribuye aleatoriamente, las diferencias que se observen, por ejemplo, en la eficacia promedio, se puede atribuir al tratamiento y no a otras características presentes en las personas.

En ciencia política, se realizan experimentos en laboratorio, utilizando distintos estímulos visuales y textuales (el capítulo 3 incluye un caso). Fuera del laboratorio, en el campo, se ha experimentado con modalidades para impulsar la participación electoral, comparando tratamientos de motivación para salir a votar (asignados al azar): contacto personal, por teléfono y por correo (Green y Gerber, 2008).

En cambio, los *datos observacionales* son aquellos cuyas variaciones no se generan en el diseño de investigación, sino que son producto de factores externos: la naturaleza, la historia, la política, etc. No hay tratamientos asignados aleatoriamente y la inferencia causal es compleja, pues las causas se confunden con las características preexistentes de las observaciones.

1.4.2 Tipos de variables

Otra distinción conceptual se refiere a las variables y sus niveles de medición (figura 1.5). Las variables son *categóricas* o *cualitativas* cuando se refieren a atributos y *métricas* o *cuantitativas* cuando expresan cantidades. Las variables categóricas se miden en un nivel nominal cuando no existe un orden en sus categorías. Por ejemplo, son nominales el área de residencia (urbana o rural) y el partido por el que una persona simpatiza. Las categóricas son ordinales cuando sí existe un orden para las categorías: clase social (baja, media, alta), nivel de instrucción (sin estudios, primario, secundario, universitario) y las categorías de respuestas en escalas Likert (muy mal, mal, regular, bien y muy bien).

Las variables métricas son de intervalo y de razón. La diferencia está en el significado del cero: para las escalas de intervalo, el cero es un valor arbitrario; para las de razón, el cero significa la ausencia de la cantidad.

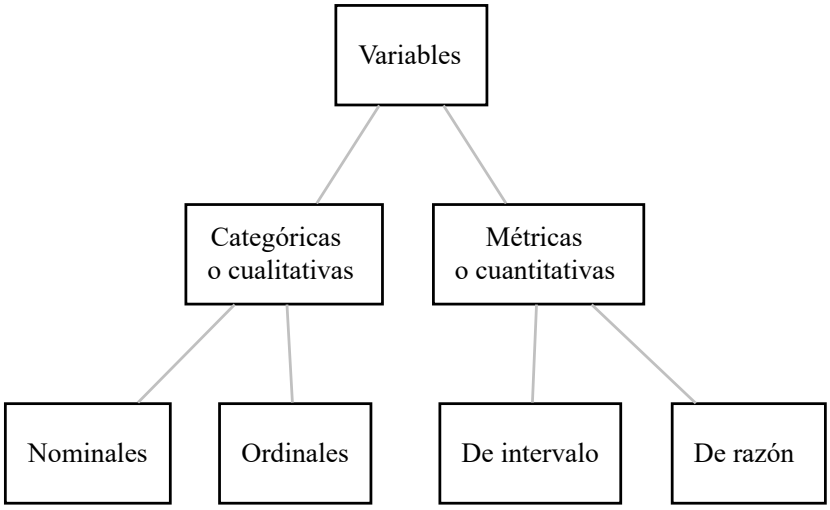


Figura 1.5 Tipos de variables y niveles de medición

Por ejemplo, en el proyecto *Polity* (https://www.systemicpeace.org/polityproject.html) se clasifican regímenes políticos en una escala que varía entre -10 (autocracia) y 10 (democracia), donde el cero no se interpreta como ausencia de régimen político, sino como un régimen intermedio entre autocracia y democracia. Tiene, por tanto, un nivel de medición de intervalo. En cambio, el número de partidos políticos que compiten en una elección es una variable métrica de razón, pues cero significa ausencia de partidos.

Los niveles de medición de las variables se deben asumir de una forma práctica y flexible, no dogmática (Velleman y Wilkinson, 1993). Una variable no es solo de un tipo, sino que se mide de una forma específica, y puede transformarse entre tipos. La democracia, por ejemplo, no es ni métrica ni categórica, sino que se puede medir como una variable métrica, en escalas -10 a 10 (proyecto *Polity*), 0 a 1 (proyecto *Varieties of Democracy*, https://www.v-dem.net/), o como una tipología categórica, en democracia, semidemocracia y autocracia (Mainwaring y Pérez-Liñán, 2013). De la misma forma, podemos medir edad en años cumplidos como escala métrica, en grupos ordinales de edad (0 a 9 años, 10 a 19 años y así sucesivamente) o en categorías nominales (jóvenes, adultos medios, adultos mayores). Los programas estadísticos permiten transformar variables entre tipos, pero solamente en el sentido de mayor a menor información: podemos recodificar la edad en años cumplidos a los grupos de edad, pero no podemos, partiendo de los grupos de edad, calcular la edad en años cumplidos.

Finalmente, la distinción entre métricas y categóricas es importante porque define qué técnica y modelo es más apropiado. En el nivel elemental, para las variables categóricas no se calculan promedios aritméticos, sino porcentajes; lo mismo ocurre con los modelos de regresión que se estudiarán en el libro: el tipo de variable dependiente define el modelo adecuado. Reconocer el tipo de variable es un paso esencial antes de entrar en la fase analítica.

1.5 Comentarios finales

Este capítulo evidenció que la investigación cuantitativa cuenta con una larga y consolidada tradición en la ciencia política, la cual ha contribuido a responder preguntas de investigación en sus diversas áreas: opinión pública, comportamiento político, política comparada, políticas públicas y relaciones internacionales. Sin embargo, deben considerarse los supuestos epistemológicos, los alcances y las limitaciones en esta forma de investigar la política.

Antes de recurrir inmediatamente a la estadística hay que pensar cuál es el mejor método para contestar la pregunta de investigación. Si interesa analizar un amplio número de casos, examinar muchas variables, generalizar, probar teorías explicativas y calcular errores medibles, entonces las técnicas estadísticas resultan apropiadas. Cuando se quieren determinar causas de efectos (*i. e.*, desarrollar teoría), detallar casos particulares y construir descripciones densas de pocos casos, otros métodos resultarían preferibles.

Asimismo, debe distinguirse cuidadosamente la estructura de los datos (transversal, serie temporal y de panel), el proceso de generación de datos (experimental y observacional), los tipos de variables (métricas y categóricas) y los niveles de medición (nominal, ordinal, de intervalo y de razón), pues para cada caso varían los modelos y métodos pertinentes.

Capítulo 2

Nociones generales de inferencia estadística

2.1 Introducción

La inferencia estadística busca obtener conclusiones sobre cantidades desconocidas, como las características de una población y las relaciones hipotéticas entre estas características (Gelman *et al.*, 2004). En el caso de la inferencia para una población, los resultados obtenidos en una muestra aleatoria se generalizan hacia la población de la cual se extrajo la muestra. Al generalizar, se puede medir la precisión (o el error) con un determinado nivel de confianza.

Seguidamente, se verán algunos conceptos fundamentales para entender cómo se realiza la inferencia para una población dentro del marco teórico de la estadística clásica, también llamada frecuentista. En cambio, las inferencias de las relaciones hipotéticas entre variables se examinarán con detalle en los capítulos posteriores destinados a los modelos de regresión.

2.2 Estimación

Un principio general de la investigación cuantitativa es que, para resolver una pregunta, debe incluirse en el análisis todo el universo o la población de casos y no una selección intencional de estos; por ejemplo, si interesa analizar el desempeño económico de las

democracias en desarrollo, deben considerarse *todos* los países que correspondan a la definición de "democracias en desarrollo", no únicamente aquellos de mayor relevancia internacional o aquellos más conocidos (Geddes, 2003).

Este universo o población responde a una delimitación conceptual de qué se quiere estudiar. Puede ser la población total de un país, el total de personas en edad de votar, el total de personas en la fuerza laboral, etc. Una población puede ser también el total de personas legisladoras de un congreso nacional –número considerablemente más pequeño que la totalidad de personas habitantes de un país– que constituye una población en un sentido conceptual: incluye todo el universo de congresistas. Asimismo, partidos políticos, leyes, decretos ejecutivos y conflictos internacionales pueden constituir poblaciones analíticas.

En ocasiones, sin embargo, las poblaciones son demasiado grandes para ser medidas en su totalidad, como cuando la población en un país equivale a millones de habitantes o cuando el total de leyes emitidas en un parlamento abarca miles. En casos donde es costoso o inverosímil medir el universo completo, resulta preferible extraer una muestra. No obstante, en comparación con otros métodos, la estadística posee una ventaja: si esta muestra se extrae de forma aleatoria, en la que *todos los elementos tienen la misma probabilidad de ser seleccionados*, es posible realizar inferencias estadísticas para la población determinada, es decir, generalizar a partir de la muestra.

Debe tenerse presente que las inferencias se harían únicamente respecto al universo del cual se extrajo la muestra. Así, una muestra aleatoria de votantes en un cantón permite inferencias para la población de votantes en el mismo cantón, no sobre las personas votantes de otros cantones, ni tampoco personas que habitan en el cantón pero que no son votantes (*e. g.*, personas extranjeras en algunos países). Asimismo, las modalidades con las que se contactan a las personas en una encuesta delimitan cuál es la población. Una encuesta telefónica permite inferir solamente a la población que tiene teléfono y no a todas las personas que habitan en el territorio. Lo mismo ocurre con las encuestas en línea: en principio, permiten analizar únicamente personas con acceso a internet (sobre las modalidades de entrevista y contacto, ver Berinsky, 2017).

Los diseños de muestreos son variados y pueden resultar complejos (Kish, 1965; ver también Hernández Rodríguez, 2012, pp. 9-21). En ocasiones es útil realizar muestreos en varias etapas, primero, seleccionando conglomerados de viviendas en un país, luego, viviendas en los conglomerados y, por último, personas dentro de las viviendas.

En este capítulo se asume el muestreo simple al azar. Teóricamente consiste en tener un listado completo de la población y seleccionar en este una muestra mediante algún procedimiento que garantice la aleatoriedad.

En la inferencia estadística existen dos conceptos claves: parámetro y estimador. Se denomina *parámetro* al valor desconocido que se quiere estimar mediante la inferencia estadística. El parámetro es una característica de la población o una relación hipotética entre variables, como los coeficientes de regresión (que se estudian en posteriores capítulos). No obstante, un parámetro no es lo mismo que una medición poblacional, por ejemplo, censal, pues una medición en un censo es vulnerable a errores (como ocurre cuando se excluyen viviendas, por ejemplo). El parámetro es una cantidad teórica, no empírica, sin ningún error de medición. Convencionalmente los parámetros se denotan con letras griegas (θ, μ, σ, β u otras). Por su parte, el *estimador* es el cálculo o fórmula que se utiliza para estimar el valor desconocido o parámetro con base en un conjunto de valores muestrales extraídos de una población.

Para comprender mejor, imaginemos que interesa conocer la edad promedio de las personas que habitan un país. Denotamos este parámetro o cantidad desconocida como μ. Para conocer el valor, descartamos la realización de un censo por su elevado costo y preferimos realizar una encuesta por muestreo. Utilizamos un procedimiento que nos permita seleccionar una muestra de personas al azar a partir de un marco muestral o listado de la población. Como la muestra se obtiene de forma aleatoria, se puede calcular el valor promedio de la edad e inferirlo a la población del país.

¿Cuál estimador debería utilizarse para estimar el parámetro μ? Esta es una pregunta complicada, pues es mediante la teoría estadística matemática, con métodos como la máxima verosimilitud, que se obtienen los estimadores. Sin ahondar en las demostraciones, la teoría indica que cuatro propiedades definen a los mejores estimadores: deben ser insesgados, consistentes, suficientes y eficientes (Wackerly *et al.*, 2002).

Insesgado significa que el estimador es, en promedio, igual al parámetro. A este "en promedio" se le llama valor esperado o esperanza matemática. Formalmente se expresa:

$$E(\hat{\theta}) = \theta$$

La expresión anterior indica que, si θ es un parámetro y $\hat{\theta}$ su estimador, entonces el valor esperado de $\hat{\theta}$ (o sea, el promedio de $\hat{\theta}$) es igual al parámetro θ. Esto implica que

el estimador es insesgado. En cambio, si el valor esperado del estimador es diferente del parámetro, $E(\hat{\theta}) \neq \theta$, se dice que el estimador es sesgado. Por lo tanto, se puede definir el sesgo como la diferencia entre el valor esperado del estimador y el parámetro:

$$Sesgo = E(\hat{\theta}) - \theta$$

Por ejemplo, la teoría estadística nos dice que el promedio aritmético, denotado $\bar{x}$, es el estimador insesgado de μ. Es decir, puesto que el promedio aritmético se calcula

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

donde x_i es cada valor y n es el tamaño de muestra, entonces, $E(\bar{x}) = \mu$. En cambio, otra fórmula, como el promedio geométrico, resulta sesgado para estimar μ.

El promedio geométrico se calcula:

$$\bar{x}_g = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}} = \sqrt[n]{x_1 * x_2 * x_3 * ...x_n}$$

$E(\bar{x}_g) \neq \mu$, pues el valor esperado del promedio geométrico es un estimador sesgado de la media μ.

La propiedad de sesgo implica una abstracción: imaginamos que extraemos múltiples muestras de una población; se aplica la fórmula del estimador a cada muestra y se observa si el promedio de los resultados es igual o no al parámetro. En la práctica, sin embargo, se extrae solo *una* muestra y no se puede determinar si el estimador es igual o no al parámetro, pues este es, en principio, una cantidad desconocida. Entonces, ¿cómo saber cuáles estimadores son insesgados? Tenemos dos opciones. Una es recurrir a las demostraciones matemáticas que constituyen la teoría estadística (Wackerly *et al.*, 2002). La otra es examinar el comportamiento de los estimadores mediante simulaciones de datos (Mooney, 1997).

Las simulaciones, también llamadas experimentos Monte Carlo, fueron creadas por John von Neumann y Stanislaw Ulam para resolver problemas complejos de probabilidad, como la reacción nuclear en cadena, cuando trabajaban en el Proyecto Manhattan en el que se crearon las primeras bombas atómicas (Bhattacharya, 2021). El método Monte Carlo permite crear poblaciones con programas computacionales, con lo cual es posible predeterminar el valor de los parámetros que, en circunstancias habituales, son

desconocidos. De esta forma se puede evaluar el resultado de estimaciones en múltiples muestras versus el valor real del parámetro que se definió.

Así, en el programa estadístico R definimos ("creamos") una población con distribución normal con media 43 y desviación estándar de 17. Cuando se ejecutan las siguientes líneas, se obtiene una muestra aleatoria de 100 observaciones de la población definida de media 43 y desviación estándar 17:

```
mu<-43
sigma<-17
n<-100
rnorm(n, mean=mu, sd=sigma)
```

Como estamos creando una población, se conoce *a priori* el valor del parámetro, $\mu = 43$, lo cual en otros contextos es imposible. En consecuencia, con esta población creada, se puede evaluar qué tan buenas son las estimaciones de los promedios aritmético y geométrico al comparar los valores muestrales respecto al parámetro. Para este ejemplo, establezco un valor inicial con `set.seed()` que permite replicar los resultados aleatorios. Además, utilizo el paquete `psych` que tiene programada la función para el cálculo de la media geométrica, `geometric.mean()`. De la población definida de media 43 y desviación estándar 17, extraigo seis muestras y, en cada una de ellas, calculo la media con ambos estimadores: el promedio aritmético y el promedio geométrico.

```
library(psych)
set.seed(229)
mediaarit<-0
mediageo<-0
for(i in 1:6){
mediaarit[i]<-mean(rnorm(n, mu, sigma))
mediageo[i]<-geometric.mean(rnorm(n, mu, sigma))
}
```

Examinemos los resultados. Para el promedio aritmético hay tres estimaciones mayores al valor real y tres menores a este. El promedio (o valor esperado) de las estimaciones con el promedio aritmético es 43.256, muy similar al valor real, $\mu = 43$, y el sesgo (la diferencia entre el valor estimado de las estimaciones y el valor real) es reducido: 0.256.

```
mediaarit
```

```
## [1] 40.83163 42.33777 42.69483 44.49376 45.21727 43.95929
```

```
mean(mediaarit) #valor esperado
```

```
## [1] 43.25576
```

```
mean(mediaarit)-mu #sesgo
```

```
## [1] 0.2557561
```

Por el contrario, el promedio geométrico muestra estimaciones sistemáticamente menores al valor real en las seis muestras; o sea, tiende a subestimar el promedio. El valor esperado de las estimaciones es 40.296 y el sesgo -2.704.

```
mediageo
```

```
## [1] 39.29473 39.18086 41.27546 40.54171 39.75604 41.72544
```

```
mean(mediageo) #valor esperado
```

```
## [1] 40.29571
```

```
mean(mediageo)-mu #sesgo
```

```
## [1] -2.704294
```

En resumen, por medio de las simulaciones Monte Carlo y sin recurrir a demostraciones matemáticas, podemos observar que el promedio geométrico es un estimador sesgado de la media. Las simulaciones, de esta forma, nos ayudan a observar cuánto afecta escoger el estimador equivocado.

Otro caso conocido de estimador sesgado es la desviación estándar (σ) cuando se utiliza la fórmula:

$$s' = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

En términos formales, $E(s') \neq \sigma$. Sin embargo, con un poco de matemática, se encuentra que el estimador de la desviación estándar resulta insesgado si se calcula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Es decir, $E(s) = \sigma$. Por lo tanto, al trabajar con muestras, se debe utilizar el estimador insesgado s de desviación estándar, no s'.

En este punto es importante tener presente que se escogen los estimadores insesgados en tanto interese inferir para poblaciones a partir de muestras. Como evidenció el ejemplo de los promedios, con un estimador sesgado los resultados difieren sistemáticamente (es decir, más allá de un error aleatorio) de los valores reales. Pero si el objetivo no es la inferencia desde la muestra a la población, las consideraciones son otras. Por ejemplo, ciertas fórmulas electorales utilizan el promedio geométrico (Shugart y Taagepera, 2017). En encuestas a personas expertas se recomienda utilizar la mediana y la moda para agregar las respuestas, no el promedio aritmético (Lindstädt *et al.*, 2020). Puesto que en ninguno de estos casos hay inferencia estadística de muestras a poblaciones, utilizar estimadores sesgados respecto a la media no es problemático.

Una segunda propiedad de los buenos estimadores es la *consistencia*. Esta indica que conforme el tamaño de muestra aumenta, el estimador se aproxima al valor real, lo que también se conoce como la ley de los grandes números. Si se piensa en una moneda perfectamente equilibrada, cada cara tiene una probabilidad de 0.5 de salir en un lanzamiento. Es posible, sin embargo, que en múltiples lanzamientos se obtenga una misma cara. Simulando este experimento en R, se generan rondas desde 10 hasta 1000000 de lanzamientos donde la probabilidad real es 0.5:

```
set.seed(178)
probabilidades<-c(mean(rbinom(10,1,0.5)),
                  mean(rbinom(100,1,0.5)),
                  mean(rbinom(1000,1,0.5)),
                  mean(rbinom(10000,1,0.5)),
                  mean(rbinom(100000,1,0.5)),
                  mean(rbinom(1000000,1,0.5)))
probabilidades
```

```
## [1] 0.400000 0.530000 0.486000 0.505500 0.498480 0.500736
```

En diez lanzamientos, se obtuvo cuatro veces una misma cara, una proporción de 0.40. Sin embargo, conforme aumenta el número de lanzamientos, la proporción de una misma cara se aproxima con mayor exactitud a la proporción real de 0.5. Esto refleja la propiedad de consistencia.

Se espera también que los estimadores sean *suficientes*, por lo cual se entiende que estos utilicen toda la información disponible. Por ejemplo, la moda no es un estimador suficiente de la media, ya que no utiliza toda la información de una serie de datos, solamente incluye las observaciones que contienen el valor repetido con mayor frecuencia. Asimismo, el rango, una medida de variabilidad que consiste en la diferencia entre el valor máximo y el mínimo, no es un estimador suficiente de la desviación estándar, pues se calcula con solo dos valores. Por el contrario, el promedio aritmético es suficiente, ya que suma todos los valores de la muestra en su estimación de la media.

Finalmente, los estimadores deben ser *eficientes*, en el sentido de que produzcan menor variancia o menor error en la estimación. Cuanto mayor sea la eficiencia de un estimador, menor el error y mayor la precisión en la estimación. Esto conlleva a discutir el error en la inferencia estadística.

2.3 Cálculo de errores

Cuando en estadística se habla de error, no debemos entenderlo como equivocaciones al realizar cálculos y fallos materiales; aunque posibles, no son medibles. El error, en términos de la inferencia estadística, se refiere a la imprecisión al estimar parámetros.

La imprecisión en la estimación está directamente relacionada con la variabilidad de un fenómeno. Cuanto más variable, diverso o heterogéneo es lo que queremos medir, más difícil es realizar inferencias. En el caso del estudio de la opinión pública, si todas las personas pensaran igual, bastaría aplicar un cuestionario a una de ellas, preguntarle qué piensa del gobierno y por quién va a votar para automáticamente saber lo que piensa el resto de la población. Sobra decir que no es así en la realidad... Las actitudes y preferencias políticas son muy distintas, muy variables. No basta con conocer la opinión de una persona, ni siquiera de unas pocas, para determinar el resto. Y cuanto más diversas sean estas opiniones, más difícil será generalizar. En resumen, *el error es proporcional a la variabilidad.*

Como se decía en el capítulo 1, una fortaleza de la estadística es que permite medir los errores en sus estimaciones. Existen, sin embargo, otros errores que no provienen de la inferencia estadística, sino de problemas en el trabajo de campo de las encuestas o en la aplicación de otras técnicas de recolección de datos: duplicados y faltantes en el marco muestral o listado del cual se extrae la muestra (lo cual implica que las probabilidades

de selección no son iguales para todos los elementos), problemas de cobertura (omisiones en el trabajo de campo de elementos que debían medirse), no respuesta y rechazo de encuestas, errores en la codificación, entre otros (Kish, 1965; Krosnick, 1999). Por ejemplo, se han identificado problemas para determinar la riqueza de las personas de rentas mayores, incluso si se aplican diseños muestrales apropiados. En consecuencia, las estimaciones de la desigualdad económica tienden a estar sesgadas en el sentido de que subestiman la riqueza total de los estratos más ricos.

Desafortunadamente, como los valores reales (parámetros) son desconocidos, estos sesgos y errores difícilmente se pueden medir, por lo que la calidad en el trabajo de campo es la única garantía para reducirlos. En la siguiente sección se enseña cómo calcular los errores que sí se pueden estimar, es decir, aquellos relacionados con la inferencia estadística, a través de los intervalos de confianza.

2.4 Intervalos de confianza

Una forma de medir la precisión de las estimaciones es a través de los intervalos de confianza, un aporte del estadístico polaco Jerzy Neyman (1894-1981). Veremos la construcción de intervalos de confianza para dos casos, promedios y porcentajes, pero los intervalos se pueden calcular para muchos otros estimadores, como la desviación estándar y los coeficientes de regresión.

2.4.1 Intervalo de confianza de una media

Para ejemplificar el cálculo del intervalo de confianza de un promedio, pensemos en una encuesta en un país, con la cual se quiere inferir el promedio de las edades de las personas electoras (mayores de 18 que pueden votar). Se extrae una muestra aleatoria de 1500 personas entrevistadas en modalidad cara a cara. El universo de estudio son las personas electoras en el momento en que se realizó la encuesta.

Como la teoría estadística indica que el promedio aritmético es el mejor estimador de la media (es insesgado, consistente, suficiente y eficiente), se calcula este promedio con base en los datos de la muestra. Se obtiene que el promedio muestral es 43.5 años (con desviación estándar de 17.3); esta se denomina una *estimación puntual*.

¿Cómo saber qué tan precisa es la estimación de 43.5 años respecto al valor real (parámetro) de la media de la población que es desconocida? Para responderlo se calcula el margen de error, con el cual se construyen intervalos de confianza. Primero, se obtiene el error estándar, definido como la desviación estándar (σ) entre la raíz cuadrada del tamaño de la muestra (n), es decir:

$$\textit{Error estándar de la media} = \frac{\sigma}{\sqrt{n}}$$

Como se mencionó, los errores son proporcionales a la variabilidad: a mayor desviación estándar, mayor error; esto, sin embargo, se compensa con el tamaño de muestra que está en el denominador: a mayor muestra, menor error. Puesto que el valor de σ se desconoce, porque es también un parámetro, puede utilizarse la desviación estimada s a partir de la muestra, lo cual convierte a la siguiente expresión en un estimador del error estándar (Wooldridge, 2023):

$$\textit{Estimación del error estándar de la media} = \frac{s}{\sqrt{n}}$$

En una segunda instancia, el error estándar se multiplica por un valor $\pm z$ que proviene de la distribución normal estándar. Con esto se obtiene el margen de error:

$$\textit{Margen de error de la media} = \pm z * \frac{\sigma}{\sqrt{n}}$$

Para el margen de error, el valor z se escoge, no se calcula, según el nivel de confianza con el cual se quiere trabajar. Un nivel de confianza de 95 % (el más utilizado) implica que, en el rango de aproximadamente dos errores estándar en la distribución normal estándar, en específico, ± 1.96 errores estándar, el margen de error contiene 95 % de las estimaciones de la media (figura 2.1). Se utiliza la distribución normal porque, según el *teorema del límite central*, los promedios estimados siguen esta distribución de probabilidad, si la muestra es grande. Es decir, más allá de cómo se comporta la variable de interés (en el ejemplo, la edad), podemos asumir cómo se distribuyen *los promedios* de esta variable y, en consecuencia, podemos calcular el error al estimar este promedio.

Téngase presente que, aunque habitualmente el nivel de confianza usado es 95 %, no existe una justificación teórica detrás de ello, sino una convención. Es posible escoger otros niveles de confianza y calcular el valor z correspondiente.

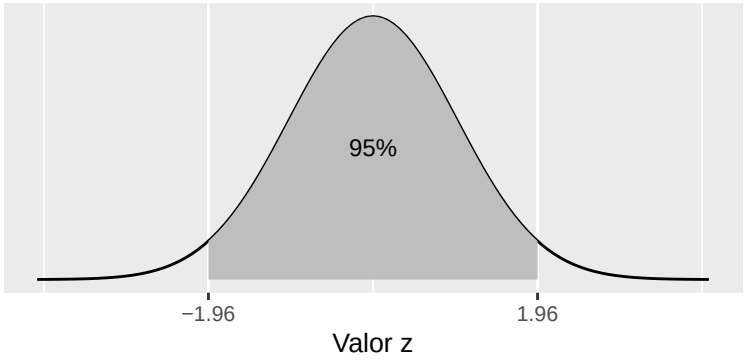


Figura 2.1 Probabilidad acumulada en la curva normal (95 %)

Por ejemplo, si se prefiere el nivel de confianza de 99 %, se multiplicaría por ±2.58 el error estándar, porque el rango de ±2.58 errores estándar contiene el 99 % de los promedios; con ±1.645 errores estándar, el 90 % (figura 2.2).

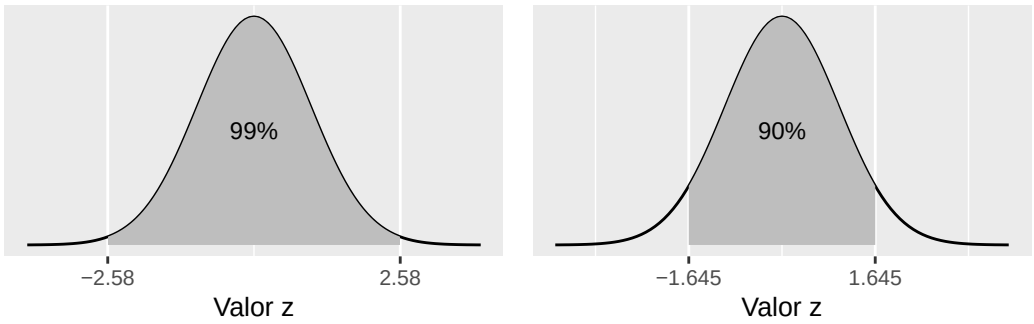


Figura 2.2 Probabilidades acumuladas en la curva normal (99 % y 90 %)

En general, es fácil obtener el valor necesario de z para cualquier nivel de confianza (en escala de 0 a 100 %) con el siguiente código en R:

```
confianza<-99
qnorm((1+(confianza/100))/2)
```

```
## [1] 2.575829
```

```
confianza<-90
qnorm((1+(confianza/100))/2)
```

```
## [1] 1.644854
```

Claramente, un mayor nivel de confianza incrementa el margen de error y habría que aumentar la muestra para contrabalancear el mayor error. Estas son decisiones que la persona investigadora debe ponderar al escoger un nivel de confianza.

En el ejemplo, desconocemos el valor de σ, por lo que usamos desviación estándar de la muestra que es 17.3 años. Entonces, el margen de error con una confianza del 95 % es:

$$\pm 1.96 * \frac{17.3}{\sqrt{1500}} = \pm 0.88$$

Para obtener el intervalo de confianza, se suma y resta el margen de error a la estimación puntual:

- Límite inferior: 43.5 – 0.88 = 42.6 años
- Límite superior: 43.5 + 0.88 = 44.4 años

Estos valores constituyen el intervalo [42.6, 44.4], el cual se interpreta de la siguiente manera: con un nivel de confianza de 95 %, el intervalo [42.6, 44.4] contiene el valor real (el parámetro) del promedio de edad de las personas electoras en el país donde se realizó la encuesta. Por su parte, la confianza dice que, si se extraen 100 muestras de idéntico tamaño, 95 de los 100 intervalos (es decir, 95 %) contendrían el valor real (parámetro) del promedio.

En resumen, tenemos que la fórmula para calcular los intervalos de confianza al 95 % para una media es:

$$\bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$$

Puede advertirse que el margen de error corresponde a una estimación en particular, en este caso de la media aritmética. El error no sería igual, por ejemplo, al estimar el promedio de calificaciones otorgadas al gobierno, con la misma muestra y nivel de confianza, pero con una variabilidad (desviación estándar) distinta. Por ello, pese a que la expresión es común, *no existe el margen de error de la encuesta*, porque cada estimación tiene su margen de error relacionado con su propia variabilidad.

2.4.2 Intervalo de confianza de un porcentaje

Ahora estimaremos el error de un porcentaje para ilustrar cómo el margen de error es diferente del calculado para el promedio. En la misma encuesta del ejemplo de las edades, se encontró que 64 % de las personas están muy satisfechas con la vida. Para

saber qué tan precisa es la estimación respecto al nivel real de satisfacción (parámetro) se calcula el margen de error.

Para el porcentaje, p, el margen de error se estima de forma similar al caso de la media, al dividir la desviación estándar entre la raíz cuadrada del tamaño de la muestra. La diferencia está en que la desviación estándar de un porcentaje es $\sqrt{p * (100 - p)}$, y se estima como $\sqrt{\hat{p} * (100 - \hat{p})}$, donde $\hat{p}$ es el porcentaje muestral ya que desconocemos el porcentaje real p.[1] Con $\hat{p} = 64\,\%$, la desviación estándar es $\sqrt{64 * (100 - 64)} = 48$.

También se utilizan los valores z en los errores de los porcentajes, dado un teorema que asegura que la distribución binomial –la que describe proporciones y porcentajes– se aproxima a la distribución normal (en muestras grandes). Entonces:

$$\textit{Margen de error del porcentaje} = \pm z * \sqrt{\frac{p * (100 - p)}{n}}$$

Para el ejemplo, el margen de error del porcentaje desconocido de satisfacción con la vida, con un 95 % de confianza, se estima así:

$$\pm 1.96 * \sqrt{\frac{64 * (100 - 64)}{1500}} = \pm 2.4$$

- Límite inferior: 64 % – 2.4 = 61.6 %
- Límite superior: 64 % + 2.4 = 66.4 %

Por lo tanto, el intervalo [61.6 %, 66.4 %] contiene el valor real de personas muy satisfechas con la vida, con una confianza del 95 % (de 100 muestras del mismo tamaño, 95 intervalos contendrían el valor real del porcentaje).

En síntesis, el intervalo de confianza al 95 % para el porcentaje se calcula:

$$\hat{p} \pm 1.96 * \sqrt{\frac{p * (100 - p)}{n}}$$

Tanto para la media como para el porcentaje, puede observarse que el error disminuye cuanto mayor sea la muestra, si la desviación estándar se mantiene constante. Sin embargo, la relación no es directamente proporcional, pues hay una raíz cuadrada para el tamaño de la muestra. Esto implica que en cada aumento en el tamaño de la muestra,

[1] Para calcular el margen de error de una proporción (valores de 0 a 1) se estima la desviación estándar con $\sqrt{\hat{p} * (1 - \hat{p})}$, donde $\hat{p}$ es la proporción muestral.

la reducción del error es cada vez más menor. En términos prácticos, un exceso de muestra puede transformarse en un incremento de costos en el trabajo de campo que no se traduce en una mejoría real en la precisión. La figura 2.3 ilustra el margen de error del ejemplo del porcentaje, calculado para distintos tamaños de muestra. Aumentar la muestra de 300 a 600 disminuye considerablemente el error de ±5.4 a ±3.8 puntos porcentuales. Duplicarla de 600 a 1200 también tiene un impacto visible, al pasar de ±3.8 a ±2.7 puntos porcentuales. Sin embargo, incrementarla de 1200 a 2400 es poco beneficioso, pues el error disminuye de ±2.7 a ±1.9. En resumen, hay un punto en el que no resulta eficiente –precisión versus costo– aumentar la muestra.

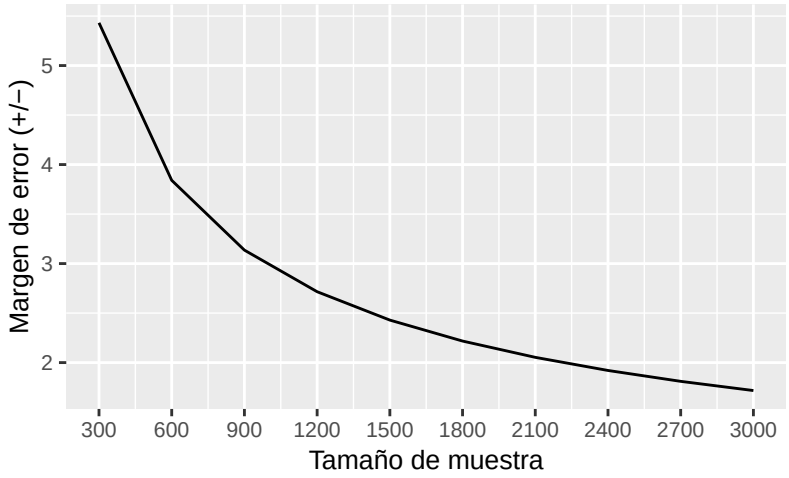


Figura 2.3 Relación entre el tamaño de muestra y el margen de error

Además, independientemente del nivel de confianza, es importante notar que en las fórmulas del margen de error *el tamaño de la población no influye*. Solo en el caso de poblaciones finitas (*i. e.*, pequeñas) se introduce el tamaño de la población N en el factor de corrección por finitud que multiplica el error estándar (Kish, 1965, pp. 43-45):

$$\sqrt{1 - \frac{n}{N}}$$

Puede verse que conforme crece N, el factor de corrección tiende a 1, y esto lo convierte en trivial en poblaciones grandes que se asumen infinitas, como lo ejemplifica la figura 2.4, para distintos tamaños de población, con una muestra constante de 100. En cambio, si la muestra es igual a la población, $n = N$, el factor es 0 y no hay error de muestreo. Entonces, contrario a lo que la intuición supondría, los errores de la inferencia no dependen de qué tan grande es una población. Se podrían recolectar muestras de

1500 personas tanto en Costa Rica como en Estados Unidos, donde los tamaños de la población son distintos (aproximadamente 5 millones y más de 300 millones de personas, respectivamente), pero si la variabilidad es la misma (iguales desviaciones estándar), los errores serían idénticos para un nivel de confianza.

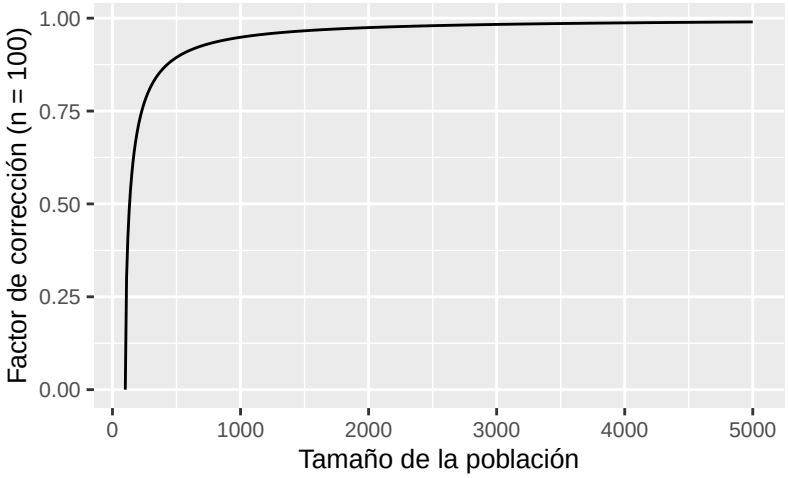


Figura 2.4 Relación entre el tamaño de población y el factor de corrección por finitud

2.5 Interpretación de las pruebas de hipótesis

Un marco analítico de inferencia estadística muy conocido –pero también criticado– se denomina pruebas de hipótesis o pruebas de significancia. Estas permiten examinar conjeturas sobre valores reales o parámetros –paradójicamente, sin llegar a conocer nunca estas cantidades–. En esta sección se ofrecen las herramientas básicas para entender las pruebas, a sabiendas de que la teoría es más elaborada, por lo que se pueden consultar otros textos especializados (Hernández Rodríguez, 2015; Wackerly *et al.*, 2002).

Las pruebas de hipótesis siguen tres pasos básicos: formulación de las hipótesis nula y alternativa, cálculo del estadístico de prueba y obtención del valor p. Para empezar, hay que tener presente que una hipótesis estadística no es lo mismo que una hipótesis teórica en una disciplina como la ciencia política. La segunda proviene de la teoría que pretende explicar el fenómeno sustantivo (elecciones, partidos, comportamiento votante, conflictos internacionales, etc.); mientras que la primera responde a reglas de procedimiento estadístico.

Por ejemplo, se ha comentado sobre el peso creciente de las personas jóvenes en la política. Podría suponerse que el electorado es joven, menor a 44 años y medio, en promedio. Esta es una hipótesis teórica. Desde el punto de vista estadístico, se establecería una *hipótesis nula*, según la cual $\mu = 44.5$, es decir, el valor real del promedio de edades es 44.5 años. Esta es una hipótesis estadística. Las hipótesis nulas en estadística se basan siempre en proposiciones de igualdad. Se examinaría, entonces, la evidencia a favor de que la media sea igual a 44.5. Complementariamente, se formula una *hipótesis alternativa* que postula que $\mu < 44.5$, es decir, que la edad promedio es menor a 44 años y medio, la cual coincide con la hipótesis teórica de la juventud del electorado.

El segundo paso es calcular el estadístico de prueba que, en el caso de la media, es:

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

donde el $\bar{x}$ es una media estimada a partir de una muestra, μ_0 es el valor hipótetico del parámetro según las hipótesis, σ es la desviación estándar y n es el tamaño de la muestra aleatoria que obtenemos para contrastar la hipótesis. Puesto que el estadístico calculado sigue una distribución normal estándar, se aprovechan las propiedades de esta conocida distribución para calcular probabilidades. Esto significa que no es necesario que la variable en sí (edad, en el ejemplo) se distribuya de forma normal; es el estadístico de prueba el que sigue esta distribución.

Según los datos del ejemplo del promedio de edad que hemos trabajado, la media estimada ($\bar{x}$) es 43.5, el parámetro hipotético (μ_0) es 44.5 y la desviación muestral (s) es 17.3 (en lugar de σ que es desconocida). Entonces, el estadístico calculado se obtiene así:

$$\frac{43.5 - 44.5}{\frac{17.3}{\sqrt{1500}}} = -2.239$$

El tercer paso consiste en calcular la probabilidad –a partir de la distribución normal estándar– de obtener el estadístico calculado o uno más extremo *al asumir que la hipótesis nula es cierta*. Es clave tener presente que no calculamos la probabilidad de que la hipótesis nula sea cierta (o falsa), sino que concluimos sobre la evidencia a favor de la hipótesis nula con base en la probabilidad de obtener el estadístico calculado o uno más extremo. Esta probabilidad (con valores entre 0 y 1, como cualquier otra) se denomina valor p (en inglés, *p-value*).

Se puede interpretar el valor p como la evidencia a favor de la hipótesis nula. Cuanto mayor sea el valor p, mayor es el respaldo para la hipótesis nula. Cuanto menor sea el valor p, menor es la evidencia para la hipótesis nula y mayor la seguridad con la cual esta se puede rechazar, en favor de la hipótesis alternativa.

Los valores p se obtienen fácilmente en R. En el caso de prueba de hipótesis de una media, se calculan las probabilidades en la distribución normal estándar con la función `pnorm()`; si se quiere examinar el área izquierda de la curva, con `pnorm(..., lower.tail=TRUE)`, o el área derecha, con `pnorm(..., lower.tail=FALSE)`. Las áreas corresponden a la probabilidad de obtener el estadístico calculado o uno más extremo, a la izquierda o a la derecha, como se ejemplifica en la figura 2.5.

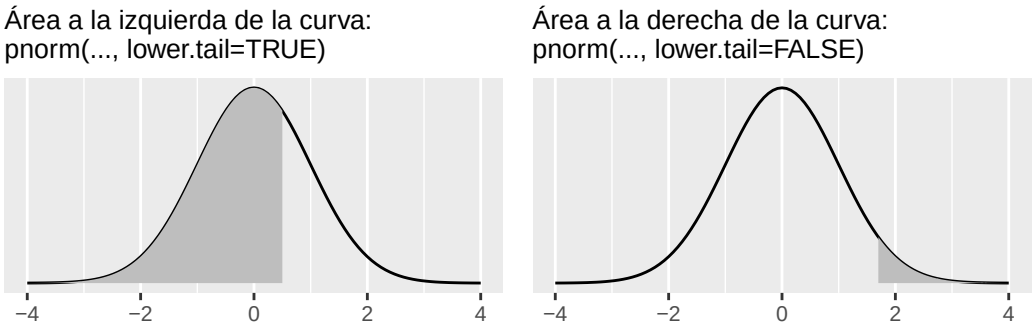


Figura 2.5 Probabilidades como áreas de la distribución normal estándar

Para la prueba de hipótesis del ejemplo, donde la hipótesis nula define que el promedio de edades es 44.5 años y la alternativa que $\mu < 44.5$, se necesita un valor p para la cola izquierda con `pnorm(..., lower.tail=TRUE)`, ya que la hipótesis alternativa indica que la media hipotética es menor al valor.

```
z<-(43.5-44.5)/(17.3/sqrt(1500))
z
```

```
## [1] -2.238719
```

```
pnorm(z, lower.tail=TRUE)
```

```
## [1] 0.01258711
```

Se obtiene el valor $p = 0.013$ (redondeado a tres decimales, que es lo usual), el cual implica que el estadístico calculado (o uno más extremo a la izquierda) es poco probable, si se asume verdadera la hipótesis nula. En términos gráficos, el área a la izquierda de

la curva es pequeña (figura 2.6). Es decir, hay poca evidencia a favor de la hipótesis nula de que el promedio de edades es 44.5 años, por lo que la hipótesis nula se puede rechazar en favor de la alternativa de que la media es menor a 44.5.

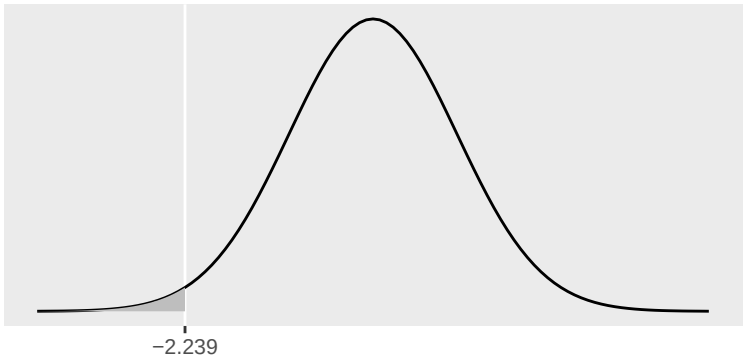


Figura 2.6 Probabilidad de obtener el estadístico -2.239 o uno más extremo a la izquierda

Si se formula, como hipótesis alternativa, que $\mu > 44.5$, o sea, que la población del padrón electoral está más bien envejecida, la hipótesis nula es también $\mu = 44.5$ y se obtiene el mismo estadístico calculado, -2.239. El valor p se obtiene también con -2.239; sin embargo, al proponer la hipótesis alternativa que la media es mayor, se considera la cola derecha de la distribución normal con `pnorm(..., lower.tail=FALSE)`.

```
pnorm(z, lower.tail=FALSE)
```

```
## [1] 0.9874129
```

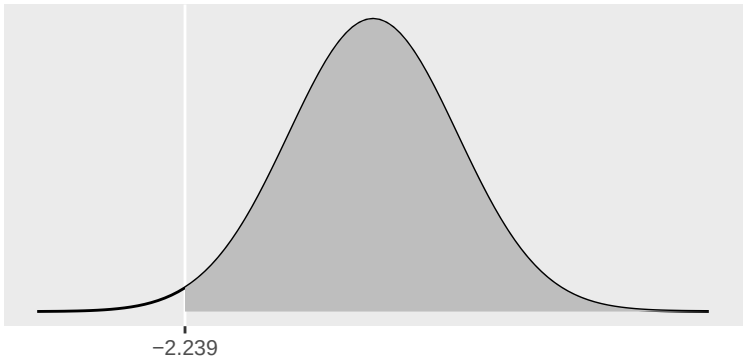


Figura 2.7 Probabilidad de obtener el estadístico -2.239 o uno más extremo a la derecha

Para la segunda hipótesis alternativa, el valor p es 0.987. Entonces, la probabilidad de obtener el estadístico calculado -2.239 o uno más extremo (a la derecha de la distribución)

es muy alta (al área bajo la curva es muy grande, figura 2.7), al asumir que la hipótesis nula es cierta. La evidencia, por lo tanto, favorece la hipótesis nula de que $\mu = 44.5$. En otras palabras, si la hipótesis nula es cierta, es muy probable obtener el valor calculado. La mejor decisión, en este caso, es no rechazar la hipótesis nula.

En ambos casos obtenemos un valor p con el cual tenemos que decidir si se rechaza o no la hipótesis nula –no se decide sobre la hipótesis alternativa–. A mayor valor p, mayor respaldo para la hipótesis nula. Ahora bien, ¿cómo saber que un valor p es suficientemente pequeño para poder rechazar la hipótesis nula? Para ello existen *niveles significancia*: umbrales que se escogen *a priori* para decidir cuándo se rechaza o no la hipótesis nula. Estos niveles de significancia suelen a ser bajos, 0.01, 0.05, 0.10, de forma que se reduzca la posibilidad de rechazar equivocadamente una hipótesis nula que es cierta (el denominado error tipo 1). Por ejemplo, con un umbral de 0.40, si obtenemos un valor p de 0.35 deberíamos rechazar la hipótesis nula; sin embargo, tenemos una probabilidad alta de 0.35 de obtener el estadístico calculado o uno más extremo al asumir la hipótesis nula verdadera. Rechazar la hipótesis nula, como sugeriría el umbral de 0.40, sería muy arriesgado.

El nivel de significancia de 0.05 (o 5 %) es el más común y resulta equivalente al nivel de confianza de 95 % que ya conocíamos.[2] Cuando el valor p resulta mayor a 0.05 no se rechaza la hipótesis nula, mientras que si es menor se rechaza. Este es un criterio útil, pero no debe adoptarse como una regla férrea, pues es solo una convención.

En resumen, para las pruebas de hipótesis de una media con R:

- Se formulan las hipótesis nula y alternativa.
- Se obtiene el estadístico calculado.
- Si la hipótesis alternativa es valor real (μ) < valor hipotético (μ_0), se calcula el valor p con `pnorm(..., lower.tail=TRUE)`.
- Si la hipótesis alternativa es valor real (μ) > valor hipotético (μ_0), se calcula el valor p con `pnorm(..., lower.tail=FALSE)`.
- Un valor p pequeño (o menor al nivel de significancia escogido) permite rechazar la hipótesis nula de que el valor real es igual al valor hipotético ($\mu = \mu_0$), en favor de la hipótesis alternativa.

[2] Formalmente, los intervalos de confianza se definen como $P(\hat{\theta}_I \leq \theta \leq \hat{\theta}_S) = (1 - \alpha)$, donde $\hat{\theta}_I$ y $\hat{\theta}_S$ son los límites inferior y superior del parámetro desconocido θ, α es el nivel de significancia y $(1 - \alpha)$ es el nivel de confianza (Wackerly *et al.*, 2002, p. 380).

2.6 Comentarios finales

Se han visto, en este capítulo, aspectos fundamentales de la inferencia estadística clásica, particularmente los conceptos de parámetro, estimador, sesgo, consistencia, suficiencia y eficiencia, así como el cálculo de los intervalos de confianza de medias y porcentajes. Se introdujeron también las pruebas de hipótesis para la media, aunque hay pruebas para otros estadísticos, así como pruebas de dos colas que, por espacio, no se abarcaron. De cualquier forma, los valores p se usarán en los próximos capítulos para evaluar resultados estadísticos; con estas posteriores aplicaciones se apreciaría su utilidad para obtener conclusiones relevantes.

Sin embargo, debe conocerse que no hay una teoría estadística única ni universalmente aceptada. Existe, además, un paradigma denominado bayesiano, el cual, aunque basado en un viejo teorema del reverendo Thomas Bayes (c. 1701-1761), ha visto crecer sus aplicaciones con el desarrollo computacional, necesario para muchos de sus cálculos.

La perspectiva bayesiana permite incorporar creencias previas, que se actualizan con datos, para obtener probabilidades posteriores, a través de una elegante y poderosa fórmula:

$$Probabilidad\ posterior \propto Datos * Probabilidad\ previa$$

Es decir, en el paradigma bayesiano, el conocimiento posterior sobre los parámetros es proporcional a la multiplicación de los datos (formalmente, la función de verosimilitud) por el conocimiento previo sobre los parámetros (ver Jackman, 2004).

La estadística clásica y bayesiana parten de diferentes premisas fundamentales que las hacen irreconciliables o inconmensurables (Kuhn, 1996). Para la estadística bayesiana, los parámetros son variables aleatorias que siguen una distribución de probabilidad, no valores fijos y desconocidos como en la estadística clásica. Las divergencias conceptuales, además, se manifiestan en controversias intelectuales. Ronald Fisher, pionero de la estadística clásica, caracterizó la perspectiva bayesiana –entonces conocida como teoría de la probabilidad inversa– como un "error" que "detiene el progreso hacia la precisión de conceptos estadísticos" (Fisher, 1922, p. 311, traducción propia). A su vez, la estadística bayesiana confronta con dureza las bases teóricas de la inferencia clásica. Dennis Lindley, por ejemplo, denunció la desconexión de la estadística frecuentista con la teoría de probabilidad, lo que hizo que la primera desarrollase "conceptos incoherentes como los intervalos de confianza" (Lindley, 2000, p. 311, traducción propia).

Una crítica recurrente del enfoque bayesiano hacia el frecuentista es la asignación arbitraria de un nivel de significancia. Se cuestiona por qué las conclusiones sustantivas de una investigación serían distintas entre un resultado con un valor p de 0.049 y otro de 0.051 y por qué los niveles de confianza convencionales (*e. g.*, 95 %) tendrían la misma interpretación en disciplinas distintas, como agronomía (donde la teoría nació) y ciencia política (donde la teoría se importó) (Gelman y Stern, 2006; Gill, 1999).

Ante las abundantes críticas (no solo bayesianas), la Asociación Estadounidense de Estadística (*American Statistical Association*) ha alertado sobre cuál debe ser la interpretación correcta de los valores p (Wasserstein *et al.*, 2019). Un valor p, vimos, no es la probabilidad de que ocurra la hipótesis nula, sino la probabilidad de obtener el estadístico calculado o uno más extremo al asumir la hipótesis nula como cierta. Además, aunque los valores p son útiles, no deben ser el único criterio para tomar una decisión o llegar a una conclusión científica. Debe tomarse en cuenta la calidad del estudio, las magnitudes de los efectos, la importancia práctica del hallazgo y la posibilidad de replicar el resultado en estudios posteriores (Cox, 1982; Wasserstein *et al.*, 2019).

Otro aspecto polémico de la estadística clásica es el hecho de que los tamaños de muestra afectan las pruebas de hipótesis. De forma automática, grandes muestras producen valores p pequeños, es decir, resultados estadísticamente (¿artificialmente?) significativos debido a que, con mayor tamaño de muestra, menor error. Esta consecuencia es problemática cuando se trabaja con bases de datos masivas, como miles o millones de publicaciones en Facebook y en Twitter/X. Con este volumen de datos es preferible utilizar métodos y algoritmos de *big data* y minería de datos.

2.7 Ejercicios

1. El Instituto Nacional de Estadística y Censos (INEC, 2020) estimó que el ingreso per cápita promedio de Costa Rica en julio de 2020 es 326483 colones, con un margen de error de ±11423 colones calculado al 95 % de confianza. Construya el intervalo de confianza para el promedio e interprételo.

2. Según la Encuesta Nacional de Hogares, de julio de 2020, el 26.2 % de los hogares se categorizan como pobres en Costa Rica, con un margen de error de ±1.2 puntos porcentuales, calculado al 95 % de confianza (INEC, 2020). Construya el intervalo de confianza para el porcentaje e interprételo.

3. En una escala de 0 a 10, las personas encuestadas por el Centro de Investigación y Estudios Políticos (CIEP, 2020) calificaron al Tribunal Supremo de Elecciones de Costa Rica con una nota promedio de 6.6, con una desviación estándar muestral de 2.5. A sabiendas de que la muestra es de 927 personas, calcule el margen de error al 95 %, construya el intervalo de confianza e interprételo.

4. Según una encuesta preelectoral de enero hecha por el CIEP (2018), la intención de voto para el Partido Restauración Nacional era de 17 % y para el Partido Acción Ciudadana de 11 %, con una muestra de 798 personas decididas a votar. Calcule los márgenes de error al 95 %, construya los intervalos de confianza para ambas estimaciones e interprételos.

5. Comparando los intervalos de confianza obtenidos en el punto 4, ¿qué le parece que se puede concluir?

6. Los resultados del censo de 2011 del INEC (2012) dicen que la población total de Costa Rica es de 4301712 personas, pero no se indica ningún margen de error. Explique por qué.

7. Alguien postula que el estado de las carreteras de su país es deficiente y que las personas lo valorarían con una nota por debajo de 6.3. Otra persona es más bien optimista y cree que la media estaría por encima de 6.4. Una encuesta, aplicada a 600 personas, encuentra que en promedio la ciudadanía da una calificación de 6.2 con una desviación estándar de 1.2.

    a) Pruebe, calculando el valor p con R, la hipótesis nula de que la media en la valoración de las carreteras es igual a 6.3 versus la hipótesis alternativa de que es menor a 6.3, como aducía el crítico.

    b) Pruebe, con R, la hipótesis nula de que la media en la valoración de las carreteras es igual a 6.4 versus la hipótesis alternativa de que es mayor a 6.4, como opinaba la voz optimista.

Capítulo 3

Comparación de medias

3.1 Comparación de dos medias: prueba *t*

¿Es el ingreso promedio de las mujeres igual al de los hombres? ¿Varía el grado de satisfacción con la democracia entre personas jóvenes y no jóvenes? ¿Gastan más los gobiernos de izquierda o los de derecha? ¿Es mayor el desempleo en dictaduras o en democracias? Estas son preguntas de investigación relevantes en la ciencia política. Puede distinguirse que en estos casos se está cuestionando sobre una variable métrica (ingresos, grado de satisfacción con la democracia, gasto público y desempleo) en dos grupos distintos o independientes. Para ello, en la primera parte de este capítulo veremos cómo comparar dos promedios utilizando la prueba t de Student.

El nombre de esta prueba proviene del trabajo de William S. Gosset (1876-1937), matemático y químico que, al trabajar en procesos de producción en la cervecería Guinness, desarrolló métodos novedosos para la teoría estadística. La compañía, sin embargo, restringía las publicaciones científicas de sus empleados, para evitar la difusión de información confidencial. Por esta razón, Gosset se vio obligado a publicar sus resultados bajo el seudónimo "Student" o estudiante en inglés (Salsburg, 2001).

Para introducir el análisis de dos medias utilizaremos un diseño experimental. Los experimentos, como vimos en el capítulo 1, implican la asignación aleatoria de unidades experimentales en grupos de tratamiento. En ciencia política los experimentos permiten evaluar hipótesis de psicología política, comportamiento electoral, cooperación y toma de decisiones, entre otras aplicaciones (McDermott, 2002).

Un tema relevante en la experimentación son los efectos de los medios de comunicación. Consideremos un experimento de laboratorio para evaluar si un anuncio televisivo tiene efectos en la valoración de un candidato. Una muestra de 60 personas se divide aleatoriamente en dos grupos. El grupo A recibe información de un candidato (hombre): su nombre, su fotografía y un resumen de sus propuestas políticas. Al grupo B se le brinda exactamente la misma información del candidato, pero además se le expone a un video de campaña (*spot*) de medio minuto, el cual incluye imágenes del candidato y reitera las propuestas de campaña que el grupo recibió por escrito; es decir, el video no incluye información adicional. El tratamiento experimental es, por lo tanto, la exposición audiovisual (estudios experimentales similares, pero más elaborados que este ejemplo didáctico, se encuentran en: Kahn y Geer, 1994; Noggle y Kaid, 2000).

En ambos grupos se aplica un cuestionario en el que se pregunta con qué probabilidad de 0 a 100 votaría por este candidato. Datos simulados (por medio de R) recrean este posible experimento aleatorizado y los resultados se muestran en la figura 3.1. En este experimento encontramos que las notas del grupo B tienden a ser mayores que las del grupo A, aunque hay bastante dispersión en cada uno. Si calculamos los promedios, obtenemos que en el grupo B es 65.0, con desviación estándar 15.8, y en el grupo A es 49.1, con desviación estándar 18.2. ¿Será esta diferencia un efecto del video?

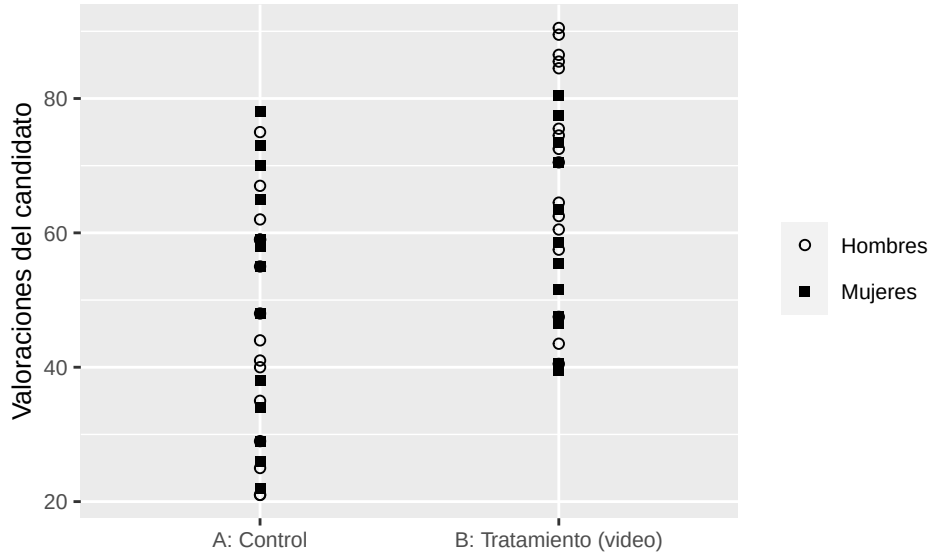


Figura 3.1 Resultados del experimento sobre efectos mediáticos (dos grupos)

Teóricamente no se puede responder esta pregunta, debido al problema fundamental de la inferencia causal (Holland, 1986): es imposible conocer un resultado con la causa presente y sin ella. En el ejemplo, no podemos saber simultáneamente qué nota le habría dado el grupo B al candidato de no haber recibido el estímulo audiovisual (el tratamiento); solo observamos uno de los dos estados en cada grupo: con tratamiento (grupo B) o sin este (grupo A). No observamos al grupo B sin tratamiento, ni al grupo A con tratamiento; son contrafácticos. Sin embargo, la estadística moderna, por medio de los experimentos aleatorizados, permite hacer inferencia causal comparando no ambos mundos, que son imposibles de observar de forma simultánea, sino contrastando promedios entre los grupos. La diferencia entre los promedios, o el valor esperado de la diferencia, se denomina el *efecto promedio del tratamiento* (Wooldridge, 2010, p. 905).

Para contrastar los dos promedios se puede proceder de distintas maneras. Primero, siguiendo el enfoque del capítulo 2, es posible construir intervalos de confianza para cada promedio y comparar si existen diferencias significativas, más allá de los márgenes de error de cada uno. Visualmente, la figura 3.2 representa los promedios en barras con los intervalos de confianza al 95 %. El intervalo [42.6, 55.6] contiene el valor real del promedio en el grupo A, mientras que el intervalo [59.3, 70.7] contiene el valor real del promedio en el grupo B. Nótese que los intervalos no se traslapan. En otras palabras, hay diferencias entre los promedios, más el margen de error de cada uno.

Aunque esta conclusión se obtiene de manera rápida y efectiva al observar el gráfico y calcular los intervalos de confianza, tiene una desventaja: cada intervalo se construye al 95 %, lo cual implica que un 5 % de los intervalos no contienen el valor real. Este se denomina el error tipo 1. Puesto que contamos con dos intervalos, tenemos dos errores tipo 1 de 5 %. Es decir, hay una inflación del error tipo 1.

Para evitar esta inflación del error tipo 1, es recomendable calcular un único intervalo de confianza, en este caso, el intervalo de confianza al 95 % de la diferencia entre los promedios, el cual se obtiene con la siguiente fórmula:

$$(\bar{x}_A - \bar{x}_B) \pm 1.96 * \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

donde $\bar{x}_A$ y $\bar{x}_B$ son los promedios de cada grupo, s_A y s_B las desviaciones estándar y n_A y n_B los tamaños de muestra. En este cálculo, teóricamente debería utilizarse el valor t en lugar del valor z, como lo hace R en los procedimientos más adelante. No obstante, ya

que la distribución t se aproxima a la distribución normal conforme aumenta el tamaño de muestra, en la práctica solamente al trabajar con muestras muy pequeñas habría diferencias relevantes y se preferiría el valor t para construir los intervalos de confianza. Por ello (y por simplicidad didáctica) se considera el valor z en el ejemplo.

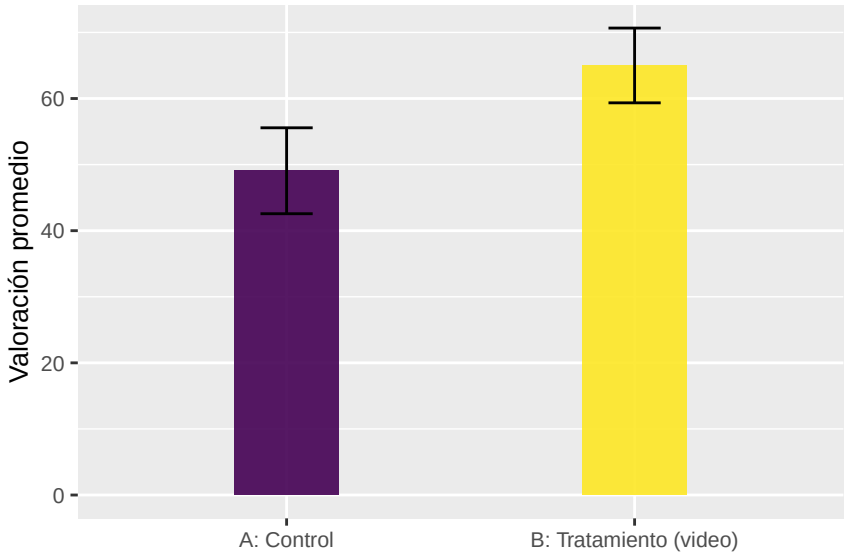


Figura 3.2 Resultados promedio del experimento sobre efectos mediáticos (dos grupos)

Con los datos del experimento tenemos:

$$(49.1 - 65.0) \pm 1.96 * \sqrt{\frac{18.2^2}{30} + \frac{15.8^2}{30}} = -15.9 \pm 8.6$$

Por lo tanto, el intervalo [-24.5, -7.3] contiene el valor real de la diferencia en los promedios. Es lógico que, si los promedios son iguales, su diferencia sea cero. Puesto que el intervalo *no* contiene el valor cero, entonces se puede concluir que los promedios no son iguales, según el intervalo de confianza del 95 %. En términos sustantivos, este procedimiento nos permite concluir que hay diferencias entre los promedios de valoración del candidato en ambos grupos experimentales.

Finalmente, con los mismos datos experimentales, podemos aplicar una prueba de hipótesis tal y como Student –mejor dicho, William Gosset– ideó. Su prueba t examina qué tan probable es obtener el estadístico resultante, o uno más extremo, al asumir que los promedios son iguales. Para ello, se formula la hipótesis nula de que el promedio del grupo de control (A) es igual al promedio del grupo de tratamiento (B), frente a la

hipótesis alternativa de que son diferentes. En R, la prueba t se realiza con la función `t.test()`, en la que se indica primero la variable respuesta y luego el tratamiento.

```
t.test(respuesta~tratamiento, data=experimento)
```

```
##
## Welch Two Sample t-test
##
## data: respuesta by tratamiento
## t = -3.6233, df = 56.914, p-value = 0.0006211
## alternative hypothesis: true difference in means between group A:
Control and group B: Tratamiento (video) is not equal to 0
## 95 percent confidence interval:
## -24.739282 -7.127385
## sample estimates:
## mean in group A: Control mean in group B: Tratamiento (video)
## 49.06667 65.00000
```

Para el ejemplo experimental, la salida muestra un intervalo de confianza de la diferencia entre los promedios muy similar al calculado con la fórmula manualmente. Además, brinda un valor $p = 0.0006$ que, al ser muy pequeño, denota poca evidencia a favor de la hipótesis nula que indica que los promedios son iguales. Es decir, podemos rechazarla. Esto nos permite concluir que la diferencia entre los promedios es estadísticamente significativa ($p < 0.001$) y que las personas que recibieron el estímulo audiovisual valoran mejor el candidato que las personas que recibieron únicamente información textual. Lo que es aún más importante: puesto que se trata de un experimento aleatorizado y controlado, las diferencias entre los promedios son el efecto causal. Es decir, el experimento nos permite concluir que la mejor valoración en el grupo del *spot* es un efecto del *spot* y no de otros factores.

Podría aducirse que hay una tercera variable confusora: el género de la persona que responde. ¿Cómo saber que es el *spot* y no las actitudes políticas de género las que producen el resultado? (Sobre la relación entre estereotipos de género y evaluaciones de candidatos y candidatas, consultar: Dolan, 2010; Taylor-Robinson y Geva, 2023). Alguien podría aducir que el resultado se debe a que en el grupo B hay mayor cantidad de hombres, los cuales valoraron mejor al candidato masculino por ser votantes del mismo género, mientras que las votantes mujeres evaluaron peor al candidato porque es

hombre y no mujer, independientemente del tratamiento. Sin embargo, si se examina la figura 3.1, mujeres y hombres están equitativamente distribuidos en ambos grupos debido a la asignación aleatoria: el grupo A tiene 15 hombres y 15 mujeres, mientras el grupo B cuenta con 16 hombres y 14 mujeres. Es decir, están balanceados, lo cual nos permite descartar la explicación alternativa de género, así como otras hipótesis provenientes de variables relacionadas con las personas participantes (*e. g.*, edad, nivel educativo, actitudes políticas y otras). Esta es la fortaleza metodológica del experimento: *los efectos confusores se cancelan gracias a la aleatorización del tratamiento.*

A pesar de sus inigualables fortalezas metodológicas, no todos los experimentos son factibles, ya que es imposible, en ocasiones, aleatorizar un tratamiento. No podríamos asignar sistemas electorales al azar en distintos países para observar sus efectos en los sistemas de partidos. Por ello se recurre a datos observacionales, como los obtenidos en encuestas. En estos estudios también se puede aplicar la prueba t de Student y calcular el intervalo de confianza de la diferencia entre promedios, aunque las conclusiones no permiten la inferencia causal.

Por ejemplo, en la encuesta de noviembre de 2020, realizada por el Centro de Investigación y Estudios Políticos (CIEP, 2020), se indagaron valoraciones alrededor de la Defensoría de los Habitantes de Costa Rica. Se preguntó, específicamente, qué nota le daría de 0 a 10 a esta institución. El promedio total es 5.7. Ahora bien, nos interesa conocer si existen diferencias significativas entre hombres y mujeres, para lo cual planteamos la hipótesis nula de que el promedio de valoración de la Defensoría de los Habitantes entre mujeres es igual al promedio entre hombres.[3] Para ello cargamos la base de datos en R, luego de haber definido un directorio de trabajo que la aloja. Los resultados de la encuesta están en un archivo con extensión ".dta" que corresponde a bases de datos en Stata.[4] En R no hay problema para abrir este tipo de archivo; simplemente se utiliza el paquete `haven` y la función `read_dta()`. Asigno la base de datos a un objeto que denomino `ciep`:

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
```

[3]La medición de identidad de género y sexo es compleja y no necesariamente binaria. La categorización mujer/hombre que se sigue aquí depende de la disponibilidad de los datos.

[4]Stata es un programa comercial de análisis estadístico muy utilizado en ciencia política y en otras disciplinas como economía y demografía.

Para evitar inflar el error tipo 1, calcularemos primero el intervalo de confianza para la diferencia de medias. A fin de obtener los promedios, desviaciones estándar y los tamaños de muestras por grupos, empleamos el siguiente código, que utiliza el paquete `dplyr`, parte de `tidyverse`, para calcular promedios por grupos:

```
library(tidyverse)
nota_dh_sexo<-ciep%>%
  group_by(sexo)%>%
  summarize(Media=mean(nota_dh, na.rm=TRUE),
            DevEst=sd(nota_dh, na.rm=TRUE),
            Muestra=sum(!is.na(nota_dh)))
nota_dh_sexo
```

```
## # A tibble: 3 x 4
##   sexo        Media DevEst Muestra
##   <dbl+lbl>   <dbl>  <dbl>   <int>
## 1  0 [Mujer]   6.16   2.55     459
## 2  1 [Hombre]  5.30   2.70     436
## 3 NA           6.33   1.15       3
```

Los resultados indican que entre las mujeres el promedio de valoración es 6.16, mientras que entre los hombres es 5.30. En la muestra, las medias son evidentemente diferentes, pero queremos saber si en la población también lo son. Por ello, calculamos el intervalo de confianza de la diferencia con los datos de mujeres y hombres:

$$(6.16 - 5.30) \pm 1.96 * \sqrt{\frac{2.55^2}{459} + \frac{2.70^2}{436}} = 0.86 \pm 0.34$$

Encontramos que el intervalo [0.52, 1.20] al 95 %, que contiene el valor real de la diferencia en la valoración de la Defensoría de los Habitantes entre mujeres y hombres, excluye el cero, de modo que las medias son estadísticamente diferentes. En otras palabras, se descarta que los promedios sean iguales.

En segunda instancia, aplicamos la prueba t con la función `t.test()`, la cual requiere que incluyamos primero la variable métrica, luego la categórica que define los grupos y, por último, la base de datos.

```
t.test(nota_dh~sexo, data=ciep)
```

```
##
## Welch Two Sample t-test
##
## data: nota_dh by sexo
## t = 4.8841, df = 882.95, p-value = 1.233e-06
## alternative hypothesis: true difference in means between group 0
and group 1 is not equal to 0
## 95 percent confidence interval:
## 0.5135639 1.2036015
## sample estimates:
## mean in group 0 mean in group 1
## 6.159041 5.300459
```

Los resultados muestran un valor p muy pequeño: 0.000001233 o, en notación científica, 1.233 por 10^{-6}. Por ende, podemos rechazar la hipótesis nula de igualdad de medias. En conclusión, las mujeres valoran significativamente mejor a la Defensoría de los Habitantes ($p < 0.001$) que los hombres, ya que los promedios son estadísticamente diferentes.

Debe agregarse que la prueba t, tal y como se aplicó, asume que las variancias (o sea, las desviaciones estándar al cuadrado) no son iguales. De hecho, en los estadísticos descriptivos se observa que las desviaciones estándar muestrales son diferentes, aunque –como cualquier estimación– contienen un margen de error. Si se quisiera asumir que las variancias son iguales, basta indicarlo en la misma función, con `var.equal=TRUE`, de lo contrario, R asume que son diferentes.

En este ejemplo, el resultado es casi idéntico, lo cual demuestra que el supuesto de variancias iguales no es problemático. De cualquier forma, la opción más segura es asumir que las variancias son distintas.

```
t.test(nota_dh~sexo, data=ciep, var.equal=TRUE)
```

```
##
## Two Sample t-test
##
## data: nota_dh by sexo
## t = 4.891, df = 893, p-value = 1.189e-06
```

```
## alternative hypothesis: true difference in means between group 0
and group 1 is not equal to 0
## 95 percent confidence interval:
## 0.514060 1.203105
## sample estimates:
## mean in group 0 mean in group 1
## 6.159041 5.300459
```

3.2 Comparación de más de dos medias: análisis de variancia (Anova)

En la sección anterior estudiamos cómo comparar dos medias y probar si son estadísticamente iguales o no. La pregunta pendiente es cómo comparar más de dos medias. La respuesta es mediante el análisis de variancia (Anova, por sus siglas en inglés) desarrollado por Ronald Fisher (1890-1962). Los aportes de Fisher son vastos, más allá del análisis de variancia. Buena parte de la inferencia clásica proviene de él, incluidos los valores p. Sin embargo, quizás su legado más importante –porque trasciende la teoría estadística– es el diseño de experimentos aleatorizados controlados.

En la sección anterior vimos que los experimentos permiten calcular los efectos causales, ya que, con la asignación aleatorizada de tratamientos, se eliminan los efectos confusores. Esta idea original de aleatorizar un tratamiento es producto del trabajo de Fisher en la finca de Rothamstead en Inglaterra.

Como relata Salsburg (2001, pp. 46-48), en Rothamstead se solía probar fertilizantes, para lo cual se dividía el terreno en dos partes y se aplicaba un tipo de fertilizante en cada una. Ahora bien, las secciones del terreno no eran homogéneas: variaban en su inclinación, composición del suelo, irrigación y demás características. Con el procedimiento usual para dividir los terrenos no se podía determinar cuál era el efecto del fertilizante porque este se confundía con los otros factores (incluidos algunos no observables) que caracterizaban a cada parte del terreno. La solución de Fisher consistió en dividir el terreno en muchas secciones pequeñas y asignar el tratamiento al azar: en unas se utilizaba un tipo de fertilizante A, en otras otro tipo de fertilizante B y en otras ninguno (grupo control). Puesto que un orden aleatorio, por definición, no sigue ningún patrón, las características del suelo, irrigación y demás se balancean en las distintas partes del terreno y sus efectos

se cancelan entre sí. Las diferencias observadas en el resultado se pueden atribuir, por lo tanto, al fertilizante o tratamiento.[5]

¿Cómo analizar los diferentes efectos promedio de los fertilizantes A y B y compararlos con el grupo de control? Para ello se utiliza el análisis de variancia de Fisher. Retomemos el ejemplo experimental de efectos mediáticos de la sección anterior, solo que, esta vez, en lugar de comparar dos grupos, uno con información y otro con un *spot*, comparamos un grupo que recibe un video positivo del candidato, otro que recibe un video negativo del candidato, hecho por un contrincante político, y un grupo control que no recibe ningún estímulo audiovisual, solo la misma información escrita que recibieron los otros dos grupos.

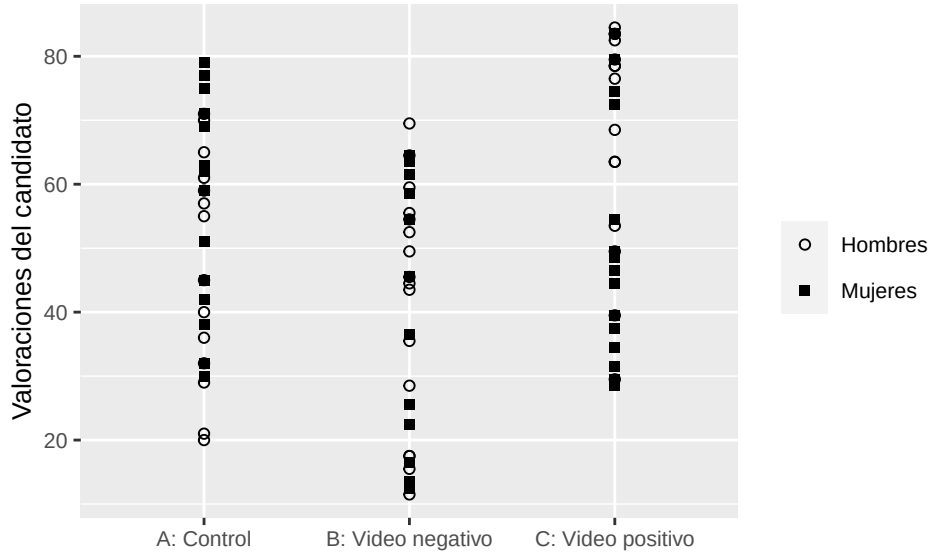


Figura 3.3 Resultados del experimento sobre efectos mediáticos (tres grupos)

La figura 3.3 muestra los resultados de este experimento simulado. A simple vista, aparecen diferencias entre los grupos. Las personas que recibieron el *spot* positivo tienden a valorar mejor al candidato que el grupo control, mientras que quienes visualizaron la propaganda negativa lo valoran peor. Mujeres y hombres aparecen distribuidos en los tres grupos de forma aleatoria, por lo que el género no incide en el resultado.

[5]Según el historiador de la estadística, Stephen Stigler (1978, p. 248), el más antiguo uso de la aleatorización que se ha conocido es un experimento de Charles Peirce y Joseph Jastrow, con naipes, 50 años antes de Fisher. Sin embargo, Fisher no solo propuso la aleatorización, sino también métodos analíticos para diseños experimentales variados.

Si se calculan los promedios para cada grupo (figura 3.4), se observan diferencias entre estos promedios grupales, particularmente entre los grupos del video negativo y del positivo; es la denominada *variabilidad entre los grupos*, mientras que el gráfico anterior (figura 3.3) mostraba la *variabilidad dentro de los grupos*. El propósito del análisis de variancia es encontrar cuándo la variación *entre* los grupos es significativamente mayor a la variación *dentro* de los grupos.

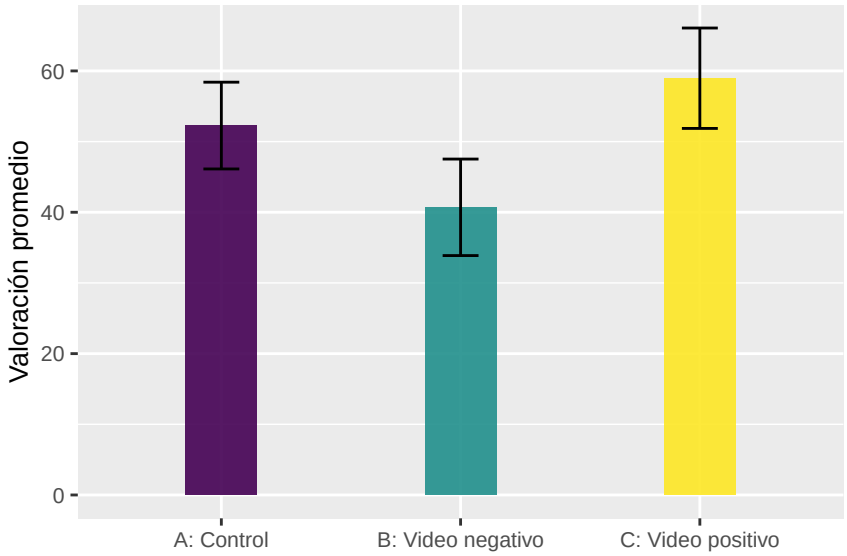


Figura 3.4 Resultados promedio del experimento sobre efectos mediáticos (tres grupos)

La hipótesis nula del Anova sostiene que los promedios son iguales; la hipótesis alternativa, que al menos un promedio es diferente. Es decir, la prueba se limita a evidenciar que existen variaciones entre los grupos, pero no establece entre cuáles medias hay diferencias significativas. En R, el Anova se realiza con la función `aov()`, indicando primero la variable métrica, luego los grupos incluidos con `factor()` y los datos. Los resultados se despliegan mejor si se incluye en la función `summary()`.

```
summary(aov(respuesta~factor(tratamiento), data=experimento))
```

```
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## factor(tratamiento)  2   5123  2561.7   7.304 0.00117 **
## Residuals           87  30514   350.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor p en este análisis es 0.001, menor a los umbrales convencionales que se utilizan para rechazar la hipótesis nula. Podemos concluir que los promedios no son iguales y que el tratamiento (video televisivo) tiene un efecto estadísticamente significativo en los promedios de valoración del candidato entre las personas participantes del experimento, *pero no se concluye cuál promedio es mayor o menor*, solo que al menos uno es diferente.

Al igual que la prueba t, el Anova puede aplicarse con datos observacionales, como los de una encuesta; la diferencia con el experimento radica en que no podemos interpretar los resultados como efectos causales. Retomando la encuesta del CIEP de noviembre de 2020, queremos analizar ahora si la valoración promedio de la Asamblea Legislativa (calificada en una escala de 0 a 10) varía según los niveles educativos de las personas. Para ello, analizamos de forma exploratoria las variaciones:

```
nota_al_educarec<-ciep%>%
  group_by(educarec)%>%
  summarize(Media=mean(nota_al, na.rm=TRUE),
            DevEst=sd(nota_al, na.rm=TRUE),
            Muestra=sum(!is.na(nota_al)))
nota_al_educarec
```

```
## # A tibble: 3 x 4
##   educarec              Media DevEst Muestra
##   <dbl+lbl>             <dbl>  <dbl>   <int>
## 1 1 [Primaria o menos]  4.13   2.91      206
## 2 2 [Secundaria]        4.64   2.40      377
## 3 3 [Universitaria]     4.5    2.20      340
```

Los resultados descriptivos indican que las calificaciones otorgadas a la Asamblea Legislativa varían poco entre personas con educación primaria o menos (4.13), personas con educación secundaria (4.64) y personas con educación universitaria (4.50). Realizamos, entonces, el Anova para probar la hipótesis nula de que los promedios son iguales. En la función `aov()` primero debemos incluir la variable métrica, luego la variable categórica que define los grupos, la cual definimos como `factor()`, y por último la base de datos que utilizamos. No incluir la variable categórica como factor puede llevarnos a resultados erróneos. Encontramos que el valor p resultante del Anova es 0.058. Si utilizamos un nivel de significancia de 0.05, que es conveniente para el análisis de una encuesta con una muestra grande, no podemos rechazar la hipótesis nula de que las medias son

iguales. En otras palabras, no detectamos variaciones en la valoración pública de la Asamblea Legislativa según niveles educativos. Esta calificación es baja (menor a 5) independientemente del nivel de instrucción alcanzado.

```
summary(aov(nota_al~factor(educarec), data=ciep))
```

```
##                    Df Sum Sq Mean Sq F value Pr(>F)
## factor(educarec)    2     34  17.212   2.851 0.0583 .
## Residuals         920   5554   6.037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 46 observations deleted due to missingness
```

3.3 Comentarios finales

Es común trabajar con promedios y muchas veces es necesario comparar las medias entre grupos independientes para determinar efectos causales o simplemente para describir variaciones. Metodológicamente, los experimentos aleatorizados controlados permiten lo primero, pues, al asignar un tratamiento al azar, los factores confusores se balancean y no influyen en el resultado. Con datos observacionales es posible comparar promedios para describir variaciones entre grupos.

Como vimos en este capítulo, con la prueba t se comparan dos medias; y con el análisis de variancia o Anova, más de dos medias. Este último método tiene, sin embargo, una limitante, ya que, aunque analiza si las medias son estadísticamente iguales, no concluye cuáles de estas son diferentes. Para ello una opción es construir intervalos de confianza para cada media, con la desventaja, antes discutida, de que infla el error tipo 1, porque cada estimación al 95 % implica que 5 de 100 intervalos (5 %) no incluyen el valor real. Otra alternativa es calcular pruebas t entre las medias, pero esto también infla el error tipo 1. Además, con muchos grupos, el número de pruebas t entre medias aumentaría exponencialmente. La alternativa más adecuada es aplicar contrastes múltiples (consultar Agresti y Franklin, 2013, pp. 693-695). Aunque la diversidad de pruebas de contrastes múltiples dificulta escoger la apropiada y, con muchos grupos, la interpretación de las comparaciones puede resultar abrumadora, los contrastes múltiples no conllevan la inflación del error tipo 1.

3.4 Ejercicios

1. Con la base de datos "CIEPnoviembre2020.dta", determine si existen diferencias entre mujeres y hombres (variable `sexo`) en torno a la valoración de la Asamblea Legislativa de Costa Rica (variable `nota_al`) mediante la construcción del intervalo de confianza para la diferencia de medias, con un nivel de confianza de 95 %.

2. Respalde el resultado del punto 1 con la prueba *t*.

3. Con la base de datos "CIEPnoviembre2020.dta", considere si existen diferencias significativas en la valoración de la Defensoría de los Habitantes (variable `nota_dh`) según el nivel educativo de las personas (variable `educarec`). Utilice un nivel de significancia de 0.01.

4. Repita el análisis de la valoración de la Asamblea Legislativa entre mujeres y hombres, punto 1, pero utilizando el análisis de variancia (Anova) en lugar de la prueba *t*. ¿Qué puede concluirse?

Capítulo 4

Medidas de asociación

4.1 Introducción

Entre la descripción y la relación causal, existe un nivel intermedio al examinar dos o más variables: la asociación. Con ella se entiende que una característica varía de forma conjunta con otra, pero que esto no basta para garantizar una relación de tipo causal, donde la causa precede al efecto.

Por ejemplo, Robert Putnam (1993), en *Making Democracy Work*, teoriza la relación entre las normas de confianza (que denomina capital social) con la cooperación interpersonal. Cuando las personas en una comunidad confían entre sí, es más probable que cooperen, ya que no prevén comportamientos oportunistas de las demás y evitarán ellas mismas ser oportunistas. A la vez, la cooperación alimenta la confianza entre las personas. Existe, pues, una asociación recíproca entre capital social y cooperación, sin una secuencia causal definida.

Las medidas de asociación abundan. Hay coeficientes específicos según los niveles de medición de las variables (nominal, ordinal y métricas) y sus combinaciones. En este capítulo se estudiarán dos métodos de asociación bivariada muy comunes: primero, la prueba *chi* cuadrado con el coeficiente V de Cramer para dos variables categóricas nominales; segundo, la correlación lineal o r de Pearson para dos variables métricas.

4.2 Tablas cruzadas entre variables categóricas

Antes de aplicar la prueba *chi* cuadrado, es necesario construir tablas cruzadas entre variables categóricas. Recordemos que este tipo de variables definen fenómenos por medio de atributos cualitativos o categorías: votantes y no votantes, grupos etarios (jóvenes, adultos medios y adultos mayores), familias de partidos políticos (socialdemócratas, conservadores, verdes y derecha radical), por mencionar algunos ejemplos. Cuando se tienen variables categóricas, estas pueden medirse nominalmente (sin orden) u ordinalmente (con orden). En este capítulo examinaremos variables categóricas en general, sin poner atención al orden. Es decir, aplica para variables nominales y para ordinales en las que el orden se obvia.

Podemos combinar dos variables categóricas en una tabla cruzada o tabla de contingencia. Estas tablas deben contener categorías exhaustivas y mutualmente excluyentes: cada observación debe pertenecer a una categoría de cada variable (exhaustividad) y solamente a una categoría (exclusividad). Por ejemplo, con los datos de la encuesta de noviembre de 2020 realizada por el Centro de Investigación y Estudios Políticos (CIEP, 2020), se construye una tabla cruzada entre las variables categóricas valoración de la gestión del gobierno (positiva y negativa)[6] y género (mujer y hombre).

Primero, se carga en R la base de datos, la cual asigno a un objeto que llamo `ciep`.

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
```

Luego, se utiliza `table()` para crear la tabla cruzada siguiendo el orden: filas, columnas. En otras palabras, la primera variable que se indique será la que defina las filas y la segunda las columnas. El resultado es simplemente un conteo en las cuatro casillas.

```
table(ciep$sexo, ciep$gestionrec)
```

```
##
##       0   1
##   0 319 182
##   1 305 145
```

[6]Por simplicidad, se construyen dos categorías basadas en las respuestas originales: positiva incluye "muy buena", "buena" y "regular"; y negativa, "mala" y "muy mala".

Cada variable se puede explorar para identificar qué significa cada valor 0 y 1, en este ejemplo, de una base de datos de Stata, por medio de la función `attr(..., "labels")`.

```
attr(ciep$sexo, "labels")
```

```
##           Mujer          Hombre Otro no binario
##               0               1               2
```

```
attr(ciep$gestionrec, "labels")
```

```
## Negativa Positiva
##        0        1
```

Podemos observar que la encuesta codificó, por un lado, las mujeres con 0, los hombres con 1 y personas no binarias con 2; por otro, la valoración negativa del gobierno con 0 y la valoración positiva con 1. Es decir, los datos muestran que 319 mujeres calificaron negativamente la gestión gubernamental, 182 mujeres la calificaron de forma positiva, 305 hombres calificaron negativamente la gestión gubernamental y 145 la calificaron de forma positiva (no hay personas declaradas no binarias en la base de datos). Con esta información podemos renombrar las filas y las columnas, al asignar la tabla a un objeto que denomino `tabla1` y al utilizar las funciones `rownames()` y `colnames()` para modificar los encabezados de la tabla, sin cambiar las variables originales.

```
tabla1<-table(ciep$sexo, ciep$gestionrec)
rownames(tabla1)<-c("Mujeres", "Hombres")
colnames(tabla1)<-c("Gestión negativa", "Gestión positiva")
```

```
tabla1
```

```
##
##           Gestión negativa Gestión positiva
##   Mujeres              319              182
##   Hombres              305              145
```

La tabla generada presenta cifras absolutas que, en realidad, son poco útiles. No permiten identificar si las mujeres, que conforman la mayor parte de la muestra, aprueban la gestión gubernamental más que los hombres o si las personas que aprueban la gestión tienden a ser hombres o mujeres. Es con valores relativos, como los porcentajes, que se pueden conocer estas relaciones.

Para obtener porcentajes, debemos combinar varias funciones de R. Primero, utilizamos `prop.table()` para calcular proporciones. Si se define `prop.table(..., 1)`, las proporciones se calculan para cada fila. En cambio, con `prop.table(..., 2)` se calculan por columna. Segundo, multiplicamos por 100 para obtener porcentajes. Tercero, añadimos `round()` para redondear decimales. Empecemos con porcentajes, por fila, con un decimal:

```
round(prop.table(tabla1, 1)*100, 1)
```

```
##
##           Gestión negativa Gestión positiva
##   Mujeres             63.7             36.3
##   Hombres             67.8             32.2
```

Podemos interpretar que, entre las mujeres, 63.7 % califica negativamente la gestión del gobierno y 36.3 % la califica positivamente. Entre los hombres, 67.8 % evalúa de forma positiva la gestión gubernamental y 32.2 % de forma negativa. Comparado entre los grupos, hay evidencia de que la valoración no varía mucho entre género.

Si queremos calcular los porcentajes por columnas, especificamos:

```
round(prop.table(tabla1, 2)*100, 1)
```

```
##
##           Gestión negativa Gestión positiva
##   Mujeres             51.1             55.7
##   Hombres             48.9             44.3
```

Vemos que entre las personas que valoran negativamente la gestión del gobierno, 51.1 % son mujeres y 48.9 % son hombres. Entre las personas que la valoran positivamente, 55.7 % son mujeres y 44.3 % son hombres.

Por último, podemos calcular porcentajes en relación con el total de la muestra sin especificar ningún número en `prop.table()`:

```
round(prop.table(tabla1)*100, 1)
```

```
##
##           Gestión negativa Gestión positiva
##   Mujeres             33.5             19.1
##   Hombres             32.1             15.2
```

Los porcentajes por total se interpretan así: 33.5 % de la muestra son mujeres que valoran negativamente la gestión del gobierno, 19.1 % son mujeres que valoran positivamente la gestión, 32.1 % son hombres que valoran negativamente la gestión del gobierno y 15.2 % son hombres que valoran positivamente la gestión. Este cálculo de porcentajes según el total es útil para construir tipologías o perfiles, más que para encontrar variaciones entre variables.

4.3 Prueba *chi* cuadrado

Siguiendo con el ejemplo anterior, ¿hay alguna relación entre el género y la valoración de la gestión del gobierno? Como es común en el análisis estadístico, luego de observar los datos tabulados, se busca aplicar una prueba que permita inferir sobre la asociación entre las variables. Para responder esta pregunta, utilizamos la prueba *chi* cuadrado de independencia para dos variables categóricas, ideada por el estadístico Karl Pearson (1857-1936), aunque Ronald Fisher (el mismo que vimos en el capítulo 3 como pionero de los diseños experimentales) corrigió el cálculo.

Antes de adentrarse en la prueba estadística, es necesario aclarar primero qué se entiende por independencia. Existe independencia estadística entre dos variables categóricas cuando una no varía respecto a la otra. En tablas cruzadas, la independencia se observa cuando los porcentajes de una variable no cambian entre las categorías de la otra variable. Con datos hipotéticos, el cuadro 4.1 ejemplifica independencia de la variable y respecto a la variable x, pues no importa la categoría de x, el porcentaje de y es igual. La conclusión sería la misma con porcentajes calculados por columnas.

Cuadro 4.1 Ejemplo de independencia entre dos variables categóricas

	Categoría y1	Categoría y2	Total
Categoría x1	120 (60 %)	80 (40 %)	200 (100 %)
Categoría x2	90 (60 %)	60 (40 %)	150 (100 %)

Con base en este principio de independencia, la prueba *chi* cuadrado considera si una tabla hipotética de variables independientes es significativamente diferente de la tabla observada de datos reales. La hipótesis nula de la prueba establece que las variables son estadísticamente independientes; mientras la hipótesis alternativa indica las variables no son estadísticamente independientes (*i. e.*, son dependientes o están asociadas).

Como vimos en el capítulo 2, un valor p denota la evidencia a favor de la hipótesis nula. Si es cercano a cero, la evidencia es pobre y podemos rechazar la hipótesis nula que asume que las variables son independientes. Si el valor p que se obtiene es mayor al nivel de significancia predeterminado (por ejemplo, 0.05), entonces la hipótesis nula no se debe rechazar y se concluye que las variables son independientes.

La lógica del método es la siguiente. La teoría de probabilidades dice que dos eventos A y B son independientes si la probabilidad de que ocurran A y B es igual a la probabilidad de A multiplicada por la probabilidad de B (ver Hernández Rodríguez, 2015). En otras palabras, si A y B son independientes:

$$Pr(A\ y\ B) = Pr(A) * Pr(B)$$

Para ejemplificar, imaginemos que se quiere saber cuál es la probabilidad de obtener "2" en un dado de seis caras y luego "5". Estos eventos son independientes entre sí, pues un resultado no depende del otro. Puesto que son independientes, entonces:

$$Pr(2\ y\ 5) = Pr(2) * Pr(5) = \frac{1}{6} * \frac{1}{6} = 0.03$$

Regresando a los datos de la tabla cruzada entre género y valoración de la gestión del gobierno, si se piensa que las probabilidades de ser mujer y la probabilidad de valorar de manera negativa la gestión del gobierno son independientes, entonces la probabilidad de ocurrencia de ambos eventos debe ser igual a su producto:

$$Pr(mujer\ y\ negativa) = Pr(mujer) * Pr(negativa) =$$

$$\frac{mujeres}{mujeres + hombres} * \frac{negativas}{negativas + positivas} =$$

$$\frac{319 + 182}{319 + 182 + 305 + 145} * \frac{319 + 305}{319 + 305 + 182 + 145} =$$

$$\frac{501}{951} * \frac{624}{951} = 0.527 * 0.656 = 0.346$$

Por lo tanto, tenemos que la probabilidad de ser mujer en la muestra es 0.527, la probabilidad de tener una opinión negativa de la gestión del gobierno es 0.656 y la probabilidad de ser mujer y tener una opinión negativa de la gestión del gobierno es 0.346, si los eventos fuesen independientes. Si se multiplica esta última probabilidad

por el total de personas en la muestra (0.346 por 951), calculamos el valor esperado: 329 (redondeado). En otras palabras, si las variables fuesen independientes, habría 329 mujeres con una opinión negativa del gobierno. Al repetir el mismo cálculo para las otras tres celdas, se construye el cuadro 4.2 de valores observados (los que existen) y valores esperados (los que resultan asumiendo independencia).

Cuadro 4.2 Valores observados y esperados

	Gestión negativa		Gestión positiva	
	Observado	Esperado	Observado	Esperado
Mujeres	319	328.732	182	172.268
Hombres	305	295.268	145	154.732

El siguiente paso es calcular el estadístico *chi* cuadrado (χ^2) utilizando los valores observados y esperados:

$$\chi^2 = \sum_{j=1}^{k} \frac{(valor\ observado_j - valor\ esperado_j)^2}{valor\ esperado_j}$$

Con los datos del ejemplo se calcula:

$$\chi^2 = \frac{(319 - 328.732)^2}{328.732} + \frac{(182 - 172.268)^2}{172.268} + \frac{(305 - 295.268)^2}{295.268} + \frac{(145 - 154.732)^2}{154.732} = 1.771$$

¿Cómo interpretar el estadístico *chi* cuadrado calculado? Si la diferencia entre valores observados y esperados es pequeña, es decir, si los valores observados se asemejan a los valores esperados que asumen independencia, el estadístico *chi* cuadrado calculado sería pequeño –de hecho, si fuesen iguales, el estadístico sería cero–. Esto indicaría que las variables son independientes, pues los valores reales son similares a los valores esperados cuando se asume independencia. Pero si la diferencia entre observados y esperados es amplia, porque los valores reales son muy distintos de los valores esperados que asumen independencia, entonces el *chi* cuadrado resultante sería grande. Por lo tanto, un estadístico *chi* cuadrado grande es un indicio de no independencia.

Ahora bien, ¿cómo saber si es suficientemente grande para poder rechazar la hipótesis nula? Para ello calculamos el valor p en R con la función `chisq.test()`. Indicamos el argumento `correct=FALSE` para que no aplique la corrección de Yates, que R hace por defecto, aunque es apta solo en muestras pequeñas para tablas de dos columnas con dos filas. Sin embargo, al trabajar tablas más grandes (con más de dos filas o dos columnas),

la corrección de Yates es irrelevante y no se especificaría nada en el argumento `correct`. Para el ejemplo sería:

```
chisq.test(tabla1, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla1
## X-squared = 1.7707, df = 1, p-value = 0.1833
```

El resultado contiene el valor p de la prueba, pero además el estadístico *chi* cuadrado que calculamos manualmente (ambos coinciden). En este ejemplo, el valor p es 0.1833, por lo que no podemos rechazar la hipótesis nula y concluimos que las variables son independientes. Este resultado coincide con el hecho de que los valores esperados (que asumen independencia) son muy similares a los observados. Además, cuando se comparan los porcentajes (por filas o por columnas) no hay variaciones relevantes entre los grupos. En síntesis, de forma sustantiva se concluye que la aprobación gubernamental es independiente del género.

4.4 Coeficiente V de Cramer

Si bien es útil para descartar independencia, la prueba *chi* cuadrado no permite concluir qué tan asociadas están dos variables. En otras palabras, no mide la magnitud de la relación. El coeficiente V de Cramer cubre esta carencia. La V de Cramer es una medida no paramétrica, es decir, que no asume una distribución de probabilidad, la cual se calcula con base en el estadístico *chi* cuadrado. Sus valores teóricos están entre 0 y 1, donde 0 significa que no hay relación y 1 que la relación es perfecta. La fórmula es:

$$V\ de\ Cramer = \sqrt{\frac{\chi^2}{n * min(filas - 1, columnas - 1)}}$$

donde χ^2 es el estadístico *chi* cuadrado y n es el total de observaciones en la tabla, el cual se multiplica por el mínimo entre el número de filas menos uno y el número de columnas menos uno, es decir, se multiplica por la resta que genere el número menor. Por ejemplo, para una tabla de cinco filas y tres columnas, n se multiplica por dos, ya que 3 columnas – 1 = 2 es menor a 5 filas – 1 = 4.

Siguiendo con el ejemplo de aprobación gubernamental y género, el coeficiente se obtendría así:

$$V\ de\ Cramer = \sqrt{\frac{1.771}{951 * min(2-1, 2-1)}} = \sqrt{\frac{1.771}{951 * 1}} = 0.043$$

En R podemos calcular la V de Cramer manualmente de esta forma:

```
sqrt(chisq.test(tabla1, correct=FALSE)$statistic/
     sum(tabla1)*min(c(nrow(tabla1)-1,ncol(tabla1)-1)))
```

```
##  X-squared
## 0.04315043
```

También, podemos utilizar el paquete `rcompanion` que tiene ya programada la función `cramerV()`:

```
library(rcompanion)
cramerV(tabla1)
```

```
## Cramer V
##  0.04315
```

El resultado, por cualquiera de las tres formas, es 0.043. Este valor indica una muy baja relación entre género y valoración de la gestión del gobierno, lo cual coincide con la independencia que concluye la prueba *chi* cuadrado. Efectivamente las variables no están asociadas.

4.5 Correlación lineal de Pearson

En la sección anterior se estudió cómo relacionar dos variables categóricas. En cambio, para establecer la relación entre dos variables métricas, una de las medidas de asociación más utilizadas es el coeficiente de correlación lineal de Pearson, conocido así por su creador Karl Pearson, el mismo que propuso la prueba *chi* cuadrado.

El coeficiente de correlación indica qué tan fuerte es una relación lineal entre dos variables métricas. Para ello calcula la covariancia entre las dos variables, es decir, cómo varían de forma conjunta, dividiendo entre las desviaciones estándar de cada variable. Esta división estandariza o elimina el efecto de las escalas de medición de las variables. Por ejemplo, si tenemos una variable con una escala de 0 a 100 y otra de 1 a 7, con la estandarización

podemos calcular la correlación sin problema. La fórmula de la correlación de Pearson (denotada r) entonces es:

$$r = \frac{covariancia_{xy}}{s_x * s_y}$$

donde s es la desviación estándar. El cálculo, de forma puntual, es:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

El coeficiente r de Pearson oscila entre -1 y 1. Los números positivos significan que cuanto mayor sea x, mayor es y (y viceversa). Los números negativos indican una relación inversa: a mayor x, menor y (y viceversa). Una correlación igual a 0 se interpreta como correlación nula, o sea, no hay asociación. Cuanto más cercana sea la correlación a 1 o -1, mayor la asociación. En consecuencia, si $r = 1$, la relación es positiva y perfecta; y si r = -1, la relación es negativa y perfecta.

Ahora bien, para valores entre 0 y 1 o entre 0 y -1, ¿cómo saber si la asociación es alta o baja? Aunque algunos textos ofrecen reglas prácticas para la interpretación de estos valores, la fuerza de la correlación depende más bien del área de estudio. Por ejemplo, en un campo donde las investigaciones previas han registrado correlaciones consistentemente fuertes entre dos variables, un r de 0.6 en un nuevo análisis podría considerarse decepcionantemente bajo. Por el contrario, en otro contexto, donde solo se han observado correlaciones tenues, un r de 0.3 puede asumirse alto.

Una forma práctica para determinar la correlación es la visualización gráfica. Como ejemplos, utilizaremos datos de las elecciones de Costa Rica en 2018, a nivel cantonal (Tribunal Supremo de Elecciones, 2018). Para los 81 cantones de esta elección, contamos con cifras de participación electoral y voto por partidos, entre otras. Iniciamos cargando en R la base de datos que está en formato Excel con el paquete `readxl`:

```
library(readxl)
eleccion18<-read_excel("eleccionesCR2018.xlsx")
```

Como primer ejemplo, calculamos la correlación entre el voto por el Partido Acción Ciudadana (PAC) en la primera ronda (elección de febrero) y el voto por el PAC en la segunda ronda o balotaje (abril). Si el voto es estable por cantón, esperaríamos un coeficiente de correlación alto. Para obtener la correlación Pearson, se utiliza la función `cor()`, pero cuidado: hay que indicar si existen casos faltantes, los `NA` en R. En el ejemplo

no hay faltantes, porque hay porcentajes de votos para todos los cantones, por lo que en la función se especifica la opción `cor(..., use="all.obs")`. En las encuestas, es más común encontrar los valores faltantes por la no respuesta. De registrarse valores `NA`, debe precisarse `cor(..., use="complete.obs")` para excluirlos.

Para el primer ejemplo de los datos electorales cantonales calculamos la correlación así:

```
cor(eleccion18$febPACporcentaje, eleccion18$abrilPACporcentaje,
    use="all.obs")
```

```
## [1] 0.8661064
```

El coeficiente resultante es 0.866, muy cercano a 1. Es decir, hay una correlación positiva y fuerte entre el voto por cantón en primera y segunda ronda. La figura 4.1 grafica la relación entre estas dos variables con una recta para reflejar el patrón lineal de la relación. Cuanto mayor es la correlación, más cercanos están las observaciones a la recta (en el siguiente capítulo veremos cómo calcular esta recta). No obstante, también hay cantones que se alejan de la recta. Es decir, la asociación no es perfecta. Si lo fuese, el r de Pearson sería 1 y todos los datos calzarían en la recta.

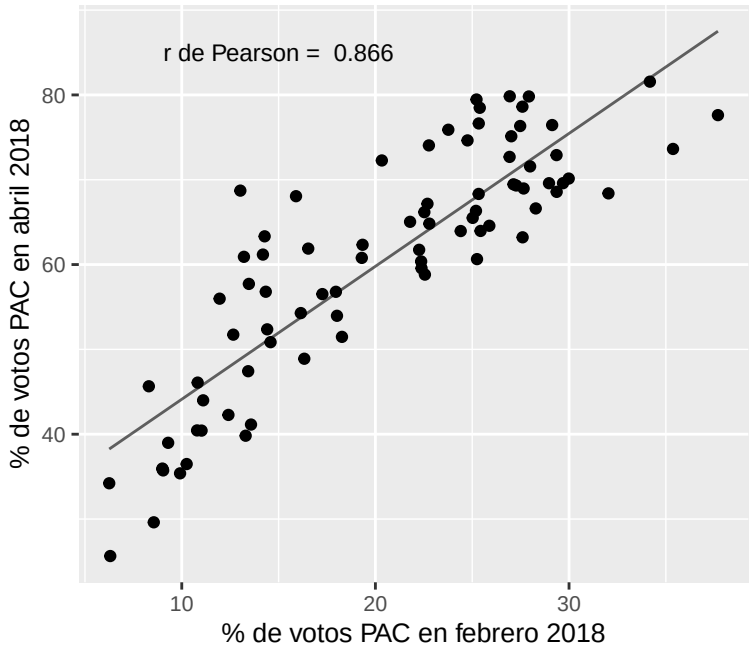


Figura 4.1 Correlación positiva en los votos por cantón en Costa Rica, 2018; datos del Tribunal Supremo de Elecciones (2018)

Por otro lado, calculamos la correlación entre los votos para el PAC en la primera ronda versus los votos para Restauración Nacional (RN). Ambos partidos encabezaron la primera vuelta en 2018, por lo que sus candidatos avanzaron a una segunda vuelta en la que el PAC prevaleció. Esta fue una campaña fragmentada entre muchos partidos, pero también polarizada ideológicamente. PAC y RN se destacaron por ubicarse en posiciones opuestas respecto al matrimonio igualitario, entre otros temas culturalmente divisores (Pignataro y Treminio, 2019). Podríamos suponer que existe una correlación negativa entre los apoyos electorales por cantón.

```
cor(eleccion18$febPACporcentaje, eleccion18$febRNporcentaje,
    use="all.obs")
```

```
## [1] -0.806489
```

Efectivamente, la correlación de -0.806 es alta y, a diferencia del primer ejemplo, negativa: a mayor porcentaje de votos para el PAC, menor porcentaje de votos para RN y viceversa. La figura 4.2 muestra esta relación negativa o inversa.

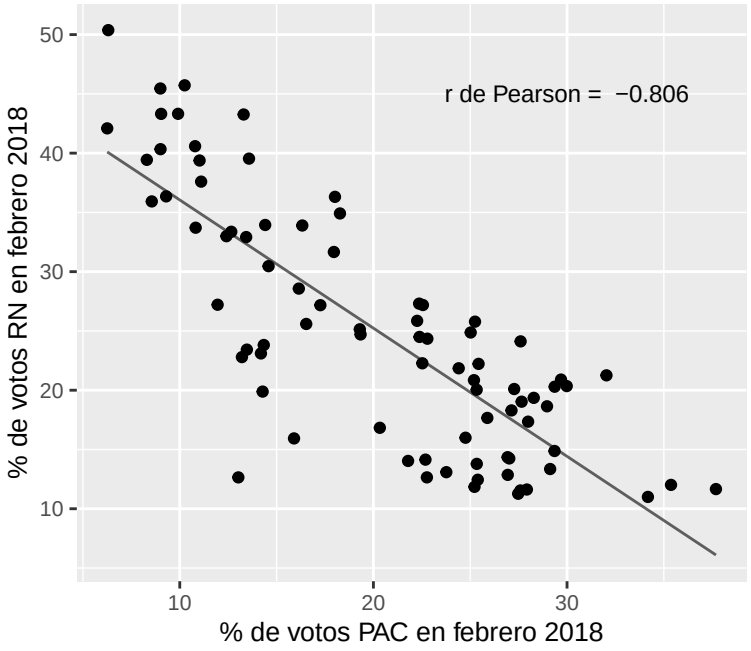


Figura 4.2 Correlación negativa en los votos por cantón en Costa Rica, 2018; datos del Tribunal Supremo de Elecciones (2018)

En los datos cantonales se registra también el número de juntas receptoras de votos por cantón, así como la participación electoral, entendida como el porcentaje de personas que fueron a votar en relación con el electorado empadronado. Podríamos probar si una mayor existencia de juntas se asocia con una mayor afluencia electoral por cantón. Aunque las escalas de medición son muy distintas (el número de juntas varía de 12 a 409 por cantón; y la participación, de 13.9 % a 79.3 %), podemos aplicar el coeficiente de correlación de Pearson, pues, como se indicó antes, esta medida estandariza las variables.

```
cor(eleccion18$juntas, eleccion18$febparticipacion,
    use="all.obs")
```

```
## [1] 0.07363294
```

Encontramos que el coeficiente es muy bajo: 0.074. Además, la visualización gráfica (figura 4.3) nos dice que no hay ningún patrón que relacione las variables. Cantones con muchas juntas de recepción de votos tienen porcentajes de participación similares a los de cantones con pocas juntas. Asimismo, hay un grupo de cantones con porcentajes de participación muy variables (de 50 % a 80 %), aunque todos tienen menos de 100 juntas. Concluimos, por lo tanto, que la correlación es nula.

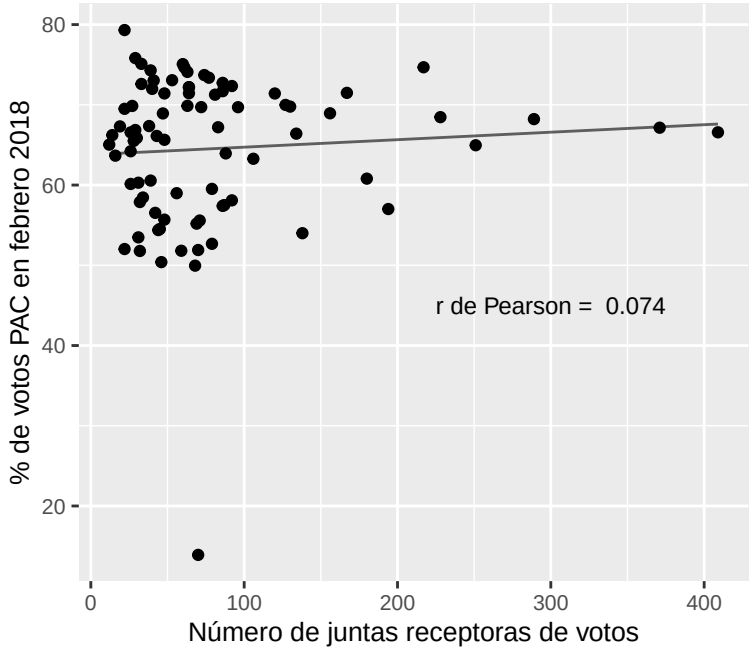


Figura 4.3 Correlación nula en los votos por cantón en Costa Rica, 2018; datos del Tribunal Supremo de Elecciones (2018)

4.6 Comentarios finales

Las medidas de asociación bivariadas son útiles para la exploración y la descripción de datos. Son parte del análisis inicial que se debería conducir siempre, antes de pasar a los métodos más sofisticados y al modelaje. Por esta razón, en el capítulo se estudiaron dos formas de analizar asociación entre variables: las tablas cruzadas con prueba *chi* cuadrado y el coeficiente V de Cramer, para dos variables categóricas; y el coeficiente r de Pearson, para dos variables métricas.

Existen otras medidas de asociación, más allá de la V de Cramer y la r de Pearson, las cuales no se abarcaron aquí. El repertorio incluye medidas específicas para variables categóricas con escala ordinal, como la *tau* de Kendall, y para correlaciones no lineales, como el coeficiente de Spearman. Estos se enmarcan en una rama de la estadística denominada no paramétrica.

4.7 Ejercicios

1. Analice la relación entre nivel educativo (variable `educarec`) y valoración de la gestión del gobierno (variable `gestionrec`) que están disponibles en la base de datos "CIEPnoviembre2020.dta" para responder las siguientes preguntas:

    a) ¿Qué porcentaje de personas con educación universitaria aprueba positivamente la gestión del gobierno?

    b) ¿Qué porcentaje de personas que valoran negativamente la gestión del gobierno tiene nivel educativo de primaria o menos?

    c) ¿Existe relación entre el nivel educativo y la valoración de la gestión gubernamental o son independientes?

    d) ¿Cuál es la magnitud de la relación entre nivel educativo y valoración de la gestión del gobierno?

2. Con la base de datos "eleccionesCR2018.xlsx", responda:

    a) ¿Qué relación existe entre la participación electoral por cantón en la primera ronda de 2018 (variable `febparticipacion`) y el balotaje (variable `abrilparticipacion`)?

b) La segunda ronda implicó apoyos al PAC y a Restauración Nacional desde los partidos que no pasaron al balotaje. ¿Cuál es la relación entre el apoyo electoral al Partido Liberación Nacional (PLN) en febrero de 2018 (variable `febPLNporcentaje`) y al PAC en abril del mismo año (variable `abrilPACporcentaje`)? ¿Qué se puede concluir?

3. Opcional. Con los datos de la tabla cruzada entre nivel educativo y valoración de la gestión del gobierno del ejercicio 1, calcule manualmente el valor del estadístico *chi* cuadrado y el coeficiente V de Cramer. Compare estos resultados con los obtenidos en R.

Capítulo 5

Regresión lineal simple

5.1 Introducción

Una de las explicaciones más poderosas sobre el voto se basa en la economía. Cuando el estado de la economía es positivo, el electorado recompensa al partido que ocupa el gobierno en las próximas elecciones; si la economía es negativa, castiga el gobierno saliente (Lewis-Beck y Stegmaier, 2013). Esta relación parece observarse en el caso de Costa Rica. La figura 5.1 grafica el porcentaje de votos válidos que obtuvo el partido que gobernaba (*i. e.*, el partido en la Presidencia de la República) en cada elección presidencial desde 1982 hasta 2018 (Tribunal Supremo de Elecciones, 2022) con el porcentaje de desempleo el año anterior a la elección (Comisión Económica para América Latina y el Caribe, 2022). El coeficiente de correlación Pearson (estudiado en el capítulo 4) es -0.4, lo cual señala una relación moderada y negativa donde a mayor desempleo (peor situación económica), menos votos recibe el partido del gobierno saliente (mayor castigo político).

La correlación entre desempleo y voto es clara. Sin embargo, más allá de esta asociación, ¿podemos estimar cuánto cambia el resultado electoral por cada punto porcentual que aumenta el desempleo? ¿Es esta relación estadísticamente significativa? Y, en general, ¿en qué porcentaje la economía explica los resultados electorales? Para responder estas tres preguntas, estudiaremos el análisis de regresión.

La regresión nació en el siglo XIX cuando Francis Galton (1822-1911), investigador en biometría y pionero de la estadística moderna, descubrió un fenómeno interesante al comparar las estaturas de personas: padres altos suelen tener hijos más bajos, mientras

padres bajos acostumbran tener hijos más altos. En conjunto, las poblaciones tienden hacia un promedio de estaturas, por lo que Galton llamó al fenómeno *regresión a la media*. Si la regresión a la media no existiera, las poblaciones humanas estarían compuestas por personas en extremo altas y bajas (Salsburg, 2001, pp. 12-13).

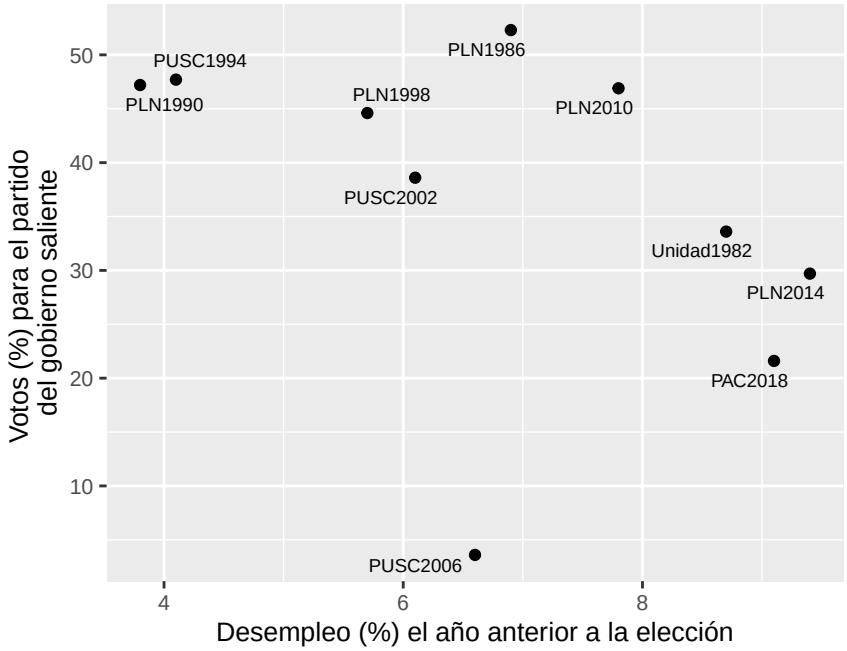


Figura 5.1 Desempleo y éxito electoral para el partido en el gobierno en Costa Rica; datos del Tribunal Supremo de Elecciones (2022) y de Comisión Económica para América Latina y el Caribe (2022)

Como notó Galton, la regresión caracteriza múltiples patrones naturales y comportamientos sociales que en ocasiones se malinterpretan en la forma de sesgos cognitivos. El psicólogo Daniel Kahneman (2012, p. 232) cuenta la historia del instructor de vuelo que prefiere no felicitar a los pilotos cuando ejecutan correctamente una maniobra, ya que luego la repiten mal; en cambio, cuando les grita por equivocarse, mejoran en el siguiente intento. El instructor asume que el castigo es beneficioso, mientras el elogio es perjudicial. En realidad, lo que actúa es la regresión a la media y no la reacción del instructor, pues cuando el cadete realizó el intento exitoso fue producto de la suerte (independientemente de que el instructor lo felicitara antes o no), mientras que a su mala ejecución seguía una buena porque su habilidad lo hacía regresar a su promedio de éxito. En otras palabras, regresa a su media.

Un ejemplo más cercano es cuando las notas de las evaluaciones en los cursos en el tiempo oscilan entre unas altas y otras bajas. El promedio de las notas refleja el rendimiento real de la persona estudiante, no los éxitos excepcionales ni los fallos espontáneos.

Aunque los ejemplos de regresión aparecen en la cotidianidad, su teoría es más reciente. Fue el ya mencionado Ronald Fisher quien sintetizó la teoría moderna de regresión, con base en la original intuición de Galton, la correlación de Pearson y la teoría de los errores de Carl Friedrich Gauss (Aldrich, 2005). Con Fisher, el modelo de regresión lineal se vincula con la teoría clásica de la inferencia estadística y con otros métodos como el análisis de variancia y la correlación lineal.

5.2 Conceptos

El análisis de regresión es un método flexible y potente. Puede responder a varios objetivos de investigación: describir, explicar y predecir. Además, permite analizar variables de todo tipo, tanto cuantitativas como cualitativas. El nivel de medición de la variable dependiente (métrica o categórica, nominal u ordinal) determina el modelo de regresión específico para el tipo de datos. En este libro, se estudian dos modelos:

- Regresión lineal, estimada por mínimos cuadrados ordinarios (en inglés, *ordinary least squares*, OLS), cuando la variable dependiente es métrica. También es llamado modelo gaussiano porque los errores asumen una distribución probabilística normal o gaussiana.
- Regresión logística, cuando la variable dependiente es categórica. Existen diversos modelos de regresión logística; en el capítulo 7 se introduce específicamente la regresión logística para variables dependientes binarias (esto es, de dos categorías).

En general, el interés al aplicar los modelos de regresión radica, primero, en conocer la magnitud del efecto promedio de una variable independiente x sobre la variable dependiente y, es decir, cuánto cambia y por cada cambio de x; segundo, en determinar si el efecto es estadísticamente significativo, más allá del error; y tercero, en establecer cuál es el poder explicativo o predictivo del modelo teórico.

Como se explicó en el capítulo 1, la investigación cuantitativa estudia los fenómenos políticos desde el enfoque de los efectos de causas (Mahoney y Goertz, 2006). Con la regresión no se intenta descubrir cuál es la causa del fenómeno, sino *cuáles son los efectos de las variables independientes*, escogidas según las teorías y los estudios previos,

sobre la variable dependiente. Así pues, a menos que realicemos un experimento, la relación entre x y y no se debe interpretar como "x causa y", sino como "x es un factor asociado a y" o "x tiene efectos sobre y".

Se pueden distinguir cinco etapas al aplicar un análisis de regresión (de cualquier tipo):

- *Especificación* o formulación de un modelo teórico que relaciona la variable independiente x con la variable dependiente y. En otras palabras, se define cuál es la función matemática que relaciona las variables.
- *Estimación* del valor de los parámetros relacionan las variables, utilizando los datos disponibles.
- *Evaluación* de la calidad del modelo, también llamada la bondad de ajuste.
- *Diagnósticos* sobre el cumplimiento de los supuestos del modelo.
- *Correcciones* en caso de incumplirse algún supuesto.

El libro se centra en las tres primeras etapas: especificación, estimación y evaluación. Las dos restantes pueden estudiarse en la bibliografía complementaria que se sugiere en el apéndice B.

5.3 Especificación

La regresión se basa en la idea de ajustar una función matemática a un conjunto de datos. Las funciones posibles son infinitas, así que el ajuste inicial debe responder al principio de parsimonia: ¿cuál es la función más simple que podemos escoger? Si repasamos los datos de elecciones en Costa Rica que introducen este capítulo (figura 5.1), se evidencia que las observaciones se aproximan a una recta descendente. Es decir, se podría utilizar una función lineal entre el desempleo y el resultado electoral. La función lineal se expresa en términos estadísticos como un modelo. Así, el modelo lineal simple se especifica de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

donde

- y_i corresponde a los valores de la variable dependiente según los datos;
- x_i son los valores de la variable independiente según los datos;
- β_0 es el parámetro (desconocido) denominado constante o intercepto;
- β_1 es el parámetro (desconocido) denominado pendiente;

- u_i son los errores –también llamados residuos– que se definen como la diferencia entre el valor observado en los datos, y_i, y el valor predicho por el modelo, $\hat{y}_i$, por lo que $u_i = y_i - \hat{y}_i$; la presencia del error implica que el modelo es probabilístico, no determinístico (ver capítulo 1);
- i es cada observación.

Para el caso del voto económico en Costa Rica, el modelo de regresión lineal se especificaría de la siguiente forma:

$$votos_i = \beta_0 + \beta_1 desempleo_i + u_i$$

donde $votos_i$ es la variable dependiente que está en función de la variable independiente $desempleo_i$, β_0 y β_1 son los parámetros desconocidos que se quieren estimar, u_i son los errores y cada i es una elección.

Con la regresión se estiman aquellos valores de los parámetros β_0 y β_1, llamados coeficientes de regresión, que minimicen los errores. Retomando la representación gráfica, se buscan parámetros que definan una recta donde la distancia entre los puntos (los valores observados) y la recta (valores predichos) sea mínima. Esta sería la recta de mejor ajuste. Para encontrar los parámetros que determinen la recta de mejor ajuste se utiliza el procedimiento de *mínimos cuadrados ordinarios* (MCO).

Las fórmulas para calcular los coeficientes mediante MCO son relativamente simples, aunque con muchas observaciones es preferible recurrir a paquetes estadísticos como R para la estimación. Por un lado, la pendiente se calcula como:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Por otro lado, el intercepto se determina con base en el valor estimado de la pendiente y las medias de las variables, $\bar{x}$ y $\bar{y}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ahora bien, retomando lo visto en el capítulo 2, ¿cómo saber si estas fórmulas de cálculo estiman correctamente los parámetros β_0 y β_1? Pues la teoría estadística, específicamente el teorema Gauss-Markov, sostiene que, bajo determinados supuestos, las fórmulas anteriores permiten obtener los mejores estimadores lineales insesgados, los cuales, en

la literatura en inglés, se conocen con el acrónimo BLUE, por *best linear unbiased estimators*. Por analogía, si para calcular una media desconocida, se sabe que el mejor estimador es el promedio aritmético; para conocer los valores de los parámetros β_0 y β_1, la teoría dice que los mejores estimadores posibles son los estimadores de mínimos cuadrados ordinarios, al ser insesgados, eficientes y consistentes bajo los supuestos del teorema Gauss-Markov.

5.4 Estimación

Con las fórmulas vistas, es fácil calcular los estimadores de mínimos cuadrados ordinarios manualmente con R. Primero, creamos los vectores de las variables independiente y dependiente, referidas a cada elección, combinadas en una única base de datos que denomino `votoeconomicoCR`:

```
partido<-c("Unidad1982", "PLN1986","PLN1990", "PUSC1994", "PLN1998",
           "PUSC2002", "PUSC2006", "PLN2010", "PLN2014", "PAC2018")
votos<-c(33.6, 52.3, 47.2, 47.7, 44.6, 38.6, 3.6, 46.9, 29.7, 21.6)
desempleo<-c(8.7, 6.9, 3.8, 4.1, 5.7, 6.1, 6.6, 7.8, 9.4, 9.1)
votoeconomicoCR<-data.frame(partido, votos, desempleo)
```

Empezando por la fórmula de la pendiente, es necesario calcular la diferencia de los valores de x_i y $\bar{x}$ y la diferencia entre los valores de y_i y $\bar{y}$, para luego multiplicar estas cantidades y sumarlas. Este es el numerador. La suma de las diferencias al cuadrado entre los valores de x_i y $\bar{x}$ es el denominador. Ambas cantidades se dividen para obtener el valor de la pendiente.

```
sum((desempleo - mean(desempleo))*(votos - mean(votos)))/
sum((desempleo - mean(desempleo))^2)
```

```
## [1] -3.217547
```

El intercepto se obtiene fácilmente con el valor de la pendiente ya estimado y con los promedios de las variables:

```
mean(votos) - (-3.217547)*mean(desempleo)
```

```
## [1] 58.52367
```

Por lo tanto, el intercepto estimado ($\hat{\beta}_0$) es 58.5 y la pendiente estimada ($\hat{\beta}_1$) -3.2.

Con los coeficientes de regresión estimados, el modelo de regresión se reescribe como una ecuación de valores predichos o ajustados de la variable dependiente con base en los valores observados de la variable independiente. En general, el modelo ajustado es:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

En el ejemplo:

$$\widehat{votos}_i = 58.5 - 3.2 desempleo_i$$

Puede verse que la ecuación no incluye un término de error, ya que esta representa un modelo ajustado para $\hat{y}_i$ (predicción), no para y_i (observación); determinar qué tan cercano es el modelo a la realidad de los datos es algo que se examinará luego.

Estos cálculos manuales de los coeficientes de regresión, aunque ilustrativos, son innecesarios, pues hay funciones programadas en R para obtener directamente los estimadores mínimos cuadrados ordinarios. Utilizamos la función `lm()` para los modelos de regresión lineal, en la cual se incluye primero la variable dependiente, luego la variable independiente y, por último, se define el conjunto de datos que se analiza. Asigno el modelo de regresión en un objeto, al que denomino `modelo1`, en el que se guardan los resultados de la estimación. Asignar los modelos en objetos es útil para luego compararlos y tabularlos. A fin de observar los resultados de la regresión, se ejecuta `summary()`.

```
modelo1<-lm(votos~desempleo, data=votoeconomicoCR)
summary(modelo1)
```

```
##
## Call:
## lm(formula = votos ~ desempleo, data = votoeconomicoCR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.688   0.003   1.895   4.080  15.977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   58.524     17.329   3.377  0.00968 **
## desempleo     -3.218      2.451  -1.313  0.22569
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.44 on 8 degrees of freedom
## Multiple R-squared:  0.1772, Adjusted R-squared:  0.07438
## F-statistic: 1.723 on 1 and 8 DF,  p-value: 0.2257
```

De la salida correspondiente, no nos interesan todos los resultados. En la siguiente sección se explica la interpretación de aquellos más relevantes. Por ahora, basta señalar que los coeficientes estimados mediante R coinciden con los calculados manualmente con las fórmulas: el intercepto es 58.5 y la pendiente -3.2.

5.5 Evaluación

Mediante la estimación de mínimos cuadrados ordinarios obtuvimos las estimaciones teóricamente insesgadas y eficientes de los parámetros. Ahora, ¿cómo se interpretan los resultados?

Primero, la constante o intercepto, $\hat{\beta}_0$, es el valor estimado de la variable dependiente y cuando x es igual a cero. En el ejemplo, el intercepto estimado es 58.5. Esto significa que, cuando el desempleo es 0 %, el partido en el gobierno obtiene 59 % de los votos válidos. Debe tenerse en consideración que el intercepto no siempre tiene una interpretación con sentido. Si en el ejemplo el intercepto fuese negativo, sería irreal, pues los votos solo adquieren valores positivos.

Más relevante resulta la interpretación de la pendiente estimada, $\hat{\beta}_1$. La pendiente significa el cambio en la variable dependiente y cuando x aumenta en una unidad. Este significado se entiende cuando se calcula algebraicamente la diferencia en el valor predicho, $\hat{y}$, cuando $x = 0$ y $x = 1$, es decir, cuando x aumenta en una unidad:

$$\begin{aligned}
\hat{y}_{x=1} - \hat{y}_{x=0} &= [\hat{\beta}_0 + \hat{\beta}_1(x = 1)] - [\hat{\beta}_0 + \hat{\beta}_1(x = 0)] \\
&= [\hat{\beta}_0 + \hat{\beta}_1 * 1)] - [\hat{\beta}_0 + \hat{\beta}_1 * 0] \\
&= [\hat{\beta}_0 + \hat{\beta}_1] - [\hat{\beta}_0] \\
&= \hat{\beta}_0 - \hat{\beta}_0 + \hat{\beta}_1 \\
&= \hat{\beta}_1
\end{aligned}$$

Puede verse que $\hat{\beta}_1$ equivale a la diferencia en el resultado (valores predichos) al cambiar de $x = 0$ a $x = 1$, o sea, cambiar x en una unidad.

En el ejemplo, la pendiente es -3.2. Entonces, si el desempleo cambia de 0 % a 1 %, el porcentaje de votos al partido en el gobierno se reduce –porque el signo es negativo– en 3.2 puntos porcentuales. Como la relación es lineal, el cambio en la variable dependiente es igual si el desempleo incrementa de 0 % a 1 %, de 2 % a 3 %, de 3 % a 4 %, etc. Bajo el supuesto de linealidad, también se puede calcular cuál sería el cambio si el desempleo aumentara en dos o más puntos porcentuales, mediante la multiplicación del coeficiente por el aumento de x. Así, con dos puntos de desempleo, el voto al gobierno disminuye 6.4 puntos porcentuales ($-3.2 * 2 = -6.4$). En conclusión, por cada punto porcentual que aumenta el desempleo, el porcentaje de votos *promedio* para el partido en el gobierno decrece 3.2 puntos porcentuales.

Además de indicar la magnitud de la relación entre x y y, la pendiente señala también la dirección. En el ejemplo, la pendiente es negativa, por ello se habla de disminución en el cambio. Si fuera positiva, la pendiente se interpretaría como aumento en y. En este sentido, es similar al coeficiente de correlación de Pearson.

Otro resultado importante es la significancia estadística de los coeficientes. La hipótesis nula establece que el coeficiente de regresión es igual a cero. El valor p indica la probabilidad de observar el valor del estadístico t o uno más extremo, si se asume la hipótesis nula cierta. En otras palabras, un valor p alto respalda la hipótesis nula y uno bajo permite rechazarla. En el ejemplo, el valor p para la pendiente es 0.226, de forma que no se puede rechazar la hipótesis nula y se considera que el coeficiente no es significativamente distinto de cero. El intercepto es significativo con un nivel de significancia de 0.05, pues su valor p es 0.0097, pero, en este ejemplo, tiene menor importancia sustantiva. La significancia estadística no equivale a relevancia.

Finalmente, la salida contiene un R cuadrado (múltiple) llamado *coeficiente de determinación* y denotado R^2, que es igual a 0.177. El coeficiente de determinación es igual al coeficiente de correlación de Pearson al cuadrado, es decir, $(r\ de\ Pearson)^2 = R^2$. Esto se puede comprobar fácilmente:

```
(cor(votoeconomicoCR$votos, votoeconomicoCR$desempleo, use="all.obs"))^2
```

```
## [1] 0.1772261
```

El R^2 se interpreta como la proporción de variancia de la variable dependiente explicada por el modelo. En el caso del voto económico en Costa Rica, dado que el $R^2 =$ es 0.177, se explica el 17.7 % de la variación en el voto para el partido en el gobierno, lo cual es poco, ya que hay 82.3 % de variancia *no* explicada.

En resumen, se concluye que el desempleo tiene un impacto de -3.2 puntos porcentuales sobre el voto para el partido en el gobierno, aunque este coeficiente no se puede distinguir estadísticamente de cero, y que el modelo explica el 18 % de la variación en el voto. Este es un resultado que reafirma la teoría del voto económico, pues la economía –medida a través del desempleo– incide en el comportamiento electoral, aunque el coeficiente no alcanza la significancia estadística según los umbrales convencionales y el modelo no explica el voto en su totalidad.

La figura 5.2 replica el gráfico de datos, pero ahora mostrando la línea de mejor ajuste, según el modelo de regresión estimado.

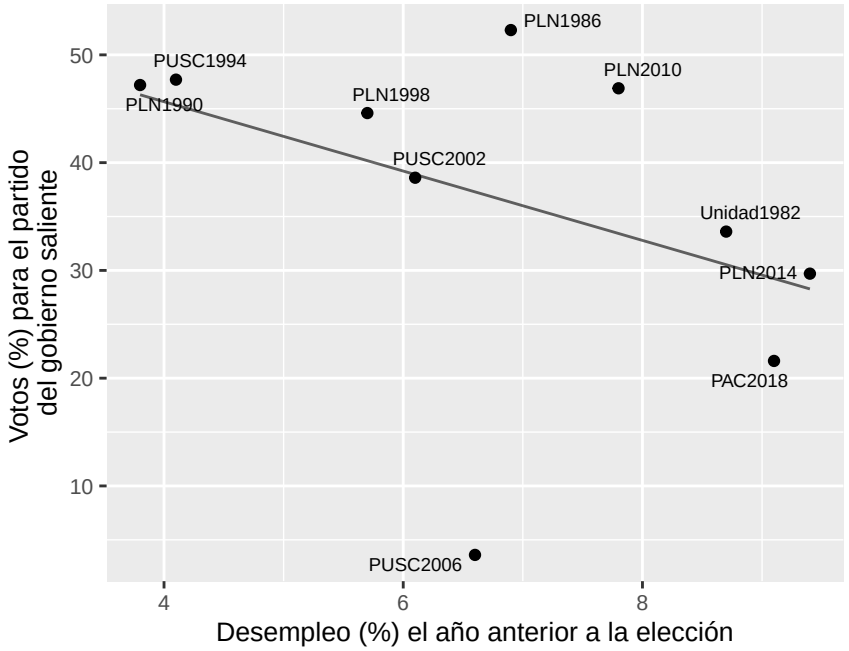


Figura 5.2 Recta de mejor ajuste según el modelo estimado

Del gráfico anterior resulta evidente que la elección del 2006 se aleja de la recta de mejor ajuste, ya que el Partido Unidad Social Cristiana (PUSC) obtuvo un porcentaje de votos más bajo de lo esperado, dado el desempleo en el año anterior a la elección.

La literatura sobre Costa Rica explica que el bajo apoyo al PUSC, pese a la situación económica positiva, se deriva de las acusaciones de corrupción contra dos expresidentes pertenecientes a este partido (Raventós Vorst, 2008). En estadística, estos casos que resultan excepcionales se denominan observaciones atípicas o desviadas (en inglés, *outliers*). Cabría, por lo tanto, preguntarse si el modelo estimado mejora al dejar por fuera esta elección particular.

En R, reestimamos el modelo con la indicación de excluir la observación atípica en el atributo `subset()` de la función `lm()`. Con `subset=(partido!="PUSC2006")` se estima el modelo de regresión para un subconjunto de datos determinados por la regla lógica de que sean distintos (signos `!=`) a la observación `PUSC2006` en el vector `partido`. En otras palabras, se ajusta el modelo con todas las observaciones, *menos* las consideradas atípicas, en este caso, PUSC en 2006.

```
modelo2<-lm(votos~desempleo, data=votoeconomicoCR,
            subset=(partido!="PUSC2006"))
summary(modelo2)
```

```
##
## Call:
## lm(formula = votos ~ desempleo, data = votoeconomicoCR, subset = (partido !=
##     "PUSC2006"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8509  -3.5638  -1.7143   0.4012  12.2475
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   63.894      9.197   6.947 0.000222 ***
## desempleo     -3.455      1.292  -2.675 0.031775 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.603 on 7 degrees of freedom
## Multiple R-squared:  0.5055, Adjusted R-squared:  0.4348
## F-statistic: 7.155 on 1 and 7 DF,  p-value: 0.03178
```

En el modelo reestimado sin la elección atípica, la pendiente es similar, -3.5, lo cual implica que el efecto –cambio promedio en el voto para el partido en el gobierno por cada punto porcentual que aumenta el desempleo– difiere poco en la estimación sin el resultado del PUSC en 2006. Sin embargo, este coeficiente es ahora significativamente distinto de cero, pues el valor p es 0.032. Además, el coeficiente de determinación aumenta considerablemente respecto a la primera estimación; ya que $R^2 = 0.505$, el segundo modelo explica el 50.5 % de la variación en el voto.

Al excluirse un caso atípico mejora no solo el ajuste del modelo (figura 5.3), sino también el respaldo a la teoría del voto económico: un mayor desempleo reduce significativamente el apoyo al partido que gobernaba.

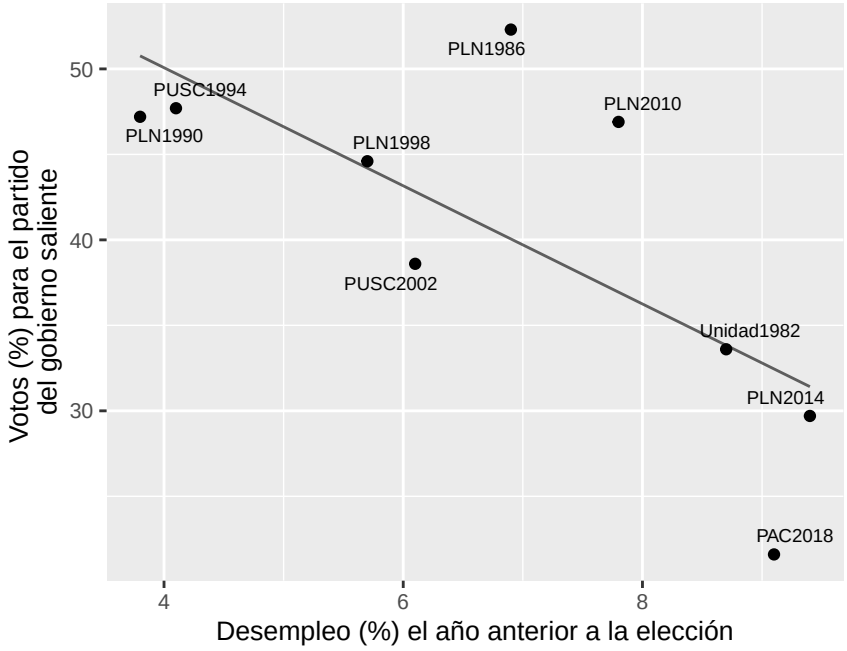


Figura 5.3 Recta de mejor ajuste sin el caso atípico del PUSC en 2006

5.6 Predicción

Con los modelos estimamos valores predichos. Estas predicciones se realizan dentro de la muestra, es decir, con los valores observados. No obstante, también podemos generar predicciones más allá. Por ejemplo, podemos estimar cuál sería el porcentaje de votos para el partido en el gobierno al asumir que el desempleo es 5 %, un valor que está en el rango de las tasas observadas de desempleo, pero que no corresponde a ninguna elección.

En R utilizamos `predict()` para generar el valor predicho, con intervalos de confianza para esta predicción con un nivel de confianza determinado. En este ejemplo, predecimos el valor estimado para un 5 % de desempleo, con un 95 % de confianza para el intervalo, de esta forma:

```
predict(modelo2, data.frame(desempleo=5),
        interval="confidence", level=0.95)
```

```
##        fit      lwr      upr
## 1 46.61747 38.39231 54.84264
```

La predicción de votos con un desempleo de 5 % es 46.6 %, con un intervalo de confianza [38.4 %, 54.8 %] al 95 %. Debido al número reducido de observaciones (nueve, en el modelo 2), la predicción es poco precisa y el intervalo demasiado amplio para una conclusión robusta. Con este intervalo de confianza no sabríamos si el partido pasa a una segunda vuelta electoral, ya que, en Costa Rica, el umbral que determina si un partido gana en primera ronda es 40 %.

Con la función `predict()` se pueden predecir múltiples valores al incluir un vector numérico en un `data.frame()`. Por ejemplo, para predecir los votos con tasas de desempleo de 5 %, 7 % y 16.5 %, ahora con un nivel de confianza de 90 %, escribiríamos:

```
predict(modelo2, data.frame(desempleo=c(5, 7, 16.5)),
        interval="confidence", level=0.90)
```

```
##         fit       lwr      upr
## 1 46.617474  40.02733 53.20762
## 2 39.706960  34.89042 44.52350
## 3  6.882019 -17.23120 30.99524
```

Hay que tener cuidado porque predecir fuera del rango de datos no es aconsejable (Lewis-Beck y Lewis-Beck, 2016). En el vector, la predicción con un desempleo de 16.5 % es arriesgada, pues implica exigirle una respuesta al modelo sobre un comportamiento que nunca ha observado. Sin embargo, es relevante, pues 16.5 % es el desempleo anual registrado en 2021, antes de las elecciones de 2022. Puesto que el partido saliente, Acción Ciudadana, obtuvo 0.7 % de los votos válidos, la predicción de 6.9 % sobreestima su éxito, aunque el intervalo de confianza contiene el valor observado de 0.7 %, lo cual es –cuanto mucho– sugerente.

5.7 Comentarios finales

Este capítulo introdujo la regresión como un método de análisis estadístico y la regresión lineal como un modelo particular para variables dependientes métricas. Con el modelo de regresión lineal, estimado con el método de mínimos cuadrados ordinarios, podemos obtener una recta de mejor ajuste que además cumple con las propiedades estadísticas deseables de ausencia de sesgo, eficiencia y consistencia. Desde el punto de vista aplicado, la regresión permite inferir el cambio en una variable dado el cambio en otra, la significancia estadística de este cambio y el poder explicativo del modelo estimado.

El caso simple –con solo una variable independiente– se presenta primordialmente con fines didácticos. En la práctica, los fenómenos políticos tienen más de un factor explicativo, es decir, son multicausales. En el ejemplo del voto económico, se sabe que la fragmentación partidaria ha aumentado en Costa Rica. Al haber más partidos compitiendo, los apoyos relativos para cada partido disminuyen. Es decir, deberíamos pensar, al menos, en dos variables explicativas del voto: desempleo y fragmentación. Además, con datos observacionales, un aparente efecto de una variable explicativa podría no serlo en realidad al existir otras variables correlacionadas. Si la inflación se correlaciona con el desempleo y con el voto para el partido en el gobierno, y dejamos por fuera la inflación, no sabríamos si el coeficiente de desempleo refleja verdaderamente un efecto del desempleo o si se confunde con el efecto de la inflación. Tanto para incluir más variables explicativas (como fragmentación partidaria) como para controlar posibles variables confusoras (como inflación), se utiliza la regresión múltiple, la cual se verá en el siguiente capítulo.

Finalmente, debe anotarse que, para que el modelo de regresión lineal, estimado por mínimos cuadrados ordinarios, ofrezca los mejores estimadores, es necesario que se cumplan determinados supuestos: linealidad en los parámetros, independencia entre el error y la variable independiente, normalidad en la distribución de los errores, variancia constante de los errores (denominada homoscedasticidad) y ausencia de correlación entre los errores. El capítulo siguiente discute algunos supuestos sobre la especificación del modelo. Asimismo, apéndice B contiene bibliografía recomendada que trata con mayor profundidad el diagnóstico de los supuestos y las medidas correctivas en caso de que algunos no se cumplan.

5.8 Ejercicios

1. Se dice que las elecciones presidenciales producen efectos secundarios sobre las elecciones legislativas (Samuels y Shugart, 2010). El siguiente cuadro presenta datos (recopilados por el autor) de las elecciones nacionales de Costa Rica desde 1953 hasta 2022. En una columna se indica el porcentaje de votos que obtuvo el partido ganador de la presidencia; en la otra, el número de legisladores pertenecientes al partido que ganó la presidencia.

    a) Pruebe, mediante un modelo de regresión estimado con R, la hipótesis de que el voto para la elección presidencial incrementa la bancada legislativa del partido en el gobierno.

    b) Interprete la pendiente.

    c) Interprete el coeficiente de determinación.

    d) Con los resultados, concluya respecto a la teoría.

Elección	Votos para presidente (porcentaje)	Legisladores del partido de gobierno
1953	64.7	30
1958	46.4	10
1962	50.3	29
1966	50.5	26
1970	54.8	32
1974	43.4	27
1978	50.5	27
1982	58.8	33
1986	52.3	29
1990	51.5	29
1994	49.6	28
1998	47.0	27
2002	38.6	19
2006	40.9	25
2010	46.9	24
2014	30.6	13
2018	21.6	10
2022	16.8	10

2. Con las fórmulas presentadas en el capítulo, calcule manualmente con R los coeficientes de regresión con los datos anteriores y compare con el resultado obtenido con la función `lm()`.

Capítulo 6

Regresión lineal múltiple

6.1 Introducción

En el capítulo 5, se estudió la regresión simple como una forma de modelar una relación lineal en la que una variable independiente x está asociada con una variable dependiente y o produce efectos en ella. Sin embargo, en la ciencia política –como en muchas otras disciplinas– difícilmente se puede asumir que *un* factor sea el responsable de un fenómeno particular, es decir, que sea su única causa. Los experimentos controlados con aleatorización, que sí permiten establecer la causa de una variable (ver capítulo 3), no son siempre factibles. En ciencia política no podemos experimentar con la asignación al azar de instituciones para determinar, por ejemplo, si las listas de elección abiertas generan mayor clientelismo que las listas cerradas, a partir de la aleatorización para que otros factores, como las culturas políticas locales, no intervengan en el resultado. Este diseño experimental no es políticamente factible ni éticamente aceptable. Por lo tanto, los estudios observacionales son una alternativa para examinar causas múltiples y la especificación de modelos de regresión con más de una variable independiente constituye un método apropiado para estos.

6.2 Modelo

El modelo lineal de regresión múltiple se expresa de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + u_i$$

En la ecuación, y_i corresponde a los valores de la variable dependiente (que es métrica o cuantitativa), x_{ji} son las variables independientes, con $j = 1, ..., k$, β_0 es el intercepto, β_1 hasta β_k son las pendientes, u_i son los errores o residuos de la estimación (la diferencia entre valores observados y predichos) y, por último, i es cada observación.

Nótese que:

- El modelo es aditivo, es decir, los términos se suman.
- Los efectos son directos, porque cada variable independiente afecta directamente la variable dependiente.
- Los puntos suspensivos indican que teóricamente se puede incorporar un número indefinido de variables independientes; el asunto de cuántas, en la práctica, se tratará luego.
- Cada variable independiente tiene un coeficiente asociado, el cual define su relación con la variable dependiente. Visualizar estas pendientes es complejo porque ya no se trata de un espacio bidimensional fácilmente graficable, como en la regresión simple, sino de uno multidimensional. Aunque en nuestras mentes no podemos imaginar un espacio de cuatro o más dimensiones, la matemática sí es capaz de resolver estos problemas multidimensionales.

Muchas de las definiciones del modelo se pueden flexibilizar para, por ejemplo, especificar que incluyen efectos multiplicativos de variables independientes entre sí, términos que se denominan *interacciones*. También es posible calcular efectos *indirectos* en modelos más complejos como los de ecuaciones estructurales. El apéndice B ofrece bibliografía complementaria para adentrarse en otras especificaciones y modelos.

6.3 Ejemplo

Un tema relevante en el estudio de la opinión pública es la aprobación o valoración de los gobiernos de turno. Por la teoría sobre la popularidad gubernamental (Bellucci y Lewis-Beck, 2011), sabemos que influyen factores tanto económicos (desempleo, inflación, crecimiento) como políticos (escándalos, manejos de crisis). También existen variaciones entre grupos sociodemográficos. Es decir, no podemos asumir que la aprobación gubernamental depende de un único factor. Por el contrario, debemos modelarla en función de múltiples variables independientes.

La encuesta de noviembre de 2020 realizada por el Centro de Investigación y Estudios Políticos (CIEP, 2020) contiene la nota que las personas en Costa Rica le dan al gobierno, en una escala donde 0 es la peor nota y 10 la mejor (variable `nota_gob`). Esta es, por lo tanto, una variable métrica que se puede modelar con regresión lineal.

Para explicar esta aprobación, se considera la valoración de la situación económica del país, pues la teoría sostiene que cuanto mejor evalúen las personas la economía, mayor es la aprobación del gobierno. No obstante, deberíamos incluir además factores sociodemográficos, pues nos interesa conocer *el efecto neto de la economía sobre la aprobación*, cuando se mantienen constantes la edad, el sexo, la provincia de residencia y el nivel educativo, pues las tasas de ocupación laboral e los ingresos varían según los perfiles sociodemográficos, los cuales a la vez podrían correlacionarse con el apoyo hacia el gobierno. En resumen, queremos saber si la valoración de la economía incide en la opinión sobre el gobierno, cuando se dejan al margen la edad de la persona, si es mujer u hombre, si vive en una provincia costera o central y el nivel educativo alcanzado.

Iniciamos cargando la base de datos:

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
```

Un paso importante, antes de iniciar cualquier análisis, es explorar las variables que se piensan incluir para determinar cómo se miden y se codifican. También es útil contar con el cuestionario de la encuesta o el manual de codificación. Por ejemplo, para una exploración de la variable dependiente, `nota_gob`, utilizamos `summary()`:

```
summary(ciep$nota_gob)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.000   5.000   4.094   6.000  10.000      14
```

Vemos que la variable `nota_gob` oscila entre 0 y 10, tiene una media de 4.1 y excluye 14 valores como `NA`, los cuales corresponden a respuesta de tipo "no sabe" o "no responde". En ocasiones, hay que recodificar los valores 99 –código convencional para la no respuesta– en `NA`. En este caso, ya están recodificados. Esta revisión descriptiva debe replicarse siempre con cada una de las variables.

La situación económica (en la base de datos, `sit_economica`) se midió con la pregunta "¿Cómo califica la situación económica del país?", para la cual las personas respondieron

con una escala de cinco puntos desde 5 ("muy bien") hasta 1 ("muy mal"). Por comodidad, incluyo esta variable como escala métrica de 1 a 5; alternativamente podría incluirse como variable categórica, pero veremos que su interpretación resulta menos simple. La variable `edad` es métrica, pues se midió en años cumplidos. El sexo es categórico, para lo cual se utiliza un código arbitrario: en esta base de datos, la variable `sexo` se codifica mujeres con 0 y hombres con 1 (no hay personas registradas como no binarias en la muestra). La provincia de residencia (variable `provinciarec`) también es categórica, donde 0 indica provincia costera y 1 provincia central. El nivel educativo (variable `educarec`) se basa en el último grado alcanzado y se agrupa en tres niveles: primaria o menos, secundaria y universitaria. Es, por lo tanto, una variable multicategórica.

Incluir variables multicategóricas en modelos de regresión (cualesquiera) requiere un tratamiento especial. Es necesario dejar por fuera una de las categorías que componen la variable, para que el modelo sea estimable, es decir, si hay C categorías, entonces se calculan $C-1$ coeficientes (solamente si se suprime la estimación del intercepto, se pueden obtener coeficientes para todas las categorías de una variable). Los programas estadísticos, por lo general, dejan por fuera la categoría codificada con el menor valor. En el ejemplo, primaria o menos es 1, secundaria es 2 y universitaria es 3. Por lo tanto, R calcularía coeficientes únicamente para secundaria y universitaria; primaria o menos se convierte en la base de comparación.

Conociendo las variables y su codificación, el modelo se formaliza como:

$$nota\_gob_i = \beta_0 + \beta_1 sit\_economica_i + \beta_2 edad_i + \beta_3 sexo_i + \beta_4 provinciarec_i$$

$$+\beta_5 secundaria_i + \beta_6 universitaria_i + u_i$$

Obsérvese que en el modelo hay un coeficiente para cada variable independiente (y dos para nivel educativo, pues tiene tres categorías), más el intercepto común. Estos siete parámetros (valores desconocidos) se estiman con el método de mínimos cuadrados ordinarios. Sin embargo, dada la multidimensionalidad, no es fácil aplicar fórmulas como en la regresión simple del capítulo 5. Ahora el álgebra lineal requerida es más compleja.

Por suerte, podemos recurrir a R para estimar estos parámetros, utilizando la misma función `lm()` de la regresión simple, pues es el mismo modelo, solo que con más variables. Así, para incluir el conjunto de variables independientes seleccionadas, adicionamos las variables en el código, al igual que en la ecuación. Las variables `sexo` y `provinciarec`

son categóricas que, al estar codificadas con 0 y 1, es posible incluirlas directamente en el modelo. Si tuvieran otra codificación (por ejemplo, 1 y 2), o si contuvieran más de dos categorías, se incluirían en el modelo con `factor()`, como sucede con la variable `educarec` que contiene tres categorías. Incluimos el modelo en el objeto `M1` para mayor comodidad y examinamos el resultado con `summary()`:

```
M1<-lm(nota_gob~sit_economica+edad+sexo+provinciarec+factor(educarec),
       data=ciep)
summary(M1)
```

Alternativamente, es práctico preparar las variables multicategóricas como factores antes de estimar el modelo. Para esta base de datos, creada originalmente en Stata, se utiliza la función `as_factor()` del paquete `haven` (función que no se debe confundir con `as.factor`).

```
ciep$educarec<-as_factor(ciep$educarec, levels="labels")
```

Esta transformación no solo evita tener que incluir la variable con `factor()` en el modelo, sino que además facilita la interpretación de los resultados porque identifica con etiquetas las categorías para las cuales se obtienen coeficientes.

```
M1<-lm(nota_gob~sit_economica+edad+sexo+provinciarec+educarec,
       data=ciep)
summary(M1)
```

```
##
## Call:
## lm(formula = nota_gob ~ sit_economica + edad + sexo + provinciarec +
##     educarec, data = ciep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0053 -2.1968  0.1499  1.9740  7.0717
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.1773989  0.3999690   5.444 6.68e-08 ***
## sit_economica           1.0318348  0.1086331   9.498  < 2e-16 ***
## edad                   -0.0002726  0.0058731  -0.046  0.96300
```

```
## sexo                  -0.3840783  0.1716876  -2.237  0.02552 *
## provinciarec           0.1249816  0.1903797   0.656  0.51167
## educarecSecundaria     0.0426653  0.2312348   0.185  0.85365
## educarecUniversitaria  0.6144686  0.2339257   2.627  0.00876 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.607 on 927 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.1046, Adjusted R-squared:  0.09885
## F-statistic: 18.06 on 6 and 927 DF,  p-value: < 2.2e-16
```

Al igual que en el capítulo anterior, nuestra interpretación del modelo se centra en la magnitud de los coeficientes, la significancia de estos y el poder explicativo. En la primera columna se presentan los coeficientes estimados, es decir, aquellos que minimizan los errores y que, a la vez, corresponden a los mejores estimadores lineales insesgados. En otras palabras, son los valores que permiten reescribir la ecuación como

$$\widehat{nota\_gob}_i = 2.18 + 1.03 sit\_economica_i - 0.0003 edad_i - 0.38 sexo_i$$

$$+0.12 provinciarec_i + 0.04 secundaria_i + 0.61 universitaria_i$$

Los valores predichos de la nota del gobierno son, por lo tanto, una combinación lineal de distintos valores de situación económica, edad, sexo, provincia, educación secundaria y educación universitaria, más un intercepto común. El intercepto se interpreta como el valor promedio de la variable dependiente cuando todos los variables independientes son cero. Para las pendientes, se mantiene la interpretación vista en el capítulo anterior: cada pendiente indica cuánto cambia y, en promedio, por cada cambio de una unidad en x_k, con la añadidura de que los cambios de cada variable independiente son marginales, es decir, se dejan las demás variables constantes o sin variar.

Empecemos con las variables métricas. Con el coeficiente 1.03 de la variable situación económica, se puede interpretar que, por cada punto que aumenta la escala de valoración de la situación económica, la nota del gobierno se incrementa 1.03, en promedio, con todas las demás variables constantes. Para la variable edad, se encuentra que, por cada año cumplido, la nota del gobierno disminuye (pues el signo es negativo) 0.0003 puntos, que es una magnitud prácticamente nula, con las demás variables constantes.

Ahora seguimos con las categóricas. Para interpretar el coeficiente de sexo, debe tenerse presente cuál es la codificación de las categorías. El coeficiente estimado que se muestra en los resultados se refiere a la categoría indicada con 1, que se compara con la categoría 0 para la cual no hay un coeficiente. En el ejemplo, el coeficiente -0.38 se refiere a hombres (codificados 1) que se compara con las mujeres (codificadas 0) para obtener la siguiente interpretación: entre los hombres, la nota promedio hacia el gobierno es 0.38 menor (porque el coeficiente es negativo), en comparación con las mujeres, manteniendo constantes las demás variables. Es decir, las mujeres otorgan una nota, en promedio, mayor. Ahora para la variable de provincia, donde 1 es provincia central y 0 provincia costera: entre habitantes de una provincia central, la valoración del gobierno es 0.12 mayor (porque el signo es positivo), en promedio, respecto a habitantes de una provincia costera; todas las demás variables se mantienen constantes.

Por último, la variable educación es un caso especial de variable categórica al estar conformada por tres categorías. Como se explicó, al incluirla como factor, R calcula coeficientes para dos categorías, secundaria y universitaria, que tienen código 2 y 3, respectivamente, y deja por fuera la categoría primaria o menos, con código 1. Los coeficientes estimados se interpretan al compararlos contra la categoría que quedó fuera, en este caso, primaria o menos. Así, el coeficiente de educación secundaria, 0.04, se interpreta como que la nota promedio del gobierno es 0.04 mayor entre personas con educación secundaria, en comparación con las personas con educación primaria o menos. A su vez, el coeficiente de universidad indica que entre personas con estudios universitarios la nota del gobierno es 0.61 mayor, en promedio, que entre personas con educación primaria o menos. Esto evidencia que cuantas más categorías contenga una variable, más engorrosa es la interpretación. Por esta razón, muchas veces conviene recodificar categorías en menos grupos. También, para variables ordinales, se puede asumir una escala, como se hizo con la valoración económica, que se interpretó de forma métrica de 1 a 5, en lugar de incluirla con categorías.

La significancia es el segundo criterio al evaluar el modelo. Se examinan los valores p asociados a cada coeficiente para determinar la probabilidad de obtener un estadístico calculado o uno más extremo, al asumir cierta la hipótesis nula que postula que el coeficiente es cero. En otras palabras, un valor p alto (mayor a un nivel de significancia escogido) sugiere mucha evidencia a favor de la hipótesis nula, por lo que esta no se rechaza y se concluye que el coeficiente no es significativamente distinto de cero.

Por el contrario, si el valor p es bajo (menor al nivel de significancia), se rechaza la hipótesis nula y se sostiene que el coeficiente es significativamente distinto de cero o, en breve, que es significativo.

Con un nivel de significancia convencional como 0.05, los coeficientes de las variables situación económica, sexo y educación universitaria son significativos, mientras que los de edad, provincia y educación secundaria no. La pendiente también es significativa, aunque en este caso no tiene relevancia teórica. En resumen, se concluye que la valoración de la economía, el sexo de la persona y el nivel educativo universitario influyen en la opinión sobre el gobierno; en tanto que la edad, la provincia de residencia y el nivel educativo secundario no están asociados con ella.

Finalmente, se interpreta el coeficiente de determinación (R^2) de la misma forma que en regresión simple: ya que el R^2 es 0.105, el modelo explica el 10.5 % de la variación en la nota del gobierno. Nótese, sin embargo, que existe también un R^2 ajustado, 0.099, que es muy similar al primero. El R^2 ajustado es una medida más apropiada en los modelos múltiples porque penaliza según el número de variables independientes. Puesto que incluir variables independientes automáticamente aumenta el coeficiente de determinación, este puede incrementarse solo por contener demasiadas variables y no porque el modelo explique bien el fenómeno. El R^2 ajustado es, por lo tanto, una medida más refinada del porcentaje de bondad de ajuste, pues corrige según el número de predictores.

Un último paso al hacer análisis de regresión es la presentación de resultados. La salida de R –en este como en otros métodos– no es un formato elegante para presentar los hallazgos estadísticos. Es preferible construir un cuadro propio siguiendo algún manual de estilo. R tiene para ello un paquete muy útil, llamado `sjPlot`, que produce cuadros de regresión en un formato similar a los que se encuentran en publicaciones académicas.

Con las siguientes líneas genero un cuadro para el modelo estimado, con la indicación de que se muestren los errores estándar y no los intervalos de confianza. Además, el cuadro se guarda en un archivo de texto en la carpeta de trabajo, según se haya definido con `setwd()` al iniciar la sesión (ver el apéndice A).

```
library(sjPlot)
tab_model(M1, show.se=TRUE, show.ci=FALSE, auto.label=TRUE,
          file="modelomultiple.doc")
```

6.4 Problemas en la especificación de modelos de regresión múltiple

Aunque la estimación de un modelo de regresión múltiple no cambia respecto al modelo simple (se utilizan mínimos cuadrados ordinarios), hay situaciones particulares en torno a la especificación, es decir, la construcción del modelo. Veamos algunos de estos aspectos.

6.4.1 ¿Cuántas variables independientes incluir?

Un supuesto fundamental del modelo de regresión lineal es que esté bien especificado (Gujarati y Porter, 2010, p. 200). Esto implica que se incluyan todas las variables necesarias y que la relación entre las variables sea la correcta. Asumiendo que acertamos en que el modelo siga una función lineal (lo cual no es siempre así), ¿cómo saber cuántas y cuáles variables independientes incluir en un modelo?

La respuesta, desde el punto de vista de la investigación politológica, es que se deben incluir las variables pertinentes según la teoría. Es decir, la teoría es el mapa desde el cual se orienta el análisis y se seleccionan las variables independientes.

Desde el punto de vista estadístico, sin embargo, hay otras consideraciones. En general, el número de variables no debe superar el número de observaciones, ya que el modelo se indeterminaría y no se podría estimar. Para ejemplificar, consideremos esta ecuación matemática:

$$y = x\beta_0 + 2$$

Si $y = 10$ y $x = 4$, entonces tenemos una ecuación con una única incógnita:

$$10 = 4\beta_0 + 2$$

En este caso, es posible despejar β_0 para conocer su valor:

$$\frac{10 - 2}{4} = \beta_0$$

$$2 = \beta_0$$

Fácilmente se encuentra que el valor desconocido es 2.

En cambio, observemos qué pasa cuando se plantea:

$$10 = 4\beta_0 + 2\beta_1$$

En esta última ecuación las soluciones –los posibles valores de las incógnitas– son infinitas. No se puede resolver una ecuación con dos incógnitas; requerimos dos ecuaciones para encontrar las dos incógnitas, es decir, un sistema de ecuaciones. Lo mismo ocurre en regresión. Cada ecuación contiene un conjunto de valores x y y, mientras que las incógnitas son los parámetros. Por lo tanto, no podemos inferir los dos parámetros (intercepto y pendiente) con una observación o par ordenado x y y.

De forma general, el número de variables (cada una con un parámetro o pendiente por estimar) no puede ser mayor al número de observaciones. Incluso sin llegar al extremo de tener más variables que observaciones, conforme el número de coeficientes por estimar se acerca al número de observaciones, el modelo de regresión pierde precisión, pues se incrementa el tamaño de los errores estándar de los coeficientes. Podríamos obtener coeficientes no significativos, los cuales en realidad sí lo son.

Teniendo presente esta restricción, el dilema metodológico radica en que, por un lado, se deben incluir todas las variables necesarias para explicar un fenómeno según el marco teórico y, por otro lado, se debe conservar la parsimonia o la simplicidad del modelo, evitando incorporar variables irrelevantes, es decir, que no aportan nada a la explicación.

6.4.2 Sesgo de variable omitida

Si bien los modelos procuran ser parsimoniosos, existe un peligro al omitir una variable relevante, ya que el estimador resultaría sesgado. A nivel conceptual, el sesgo de variable omitida consiste en atribuir el efecto de una variable independiente x_1 sobre la y, cuando en realidad el efecto es de una x_2 que no se incluyó en el modelo, pero que está correlacionada con x_1 y con y. De esta forma, se establece una relación espuria o un problema de confusión entre x_1 y y.

De acuerdo con el ejemplo desarrollado anteriormente, podría argumentarse que el efecto de la valoración de la economía en realidad proviene de la situación económica personal. Se supondría, entonces, que las personas desempleadas otorgan una peor calificación al gobierno que las personas empleadas y que el efecto de la valoración económica proviene de que las personas desempleadas son más pesimistas sobre la economía del país. En

otras palabras, la desaprobación del gobierno proviene de la valoración personal y no de una percepción sobre la economía nacional. Para probar esta hipótesis, se incluye en el modelo la variable `desempleado`, codificada 1 si la persona está desempleada y 0 si tiene empleo. En esta estimación omitimos la variable educación.

```
M2<-lm(nota_gob~sit_economica+edad+sexo+provinciarec+desempleado,
       data=ciep)
summary(M2)
```

```
##
## Call:
## lm(formula = nota_gob ~ sit_economica + edad + sexo + provinciarec +
##     desempleado, data = ciep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2233 -2.2025  0.2973  1.7746  7.2091
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.561909   0.349714   7.326 5.17e-13 ***
## sit_economica  1.030119   0.108800   9.468  < 2e-16 ***
## edad          -0.002121   0.005649  -0.375   0.7074
## sexo          -0.416703   0.172824  -2.411   0.0161 *
## provinciarec   0.183682   0.188706   0.973   0.3306
## desempleado   -0.491752   0.219785  -2.237   0.0255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.607 on 924 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.1006, Adjusted R-squared:  0.09578
## F-statistic: 20.68 on 5 and 924 DF,  p-value: < 2.2e-16
```

El modelo anterior (`M2`) indica que efectivamente la variable `desempleado` reduce la nota del gobierno, con significancia estadística. Sin embargo, el coeficiente de situación económica mantiene su significancia. El modelo, por lo tanto, sostiene que tanto la percepción nacional como la condición de desempleo influyen en la valoración gubernamental.

Recordemos, sin embargo, que en el modelo original (`M1`) se incluía la variable nivel educativo. ¿Qué pasaría si el efecto observado del desempleo proviene de personas con menor grado académico y mayor dificultad para ingresar al mercado laboral? De ser esto cierto, estaríamos observando una correlación espuria al asumir que hay un efecto del desempleo sobre la aprobación gubernamental, sin considerar el nivel educativo de las personas.

```
M3<-lm(nota_gob~sit_economica+edad+sexo+provinciarec+desempleado+educarec,
       data=ciep)
summary(M3)
```

```
##
## Call:
## lm(formula = nota_gob ~ sit_economica + edad + sexo + provinciarec +
##     desempleado + educarec, data = ciep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0477 -2.0802  0.1391  1.9194  7.3778
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.3172731  0.4074983   5.687 1.74e-08 ***
## sit_economica           1.0266043  0.1083500   9.475  < 2e-16 ***
## edad                   -0.0007561  0.0058662  -0.129  0.89747
## sexo                   -0.4479369  0.1724285  -2.598  0.00953 **
## provinciarec            0.0958961  0.1901487   0.504  0.61416
## desempleado            -0.4294898  0.2197917  -1.954  0.05099 .
## educarecSecundaria      0.0870502  0.2312768   0.376  0.70671
## educarecUniversitaria   0.6182062  0.2343189   2.638  0.00847 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.596 on 922 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.1102, Adjusted R-squared:  0.1034
## F-statistic: 16.31 on 7 and 922 DF,  p-value: < 2.2e-16
```

Efectivamente, al incluir la educación, el desempleo pierde la significancia (`M3`), lo cual confirma que el modelo sin educación (`M2`) comete un sesgo de variable omitida y que el estado de desempleo es irrelevante en la explicación.

Tanto la omisión de variables con relevancia como la inclusión de variables innecesarias acarrean problemas estadísticos. La omisión produce sesgo e inconsistencia en los estimadores; y la inclusión innecesaria, ineficiencia o mayores errores estándares (Gujarati y Porter, 2010, pp. 470-474). Puesto que saber qué incluir en el modelo no es obvio ni sencillo, la teoría es la mejor guía. Como dicen Gujarati y Porter (2010, p. 511), "la elaboración de modelos es tanto un arte como una ciencia".

6.4.3 Multicolinealidad

Es común que las variables independientes presenten cierta correlación entre sí. Por ejemplo, el nivel educativo y el ingreso están asociados, así como el interés político y el nivel de información sobre la política. Sin embargo, cuando las variables independientes están *fuertemente* correlacionadas, se genera un problema en regresión denominado multicolinealidad. Este puede afectar las estimaciones, al agrandar los errores estándar, de modo que los coeficientes podrían interpretarse como no significativamente distintos de cero, cuando, en realidad, sí lo son (Gujarati y Porter, 2010, p. 327).

La multicolinealidad puede evaluarse mediante correlaciones bivariadas, por ejemplo, con los coeficientes r de Pearson y V de Cramer (ver capítulo 4). Otra forma de diagnosticar la multicolinealidad es observar un coeficiente de determinación R^2 muy alto con coeficientes de regresión que no son significativamente distintos de cero.

Existen distintas estrategias para resolver los problemas de multicolinealidad. Lo primero es constatar que dos variables no estén midiendo lo mismo. Por ejemplo, incluir simultáneamente en un modelo el número de partidos políticos de la elección y un indicador de fragmentación partidaria (como el índice Laakso-Taagepera) induciría un problema de multicolinealidad, pues ambos miden el mismo concepto. Otra solución, más sofisticada, incluye el uso de análisis de factores (capítulo 9), el cual permite reducir el número de variables en factores no correlacionados entre sí. La regresión *ridge* es otra alternativa ante la multicolinealidad problemática.

6.4.4 Endogeneidad

La regresión supone un modelo teórico donde las variables independientes generan efectos sobre la variable dependiente. Sin embargo, ¿qué pasaría si la variable dependiente también pudiese influenciar una variable independiente? Esto se llama causalidad recíproca y es una fuente de endogeneidad.

En sentido estricto, la endogeneidad es la correlación entre las variables explicativas y los errores. Una fuente de dicha correlación (pero no la única) es la causalidad recíproca (ver Wooldridge, 2010, pp. 54-55). La endogeneidad, entendida como causalidad recíproca, es uno de los problemas más frecuentes en la ciencia política (Franzese, 2007). Para ilustrar, se sabe que los sistemas electorales condicionan el número de partidos políticos (Duverger, 1957; Shugart y Taagepera, 2017). No obstante, también se argumenta que los partidos, como actores políticos, buscan alterar los sistemas políticos para beneficiarse (Colomer, 2003). Entonces, modelar el número de partidos en función del sistema electoral conlleva un problema estadístico de endogeneidad. La relación entre democracia y desarrollo económico es otro problema clásico de endogeneidad, pues democracia y desarrollo se refuerzan mutuamente (Przeworski *et al.*, 2000).

Aunque las soluciones estadísticas ante la endogeneidad pueden resultar muy sofisticadas, una estrategia sencilla para alivianar la endogeneidad consiste en utilizar variables independientes rezagadas en el tiempo. Por ejemplo, en un estudio de crecimiento económico y democracia, se puede medir el producto interno bruto (como indicador de la variable independiente, desarrollo económico) un año antes respecto al indicador de democracia (variable dependiente). No obstante, esta es una respuesta limitada, pues no contempla la posible heterogeneidad entre los distintos casos, por ejemplo, las características históricas y culturales de los países (Pignataro, 2018).

Otro método más robusto consiste en estimar un modelo con una variable alternativa, llamada *instrumento*, que cumpla dos condiciones: estar parcialmente correlacionada con la variable independiente endógena (aquella que está influenciada por la variable dependiente); y ser exógena (no endógena) respecto a la dependiente (Wooldridge, 2010). También existen los llamados modelos no recursivos que estiman las relaciones recíprocas o la retroalimentación (*feedback*) que produce endogeneidad (Berry, 1984). En resumen, se necesitan herramientas avanzadas para resolver la endogeneidad.

6.5 Comentarios finales

En este capítulo se amplía el modelo de regresión lineal del capítulo 5 hacia uno con varios predictores o variables independientes. El modelo múltiple se puede ver como una extensión de la regresión simple o, lo que es lo mismo, la regresión simple como un caso particular de la múltiple. En cuanto a la estimación, se utiliza también el método de mínimos cuadrados ordinarios para obtener estimadores insesgados, consistentes y eficientes. No obstante, la especificación y la evaluación del modelo requieren mayor atención bajo el modelo múltiple. En particular, se deben incluir las variables necesarias, pero sin saturar el modelo con variables irrelevantes. Se debe procurar que las variables explicativas no estén demasiado correlacionadas entre sí (eliminar la multicolinealidad). Por último, las variables independientes deben ser exógenas: no deben estar influenciadas por la variable dependiente (ausencia de endogeneidad).

Se vio además que el coeficiente de determinación (R^2) no es un criterio perfecto para determinar la calidad del modelo. El número de variables independientes incide en que este sea mayor, incluso si el R^2 ajustado corrige esta inflación. La multicolinealidad puede incrementar también este coeficiente sin basarse en que el modelo sea, por sí mismo, bueno. Por lo tanto, resulta pertinente reflexionar sobre la validez de las variables seleccionadas. Al respecto, en la ciencia política Christopher Achen (2005) ha sido un fuerte crítico de los modelos que incluyen variables independientes sin discreción, por más correctos que sean los estimadores. Sugiere, por el contrario, fundamentar teóricamente las variables y realizar análisis exploratorios (simples gráficos y tabulaciones entre cada variable independiente y la variable dependiente) antes de estimar los modelos.

Incluso si el modelo está bien especificado y el R^2 es alto, existirá variancia no explicada. ¿Qué pasa con ella? Existen dos posiciones filosóficas: una supone el mundo de forma determinista, por lo que es posible explicar el 100 % ($R^2 = 1$) de un fenómeno si se incluyen todas las variables explicativas; la otra afirma que hay componentes aleatorios en el mundo, por lo cual nunca se podrán hacer predicciones perfectas donde la variancia explicada sea 100 % (King *et al.*, 1994, p. 59). En otras palabras, debemos resignarnos a coeficientes de determinación imperfectos y, en la mayoría de las veces, bajos.

Para finalizar, tengamos presente que en regresión lineal es posible construir modelos no solo con términos aditivos, como los vistos en el capítulo, sino también con componentes cuadráticos y con multiplicaciones (*i. e.*, interacciones) entre variables independientes.

Por ejemplo, ¿qué pasa con patrones curvilíneos, como entre edad y voto, en el que las personas de menor edad votan menos, las de edad intermedia votan más y las de mayor edad votan menos. Para esto, habría que incluir un término cuadrático para edad, de forma que capture el comportamiento curvilíneo entre edad y voto:

$$voto_i = \beta_0 + \beta_1 edad_i + \beta_2 edad_i^2 + u_i$$

En estos y en todos los casos, el objetivo es aproximarse al modelo "verdadero", una búsqueda, como se mencionó en este capítulo, que es parte ciencia, parte arte.

6.6 Ejercicios

1. Utilice la base de datos "CIEPnoviembre2020.dta" para estimar un modelo de regresión múltiple que explique la nota que le brindan las personas a los partidos políticos en Costa Rica (`nota_pp`), medida en una escala de 0 a 10, con base en las siguientes variables independientes: nota que le dan al gobierno (`nota_gob`) en una escala de 0 a 10, nota que le dan a la Asamblea Legislativa (`nota_al`) en una escala de 0 a 10, edad en años cumplidos (`edad`), sexo de la persona entrevistada (`sexo`, donde 0 = mujer y 1 = hombre) y provincia de residencia (`provinciarec`, donde 0 = costera y 1 = central).
2. Interprete los coeficientes, la significancia y el coeficiente de determinación con los resultados del modelo estimado en el punto 1.
3. Discuta si el modelo del punto 1 presenta problemas de sesgo de variable omitida, multicolinealidad y endogeneidad.

Capítulo 7

Regresión logística

7.1 Introducción

Los dos capítulos anteriores mostraron que el análisis de regresión es una herramienta útil y poderosa para contrastar hipótesis explicativas y modelar fenómenos políticos. En ellos se estudió el modelo lineal o gaussiano, estimado a través del método de mínimos cuadrados ordinarios (MCO), el cual se aplicó en ejemplos con variables dependientes que son métricas o cuantitativas. En muchos casos, sin embargo, la variable de interés tiene una medición categórica o cualitativa. Pueden pensarse múltiples ejemplos: la participación electoral (votar o abstenerse), la identificación partidaria (simpatizar con un partido o con ninguno), el comportamiento legislativo (a favor o en contra de un proyecto de ley) y las relaciones entre Estados (conflicto o paz), entre otros.

Para modelar las variables anteriores, que son categóricas binarias o dicotómicas (de dos categorías), la estimación por MCO resulta inadecuada. En un gráfico es fácil de ver. Si codificamos una variable dependiente categórica binaria en 0 y 1, una simulación de datos muestra que la línea de mejor ajuste según la estimación MCO está lejos de casi todas las observaciones (figura 7.1). El modelo lineal no parece minimizar el error –la distancia entre la línea y las observaciones–. Además, bajo este modelo se predicen valores fuera del rango 0 y 1, el cual constituye el límite de la variable dicotómica. Es decir, si 1 es votar y 0 abstenerse, ¿qué significa una predicción de más de 1 o de menos de 0? Ante los problemas que presenta el modelo lineal con MCO, estudiaremos una alternativa que se denomina regresión logística.

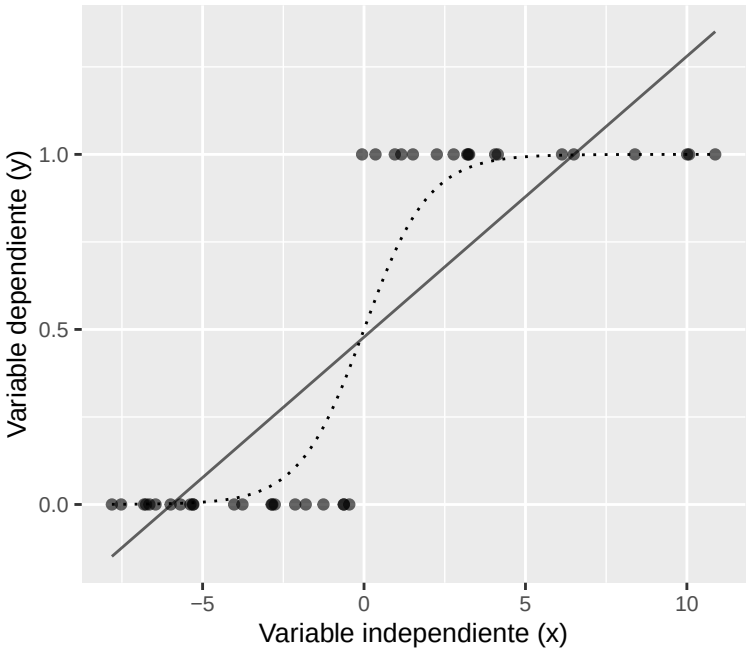


Figura 7.1 Simulación del ajuste por mínimos cuadrados ordinarios (línea) y por regresión logística (curva)

7.2 Modelo

Una recta no se ajusta bien a variables categóricas binarias. En cambio, una curva en forma de "s", como la línea punteada que aparece en la figura 7.1, se aproxima mejor a este tipo de datos.

El modelo logístico es una solución, entre varias, para ajustar una curva en forma de "s" debido a sus propiedades matemáticas y la interpretación sustantiva que permite (Hosmer *et al.*, 2013, p. 7). Similar al modelo lineal o gaussiano, se tiene una ecuación lineal de la siguiente forma:

$$w = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

donde β_0 es el intercepto, β_1 hasta β_k los parámetros de las pendientes y x_j son las variables independientes, con $j = 1, ..., k$. Hasta el momento no hay ninguna novedad respecto al modelo lineal estimado con mínimos cuadrados ordinarios.

Lo que se hace ahora diferente es utilizar una función matemática que produzca una curva como la que se visualiza en la figura 7.1. Dicha función es la siguiente:

$$f(w) = \frac{e^w}{1 + e^w}$$

donde e corresponde al número irracional que es aproximadamente 2.718.

Esta función genera valores entre 0 y 1, los cuales se pueden interpretar como probabilidades de pertenencia a una de las dos categorías de la variable dependiente. En términos de ocurrencia del fenómeno codificado, se plantea que:

$$P(y = 1) = \frac{e^w}{1 + e^w}$$

donde $P(y = 1)$ denota la probabilidad de pertenecer al grupo codificado 1. Por ejemplo, si la variable dependiente es participación en una elección, podemos codificar si la persona votó como 1 y si no votó como 0. Esta decisión es arbitraria, porque bien podríamos haber decidido que 1 fuera abstenerse. Sin embargo, la escogencia determina la forma en que se interpreta el modelo. Si 1 significa que votó, entonces $P(y = 1)$ predice la probabilidad de pertenecer al grupo de votantes. Si 1 fuese abstenerse, $P(y = 1)$ sería la probabilidad de no votar.

Al sustituir w, se obtiene el modelo logístico:

$$P(y = 1|x_1, x_2, ..., x_k) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}$$

donde los parámetros β_0 a β_k del modelo logístico se deben estimar por el método de máxima verosimilitud, no por mínimos cuadrados ordinarios.

Un aspecto conceptual importante es que la regresión logística es un *modelo lineal en sus parámetros* desde el marco de los modelos lineales generalizados (ver apéndice C). Se puede demostrar que al aplicar la transformación logito se llega a una ecuación similar (excepto por la ausencia del término del error) a la del modelo gaussiano visto en el capítulo 6:

$$ln\left[\frac{P(y = 1)}{1 - P(y = 1)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

En palabras coloquiales, el modelo lineal gaussiano y el modelo logístico son "primos", miembros de una misma familia de modelos, en la que hay un núcleo común y varían otras características (*i. e.*, la función de enlace y la distribución de los errores). En cambio, cuando hablamos de modelos no lineales, nos referimos a una familia distinta de modelos. A continuación, se verá una aplicación del modelo logístico a modo de ejemplo.

7.3 Ejemplo

En este ejemplo modelaremos el voto para el Partido Acción Ciudadana (PAC), según los datos de la encuesta de noviembre de 2020 del Centro de Investigación y Estudios Políticos (CIEP, 2020). Estudiar el voto es sustantivamente relevante debido a que el PAC obtuvo la presidencia en Costa Rica durante el periodo 2018-2022. Aunque la encuesta se realizó dos años después de la elección, permite identificar las fuentes de apoyo en la opinión pública. Específicamente, la pregunta es "¿Por quién votó usted en la segunda ronda de las elecciones presidenciales de 2018?". Puesto que nos interesa predecir el voto por el PAC, codificamos la variable de manera binaria, donde 1 es si votó por el PAC y 0 es si votó por Restauración Nacional o no votó; se excluyen las personas que no tenían edad para votar, no recuerdan por quién votaron o no responden.

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
```

Analizaremos si existen variaciones en el voto para el PAC (variable `votopac`), según la edad en años cumplidos (`edad`), si es mujer u hombre (`sexo`), si habita en una provincia central o costera (`provinciarec`) y el nivel educativo primaria o menos, secundaria y universitaria (`educarec`). Las reglas para incorporar las variables independientes son las mismas que en regresión lineal. Las métricas se incorporan sin complicaciones (`edad`). Las dicotómicas, si están codificadas con 0 y 1, se incluyen directamente (`sexo` y `provinciarec`). Las categóricas, si tienen más de dos categorías, se añaden como factores (`educarec`) y se estimarán tantos coeficientes como categorías, menos una que se convierte en la base de comparación; se pueden incluir con `factor()` en el modelo o luego de aplicar `as_factor()` antes de la estimación.

```
ciep$educarec<-as_factor(ciep$educarec, levels="labels")
```

Para el modelo logístico, primero, definimos w como:

$$w = \beta_0 + \beta_1 edad + \beta_2 sexo + \beta_3 provinciarec + \beta_4 secundaria + \beta_5 universitaria$$

Luego, sustituimos la w de la función logística que construye la curva y predice probabilidades de pertenencia al grupo 1, es decir, voto PAC:

$$P(voto\ PAC = 1) = \frac{e^w}{1 + e^w}$$

$$P(voto\ PAC = 1) = \frac{e^{\beta_0 + \beta_1 edad + \beta_2 sexo + \beta_3 provinciarec + \beta_4 secundaria + \beta_5 universitaria}}{1 + e^{\beta_0 + \beta_1 edad + \beta_2 sexo + \beta_3 provinciarec + \beta_4 secundaria + \beta_5 universitaria}}$$

Para estimar los parámetros en R se utiliza la función `glm()` referida a los modelos lineales generalizados. En particular, para estimar un modelo logístico, en la serie de atributos se indica `family=binomial`, con lo cual se define, automáticamente, la función de enlace logito. La estimación en R, por lo tanto, refleja la teoría anteriormente explicada de la regresión logística como parte del modelo lineal generalizado. Si se escribiera `family=gaussian`, obtendríamos la estimación para un modelo lineal con mínimos cuadrados ordinarios, de forma idéntica a `lm()`. En síntesis, utilizamos un modelo lineal generalizado específico: el logístico que se caracteriza por la distribución binomial con función logito.

La sintaxis de `glm()` es similar a `lm()`, en cuanto al orden (variable dependiente, variables independientes y datos), pero es importante indicar `na.action=na.exclude`, para un procedimiento posterior (la tabla de clasificación). Es esencial, además, que la variable dependiente esté codificada con 0 y 1. Incluyo el resultado en el objeto `modelopac`.

```
modelopac<-glm(votopac~edad+sexo+provinciarec+educarec, data=ciep,
               family=binomial, na.action=na.exclude)
summary(modelopac)
```

```
##
## Call:
## glm(formula = votopac ~ edad + sexo + provinciarec + educarec,
##     family = binomial, data = ciep, na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.7604  -0.9687  -0.6739   0.9384   1.9840
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.620352   0.331645  -4.886 1.03e-06 ***
## edad                    0.007045   0.005412   1.302   0.1930
## sexo                   -0.359212   0.154880  -2.319   0.0204 *
## provinciarec            0.744662   0.175127   4.252 2.12e-05 ***
## educarecSecundaria      0.440032   0.209787   2.098   0.0359 *
## educarecUniversitaria   1.650957   0.211642   7.801 6.16e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1100.2  on 796  degrees of freedom
## Residual deviance:  983.9  on 791  degrees of freedom
##   (172 observations deleted due to missingness)
## AIC: 995.9
##
## Number of Fisher Scoring iterations: 4
```

Los resultados de la regresión logística tienen mucha semejanza con los de la regresión lineal. Hay coeficientes estimados, errores estándar y valores p de significancia. La diferencia más importante entre ambos es que, en regresión logística, los coeficientes no se pueden interpretar de forma directa, como se hacía con la regresión lineal gaussiana. Es necesario exponenciar los coeficientes para interpretar su magnitud (exponenciar significar elevar el número e al coeficiente de regresión, es decir, e^{β_j}). Otra opción es calcular probabilidades marginales. Ambos procedimientos se dejan para una profundización posterior, más allá de este libro introductorio en el que nos centraremos en interpretar el signo del coeficiente, no su magnitud.

Según los resultados, edad tiene un coeficiente positivo. Esto significa que conforme aumenta la edad, aumenta la probabilidad de haber votado por el PAC en 2018. Este coeficiente, sin embargo, no es significativamente distinto de cero, según su valor p de 0.193. Para el coeficiente de sexo, debemos comparar el grupo codificado como 1 contra

el grupo codificado 0, es decir, hombres (1) con mujeres (0). El coeficiente es negativo, por lo que la probabilidad de haber votado por el PAC es menor entre hombres que entre mujeres o, lo que es lo mismo, las mujeres votaron más por el PAC que los hombres; este coeficiente es además significativamente distinto de cero al 0.05, pues su valor p es 0.020. Provincia también es una variable categórica, donde 1 indica provincia central y 0 provincia costera. Ya que el coeficiente es positivo, se concluye que entre las personas que habitan provincias centrales la probabilidad de haber votado por el PAC es mayor que entre las personas que votaron en provincias costeras. La variable es distinta de cero, pues su valor p es mucho menor a 0.001.

Por último, tenemos dos coeficientes para la variable educación, que corresponden a los niveles secundaria y universitaria, donde primaria o menos es la categoría base que queda por fuera, al estar codificada con el menor valor. Cada coeficiente se interpreta separadamente. El coeficiente de educación secundaria es positivo, por lo que la probabilidad de haber votado por el PAC es mayor entre personas con educación secundaria, en comparación con las personas con educación primaria o menos, con una significancia de 0.05, pues el valor p es 0.036. Entre las personas con educación universitaria, la probabilidad de haber votado por el PAC es también mayor, en comparación con las personas con educación primaria o menos, con significancia estadística al ser el valor p menor a 0.001.

En resumen, para interpretar los coeficientes de la regresión logística, centrándose en la dirección, tenemos que, si un coeficiente es positivo:

- En el caso de una variable métrica, se dice que aumenta la probabilidad del evento $y = 1$.
- En el caso de una variable categórica, se interpreta que la probabilidad de $y = 1$ es mayor entre el grupo codificado con 1 versus el grupo codificado con 0 (o la categoría base).

Por el contrario, si un coeficiente es negativo:

- Al aumentar la variable métrica, disminuye la probabilidad de $y = 1$.
- La probabilidad de $y = 1$ es menor en el grupo codificado con 1 versus el grupo codificado con 0 (o la categoría base).

Con los parámetros estimados, podemos reescribir el modelo ajustado así:

$$\hat{w} = -1.62 + 0.007edad - 0.36sexo + 0.74provinciarec$$

$$+0.44secundaria + 1.65universitaria$$

donde $\hat{w}$ es una predicción lineal, pero no una probabilidad. Para obtener probabilidades debemos utilizar la ya conocida función logística:

$$P(voto\ PAC = 1) = \frac{e^{\hat{w}}}{1 + e^{\hat{w}}}$$

De esta forma, podríamos calcular probabilidades estimadas específicas, por ejemplo, de que una mujer de 48 años, que vive en una provincia central y que tiene un nivel educativo universitario, haya votado por el PAC en 2018. Para ello utilizamos la ecuación lineal con los valores descritos, teniendo presente los valores 0 y 1 en las variables categóricas:

$$\hat{w} = -1.62 + 0.007 * 48 - 0.36 * 0 + 0.74 * 1 + 0.44 * 0 + 1.65 * 1 = 1.106$$

La predicción lineal obtenida la transformamos ahora en probabilidad:

$$P(voto\ PAC = 1) = \frac{e^{1.106}}{1 + e^{1.106}} = 0.751$$

En otras palabras, una mujer con las características descritas, según el modelo, tiene una probabilidad de 0.751 o 75.1 % de haber votado por el PAC en 2018.

7.4 Evaluación de la calidad del modelo

El modelo da resultados, pero falta evaluar qué tan confiables son. Una estrategia básica, pero útil, para evaluar la calidad del modelo es construir una tabla de clasificación. De la misma manera en que se calculó la probabilidad predicha para el ejemplo de la mujer hipotética, es posible calcular probabilidades para todas las personas en la base de datos. Es decir, cada observación (en el ejemplo, personas votantes) tiene un valor observado (voto PAC o no, en los datos) y una probabilidad predicha (voto PAC o no, en el modelo).

Para calcular las probabilidades predichas para todo el conjunto de datos recurrimos a la función `predict()` (estudiada ya en el capítulo 5), donde se asignan los valores predichos a un nuevo vector que llamo `probpred`:

```
ciep$probpred<-predict(modelopac, type="response")
```

Si se examina esta variable, se podrá ver que contiene valores entre 0 y 1, es decir, probabilidades predichas. Para comparar esta predicción con el valor observado hay que decidir un punto de corte para demarcar una predicción de voto PAC y no voto PAC. El punto habitual (aunque no el único válido) es 0.5. Es decir, decimos que las probabilidades mayores o iguales a 0.5 predicen que la persona votó por el PAC y las probabilidades menores a 0.5 que no votó por el PAC.

Con esta regla, construimos la variable de voto predicho:

```
ciep$votopredicho<-NA #variable vacía
ciep$votopredicho[ciep$probpred >=.5]<-"Predice que votó por el PAC"
ciep$votopredicho[ciep$probpred <.5]<-"Predice que no votó por el PAC"
```

Ahora cruzamos esta variable de comportamiento predicho (`votopredicho`) con la variable del comportamiento observado (`votopac`), con una tabla:

```
table(ciep$votopac, ciep$votopredicho)
```

```
##
##     Predice que no votó por el PAC Predice que votó por el PAC
##   0                            338                           91
##   1                            170                          198
```

La tabla muestra que hay 198 personas que votaron por el PAC y que el modelo predice correctamente como votantes del PAC. También hay 338 personas que no votaron por el PAC y el modelo coincide en la predicción del voto. No obstante, hay 170 personas que votaron por el PAC y el modelo predice que no votaron por dicho partido, así como 91 personas que no votaron por el PAC, aunque el modelo estima que sí lo hicieron.

Más práctico aún es calcular la tabla anterior con porcentajes totales:

```
round(prop.table(table(ciep$votopac, ciep$votopredicho))*100, 1)
```

```
##
##     Predice que no votó por el PAC Predice que votó por el PAC
##   0                           42.4                         11.4
##   1                           21.3                         24.8
```

Según esta tabla, 67.2 % (42.4 + 24.8) de los casos están correctamente predichos (personas votantes para las cuales la realidad y el modelo coinciden). Es decir, la tasa de clasificación correcta del modelo es 67.2 %, mientras que la tasa de error es 32.7 % (21.3 + 11.4).

Estos porcentajes de clasificación nos permiten no solo evaluar la calidad del modelo, sino también compararlo con otros modelos para ver si, al incluir variables adicionales, el porcentaje de clasificación correcta aumenta o la tasa de error disminuye. Por ejemplo, ¿cuánto mejoraría el modelo si se añade como variable independiente la valoración del gobierno anterior del PAC (2014-2018)? Si el porcentaje de clasificación correcta aumenta, la variable es relevante, ya que mejora la precisión predictiva del modelo. Si el porcentaje de clasificación correcta no mejora al incorporar la variable, el modelo pierde parsimonia y eficiencia sin un beneficio neto.

7.5 Comentarios finales

Este capítulo propuso la regresión logística como un modelo para tratar variables dependientes binarias o dicotómicas, es decir, de dos categorías. Si a este tipo de variable dependiente se aplicara el método de estimación de los mínimos cuadrados ordinarios, que vimos para el modelo lineal gaussiano, se pueden producir estimadores consistentes e insesgados (Wooldridge, 2010, p. 562). Dicho modelo se denomina el *modelo lineal de probabilidad* y, aunque hay quienes que lo prefieren frente al logístico, solamente bajo ciertas condiciones se obtiene la consistencia y la ausencia de sesgo en las estimaciones (Horrace y Oaxaca, 2006). Asimismo, las predicciones del modelo lineal de probabilidad pueden salirse del rango de 0 a 1, como vimos en la figura 7.1 al inicio del capítulo. En cambio, una curva como la logística se ajusta mejor a los datos que una recta. Otra curva similar se obtiene con el modelo *probit*, cuyos resultados difieren poco del modelo logístico (Glasgow y Alvarez, 2008, p. 516).

Es importante considerar algunas diferencias entre el modelo de regresión lineal gaussiano y el logístico. Primero, el logístico requiere un mayor número de observaciones que el lineal. Con muestras pequeñas (aproximadamente, menos de 100 observaciones), la regresión logística con máxima verosimilitud genera sesgo en los parámetros estimados (Rainey y McCaskey, 2021). Segundo, para la interpretación de la bondad de ajuste, en regresión logística, los paquetes estadísticos ofrecen varios pseudo R^2 que son conceptualmente distintos al coeficiente de determinación del modelo de regresión lineal. Estos pseudo R^2 pueden ayudar a comparar varios modelos estimados, pero no se interpretan como proporción de variancia explicada (Hosmer *et al.*, 2013, p. 182). Por este motivo, la tasa de clasificación correcta resulta más indicativa sobre la calidad del modelo. Tercero, la interpretación de los coeficientes en la regresión logística no es directa y requiere transformaciones para la comprensión de estos.

Por último, recordemos que el modelo logístico se aplicó con variables categóricas dicotómicas. Para variables categóricas de más de dos categorías, existen modelos como el logístico multinomial y el logístico ordinal, que también pueden estimarse en R. Por ejemplo, el análisis del voto en países multipartidistas (es decir, aquellos con más de dos partidos relevantes) requiere de modelos multinomiales (*e. g.*, el voto en Italia, ver Bellucci, 2006). El análisis de respuestas tipo Likert (escalas de respuesta desde "muy bien" hasta "muy mal") puede realizarse con modelos logísticos ordinales (como en Pignataro y Cascante Segura, 2017). Aunque la estimación de ambos modelos es poco problemática, la interpretación de los coeficientes resulta algo más complicada, especialmente en regresión ordinal.

7.6 Ejercicios

1. Utilice la base de datos "CIEPnoviembre2020.dta" para estimar un modelo de regresión logística que explique la valoración de la gestión del gobierno (variable `gestionrec`, donde 0 = negativa y 1 = positiva). Como variables independientes incluya evaluación de la situación económica (`sit_economica`, desde 1 = "muy mal" hasta 5 = "muy bien"), estado de desempleo (`desempleado`, donde 0 = no desempleado y 1 = desempleado), edad en años cumplidos (`edad`), sexo de la persona entrevistada (`sexo`, donde 0 = mujer y 1 = hombre), provincia de residencia (`provinciarec`, donde 0 = costera y 1 = central) y nivel educativo (`educarec`, donde 1 = primaria o menos, 2 = secundaria y 3 = universitaria).

2. Interprete los coeficientes obtenidos en el modelo en términos de la dirección y la significancia.

3. Calcule el porcentaje de predicción correcta y el porcentaje de error del modelo.

4. Con el modelo estimado, calcule la probabilidad de que un hombre de 35 años, que considera que la situación económica del país es mala, desempleado, residente de una provincia costera y con nivel educativo de secundaria, valore de forma positiva la gestión del gobierno.

Capítulo 8

Análisis de conglomerados

8.1 Introducción

Las técnicas de análisis de conglomerados (conocidos por el término inglés *clusters*) constituyen herramientas útiles para la descripción y exploración de datos multivariados. Clasifican observaciones o casos en grupos homogéneos, es decir, conformados por entidades con características similares. Estas clasificaciones, en primera instancia, resumen información, con lo cual se obtiene simplicidad descriptiva. Además, permiten elaborar tipologías multidimensionales, relevantes para la formación de conceptos en la investigación politológica (Collier *et al.*, 2008). En ciencia política, se ha recurrido al análisis de conglomerados para construir tipologías de electores (Fournier Facio, 2002; Pignataro y Cascante, 2018), identificar perfiles de participación política (Guzmán Castillo, 2021; Verba y Nie, 1972), clasificar partidos (Ennser, 2012) y sistemas de partidos políticos (Altman *et al.*, 2009), distinguir regímenes de bienestar (Martínez Franzoni, 2008) y clasificar palabras claves en análisis textual (Aruguete y Calvo, 2018).

Entre las numerosas técnicas de agrupamientos existentes, a continuación se examinará el *método aglomerante jerárquico*. Este se caracteriza por iniciar en un punto donde hay tantos grupos como casos. Los casos se aglomeran poco a poco en *clusters* cada vez más grandes hasta formar un grupo único con todos los casos (Hernández Rodríguez, 2013). El método aglomerante jerárquico se basa en el cálculo de distancias entre los valores respectivos de cada objeto o caso según una o más variables. Estas distancias se pueden determinar por distintas fórmulas; los próximos ejemplos utilizan la distancia euclidiana.

8.2 Método

El análisis de conglomerados parte del cálculo de distancias entre todos los objetos y las variables consideradas. Si se tienen dos objetos u observaciones, O_1 y O_2, como pueden ser partidos políticos, legisladores, países, etc., medidos con K variables, podemos calcular la disimilitud (*i. e.*, el grado de diferencia) entre los objetos a través de la *distancia euclidiana*:

$$distancia(O_1, O_2) = \sqrt{\sum_{j=1}^{k}(x_{1j} - x_{2j})^2}$$

La distancia euclidiana calcula cuán distintos son dos objetos. Cuánto más diferentes son, mayor la distancia euclidiana. Si los valores fueran iguales para todas las variables, habría máxima semejanza y la distancia sería cero.

Para ejemplificar, se tienen los siguientes datos para dos observaciones y cuatro variables:

	x1	x2	x3	x4
Objeto 1	7	1	6	8
Objeto 2	5	3	2	9

Con estos dos objetos, la distancia euclidiana se calcula:

$$distancia(O_1, O_2) = \sqrt{(7-5)^2 + (1-3)^2 + (6-2)^2 + (8-9)^2}$$

$$distancia(O_1, O_2) = \sqrt{(2)^2 + (-2)^2 + (4)^2 + (-1)^2}$$

$$distancia(O_1, O_2) = \sqrt{25} = 5$$

Existen otras medidas, aparte de la distancia euclidiana, como la distancia Manhattan, la métrica de Minkowski y la distancia generalizada de Mahalanobis (Aldenderfer y Blashfield, 1984), aunque la euclidiana es empleada con frecuencia. Sin embargo, la distancia euclidiana es apropiada solo para medir disimilitudes entre variables métricas; para variables categóricas hay otros coeficientes pertinentes.

Luego de calcular la matriz de distancias, el análisis de conglomerados de tipo aglomerante jerárquico inicia una clasificación automatizada, partiendo del par de objetos con menor distancia entre sí, para agrupar todos los demás objetos consecutivamente (Hernández Rodríguez, 2013, pp. 232-241).

Una vez completada la clasificación, el resultado se analiza en un *dendrograma* (gráfico de árbol) para interpretar, en conjunto con la teoría y el conocimiento previo, cuántos y cuáles son los grupos resultantes. Este procedimiento es más fácil de apreciar con un ejemplo, como el que se muestra a continuación.

8.3 Ejemplo

La base de datos "eleccionesCentroamerica.xlsx" incluye 35 elecciones presidenciales (de primera vuelta) en cinco países centroamericanos: Costa Rica, El Salvador, Honduras, Guatemala y Nicaragua. Para el ejemplo, nos interesa clasificar las elecciones según el grado de democracia electoral, de acuerdo con el índice que estima el proyecto *Varieties of Democracy* (https://www.v-dem.net/), y el porcentaje de participación electoral (recopilación propia del autor con base en fuentes varias de cada país). La premisa teórica es que ambas dimensiones no están perfectamente correlacionadas: podemos encontrar alta participación en democracias consolidadas, así como en democracias de bajo nivel y en regímenes autocráticos.

En primer lugar, importamos en R la base de datos en formato Excel con `read_excel()` y la asignamos en el objeto nombrado `ca`:

```
library(readxl)
ca<-read_excel("eleccionesCentroamerica.xlsx")
```

Un paso útil para el análisis de conglomerados es determinar una variable identificadora de los casos. En la base de datos, esta variable corresponde a `etiqueta`, una combinación de la abreviatura del país y el año de la elección:

```
ca$etiqueta
```

```
##  [1] "CR1990"  "CR1994"  "CR1998"  "CR2002"  "CR2006"  "CR2010"  "CR2014"
##  [8] "CR2018"  "ELS1989" "ELS1994" "ELS1999" "ELS2004" "ELS2009" "ELS2014"
## [15] "ELS2019" "GUA1990" "GUA1995" "GUA1999" "GUA2003" "GUA2007" "GUA2011"
## [22] "GUA2015" "GUA2019" "HON1993" "HON1997" "HON2001" "HON2005" "HON2009"
## [29] "HON2013" "HON2017" "NIC1990" "NIC1996" "NIC2001" "NIC2006" "NIC2011"
```

Para que la variable `etiqueta` nombre las filas de la base de datos, en lugar de la indexación numérica, podemos utilizar una función del paquete `tidyverse` que hace precisamente esta sustitución:

```
library(tidyverse)
ca<-column_to_rownames(ca, var="etiqueta")
```

Otro procedimiento, también aconsejable antes de iniciar propiamente el análisis de conglomerados, es estandarizar las variables cuando su escala de medición es muy distinta (Aldenderfer y Blashfield, 1984; Everitt y Hothorn, 2011). En el ejemplo, el índice de democracia electoral tiene valores entre 0 y 1, mientras que el porcentaje de participación electoral conceptualmente varía entre 0 y 100. Las escalas son bastante diferentes y conviene estandarizarlas para que las distancias euclidianas que se calculen no dependan de la escala de medición.

En R, recurrimos a la función `scale()` que estandariza variables, al restar la media y dividir entre la desviación estándar. Utilizamos únicamente las dos variables que nos interesan de esta base de datos original (`demelectoral` y `participacion`) y las asignamos a un nuevo objeto, `caEST`, que incluye estos vectores:

```
caEST<-scale(ca[, c("demelectoral", "participacion")])
```

Tenemos, por lo tanto, una base de datos con el índice de democracia electoral y el porcentaje de participación electoral estandarizados, con la etiqueta de la elección que nombra las filas.

```
head(caEST)
```

```
##         demelectoral participacion
## CR1990     1.531229     1.7196798
## CR1994     1.547737     1.6527058
## CR1998     1.542234     0.5923086
## CR2002     1.575248     0.4828747
## CR2006     1.597258     0.1354980
## CR2010     1.624770     0.5094980
```

El segundo paso, luego de la preparación de los datos, es calcular la matriz de distancias con la misma fórmula euclidiana que vimos antes. Por supuesto, es más eficiente calcularla en R, con la función `dist()`:

```
distancias<-dist(caEST, method="euclidean")
```

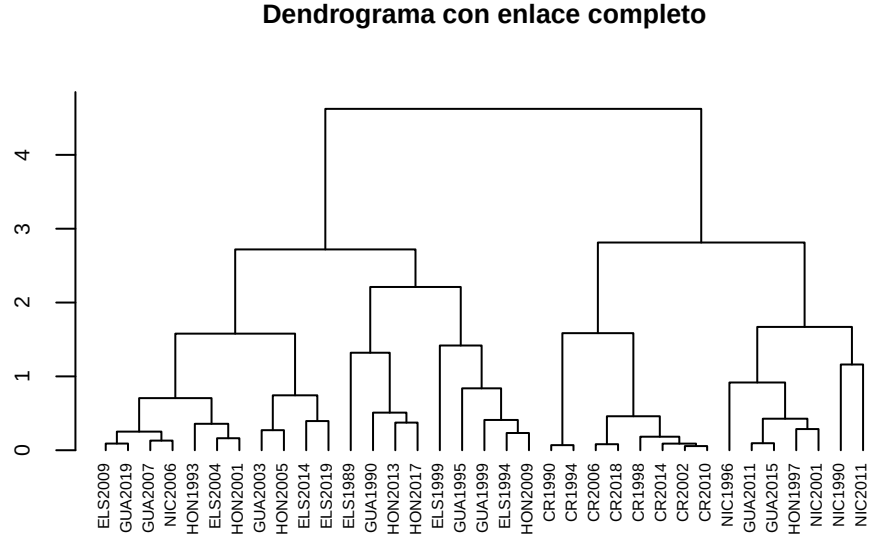
Con la matriz de distancias calculada, se procede al análisis de conglomerados con la función `hclust()`. En esta hay que escoger un método o tipo de vinculación para crear

los conglomerados. Empezamos por el método completo (*complete*), también llamado vecino más lejano, ya que utiliza la regla de buscar la mayor distancia entre objetos de un grupo y los restantes objetos para decidir el agrupamiento.

```
conglomerados1<-hclust(distancias, method="complete")
```

La forma de analizar los conglomerados en el análisis jerárquico es por medio de un dendrograma, es decir, una representación en forma de árbol del agrupamiento. En R creamos el dendrograma con `plot()`, con algunas especificaciones de formato. Con `cex` se controla el tamaño de fuente de las etiquetas de los casos, con `cex.axis` el tamaño de los valores del eje vertical y con `cex.main` el tamaño del título; al indicar un número menor a uno, disminuye el tamaño respecto a la configuración base. Con `hang` se posicionan las etiquetas de los casos; al precisarse `hang=-1`, se colocan las etiquetas debajo del eje horizontal. Con `sub=`, `xlab=` y `ylab=` se eliminan el subtítulo, el título del eje horizontal y el título del eje vertical para simplificar la presentación.

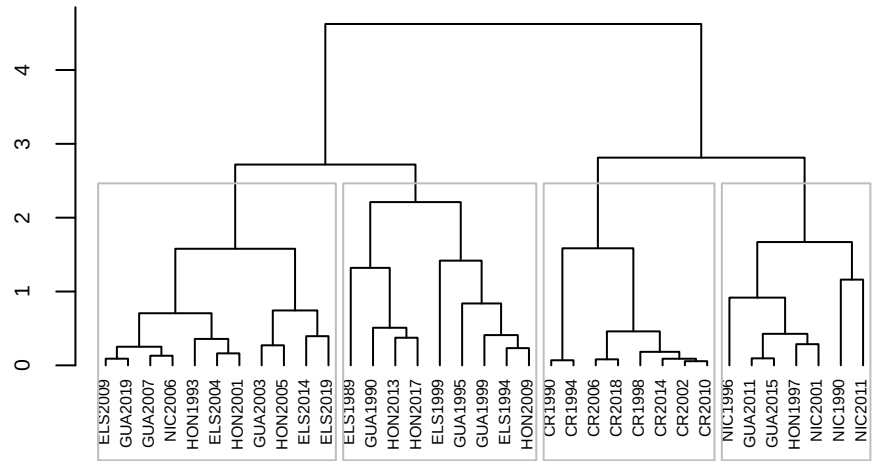
```
plot(conglomerados1, cex=.5, cex.axis=.7, cex.main=.8, hang=-1,
     main="Dendrograma con enlace completo", sub="", xlab="", ylab="")
```



El dendrograma se lee de abajo hacia arriba de la siguiente forma: los casos se agrupan según su semejanza en pares y luego en grupos mayores, hasta que todos se juntan en un único grupo de todos los casos. El arte de leer los dendrogramas, por lo tanto, consiste en saber dónde detenerse, es decir, en cuál punto la agrupación es útil y parsimoniosa, pues genera un número pequeño de grupos con casos semejantes entre sí.

La observación del gráfico y el conocimiento de los casos sugiere la existencia de cuatro grupos. Esto se consigue haciendo un corte en la altura del eje vertical alrededor de 2.5. La función `rect.hclust()` genera este corte visualmente, al indicar dónde queremos que se realice. El resultado es el mismo dendrograma, pero con rectángulos que conforman los grupos según el punto de corte escogido.

```
plot(conglomerados1, cex=.5, cex.axis=.7, cex.main=.8, hang=-1,
     main="Dendrograma con enlace completo", sub="", xlab="", ylab="")
rect.hclust(conglomerados1, h=2.5, border="grey")
```



Este dendrograma se puede guardar en la carpeta de trabajo con las funciones `jpeg()` y `dev.off()`, las cuales encierran el código del gráfico, de la siguiente forma:

```
jpeg(file="dendrogramaCA.jpg", width=5, height=4.2, units="in", res=300)
plot(conglomerados1, cex=.5, cex.axis=.7, cex.main=.8, hang=-1,
     main="Dendrograma con enlace completo", sub="", xlab="", ylab="")
rect.hclust(conglomerados1, h=2.5, border="grey")
dev.off()
```

Otra manera de agrupar los casos es por medio de la creación de una nueva variable que indique la pertenencia a los grupos, según el punto de corte escogido visualmente, con la función `cutree()`:

```
corte1<-factor(cutree(conglomerados1, h=2.5))
corte1
```

```
##  CR1990  CR1994  CR1998  CR2002  CR2006  CR2010  CR2014  CR2018 ELS1989 ELS1994
##       1       1       1       1       1       1       1       1       2       2
## ELS1999 ELS2004 ELS2009 ELS2014 ELS2019 GUA1990 GUA1995 GUA1999 GUA2003 GUA2007
##       2       3       3       3       3       2       2       2       3       3
## GUA2011 GUA2015 GUA2019 HON1993 HON1997 HON2001 HON2005 HON2009 HON2013 HON2017
##       4       4       3       3       4       3       3       2       2       2
## NIC1990 NIC1996 NIC2001 NIC2006 NIC2011
##       4       4       4       3       4
## Levels: 1 2 3 4
```

Una tabla muestra que, con el punto de corte escogido, las elecciones se aglutinan en cuatro grupos.

```
table(corte1)
```

```
## corte1
##  1  2  3  4
##  8  9 11  7
```

La función `cutree()` permite crear grupos no solo con punto de corte, sino también con un número de grupos deseado. Esto resulta particularmente pertinente cuando existen expectativas teóricas sobre el número de grupos. Por ejemplo, si se quisieran tres grupos, en lugar de basarse un punto de corte, se utiliza el argumento `k=3`:

```
corte2<-factor(cutree(conglomerados1, k=3))
table(corte2)
```

```
## corte2
##  1  2  3
##  8 20  7
```

Volviendo a la aglomeración con punto de corte 2.5, para interpretar los conglomerados es práctico calcular los promedios de cada variable según el grupo. Para ello podemos utilizar la función `aggregate()`, con la combinación de la base de datos de las variables estandarizadas y el factor que agrupa las elecciones.

```
aggregate(caEST, list(corte1), mean)
```

```
##   Group.1 demelectoral participacion
## 1       1    1.5697460     0.7122041
## 2       2   -1.0603843    -1.0863325
## 3       3   -0.1765044    -0.3147065
## 4       4   -0.1532802     1.0773044
```

Vemos que el primer grupo incluye elecciones con índices de democracia altos y con participación electoral por encima de la media (como las variables están estandarizadas, la media de las variables es cero). El segundo grupo está formado por elecciones de baja democracia y baja participación. El tercer grupo también tiene niveles de democracia y de participación por debajo de la media, pero no tan bajos como en el grupo dos. El cuarto grupo contiene elecciones de alta participación (mayor que en el grupo uno), pero con bajos índices de democracia.

Con el agrupamiento resultante podemos observar las elecciones en su escala original (no estandarizada), como se muestra en la figura 8.1. El primer grupo (a la derecha del gráfico) está constituido por elecciones en Costa Rica, que se diferencian del resto por sus mayores niveles de democracia electoral, aunque la participación electoral es variada, entre alta y media. El segundo grupo (en la parte inferior izquierda) tiene elecciones de El Salvador, Honduras y Guatemala donde la democracia y la participación son bajas. El tercer grupo está en el centro del gráfico: la participación y la democracia son intermedias. El cuarto grupo está en la parte superior a la mitad, pues son elecciones de Guatemala, Honduras y Nicaragua con mayor participación, pero con niveles medios de democracia electoral.

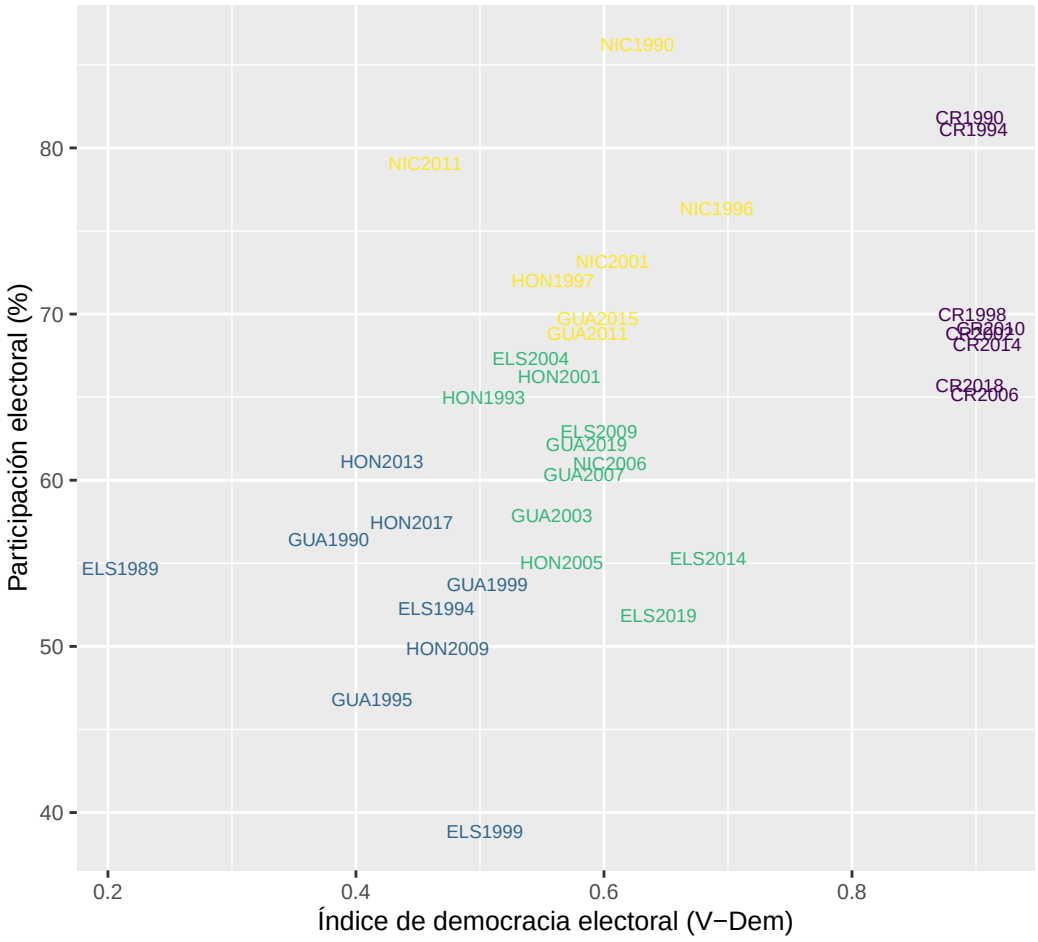
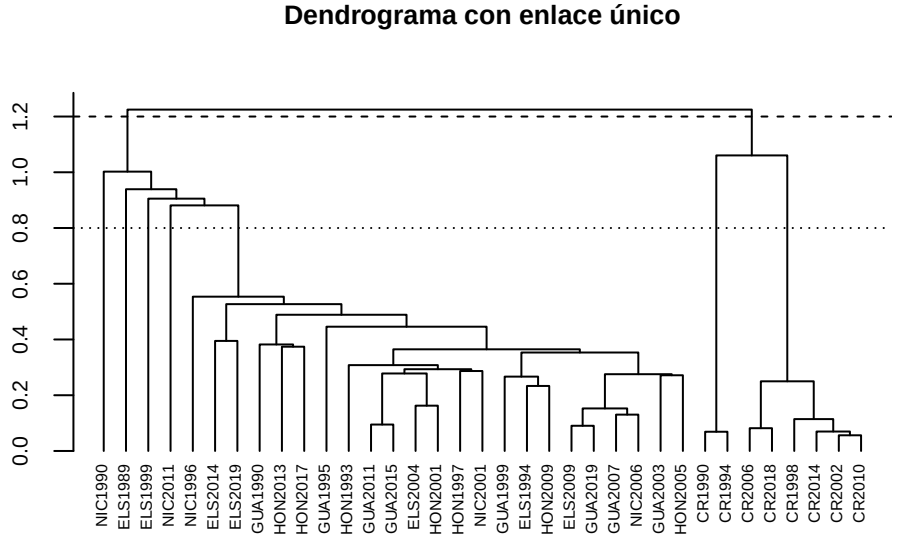


Figura 8.1 Elecciones clasificadas según el análisis de conglomerados

Este análisis se basó en la aglomeración jerárquica con el enlace completo. Sin embargo, existen otros tipos de enlace. El enlace único, llamado también vecino más cercano, calcula distancias mínimas entre los objetos de un grupo y los restantes, a diferencia del enlace completo que utiliza distancias máximas. El enlace promedio computa, precisamente, promedios entre las distancias mínima y máxima. No obstante, hay que destacar, todos los métodos agrupan según las semejanzas de los objetos, calculadas con las distancias euclidianas.

Para obtener el dendrograma, según el enlace único, aprovechamos la misma matriz de distancias ya calculada y cambiamos el método de completo (*complete*) a único (*single*) en la función `hclust()`:

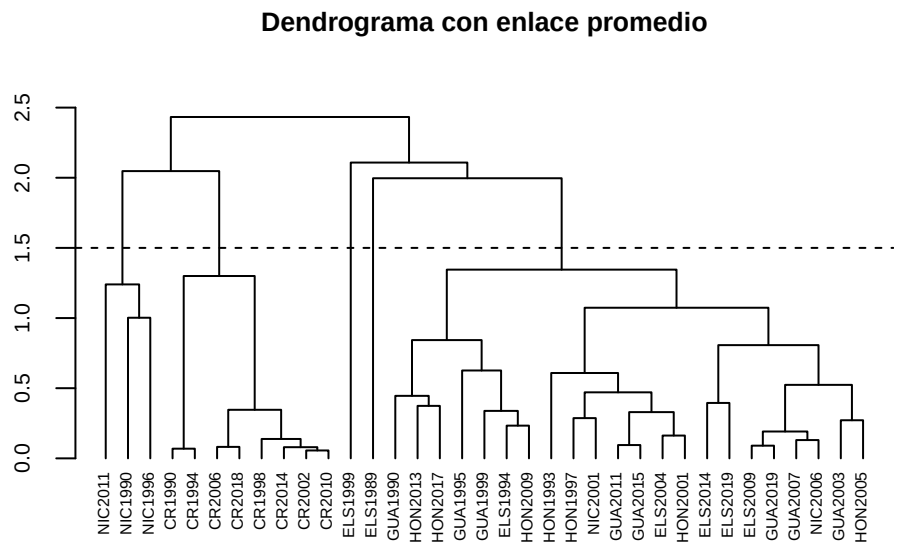
```
conglomerados2<-hclust(distancias, method="single")
plot(conglomerados2, cex=.5, cex.axis=.7, cex.main=.8, hang=-1,
     main="Dendrograma con enlace único", sub="", xlab="", ylab="")
abline(h=1.2, lty="dashed") #corte gráfico en 1.2
abline(h=0.8, lty="dotted") #corte gráfico en 0.8
```



La determinación de los conglomerados en el dendrograma con el enlace único parece menos obvia. Podríamos quedarnos con dos grupos, haciendo un corte en 1.2, pero entonces habría una gran desproporción en el tamaño de los grupos, por un lado, las elecciones de Costa Rica y, por el otro, todas las demás de Centroamérica. Si bajamos la altura y cortamos en 0.8, queda un grupo grande y varios grupos muy pequeños y de casos únicos. En este ejemplo, el enlace único no crea una clasificación parsimoniosa de los casos. De hecho, evidencia un problema común del tipo de enlace simple: formar un grupo grande y luego añadir casos uno por uno (Aldenderfer y Blashfield, 1984, p. 39).

Por último, generamos el dendrograma con el enlace promedio (*average*).

```
conglomerados3<-hclust(distancias, method="average")
plot(conglomerados3, cex=.5, cex.axis=.7, cex.main=.8, hang=-1,
     main="Dendrograma con enlace promedio", sub="", xlab="", ylab="")
abline(h=1.5, lty="dashed") #corte gráfico en 1.5
```



El dendrograma con enlace promedio sugiere que podríamos realizar el corte en 1.5 para obtener cinco grupos: uno con elecciones de alta participación en Nicaragua, otro con elecciones de Costa Rica, otro con la elección de 1999 en El Salvador, otro con la elección de 1989 en El Salvador y otro con el resto de las elecciones. Al igual que con el enlace único, hay grupos muy grandes y otros muy pequeños –y es un contrasentido respecto al propósito de conglomerar–. En conclusión, la comparación de los dendrogramas con tres métodos de enlace apunta a que el enlace completo provee la mejor clasificación de las elecciones centroamericanas, según su nivel de democracia y porcentaje de participación.

El cuadro 8.1 ofrece un resumen de los tipos de enlace según la nomenclatura usada y otra alternativa que también se puede encontrar en la literatura. La tercera columna indica, además, la forma de especificar el método en la función `hclust()` de R. Aunque en el ejemplo anterior el enlace completo ofrece la mejor representación de los datos, no siempre sucede así y resulta oportuno comparar los dendrogramas con los tres vínculos y decidir cuál es preferible en cada análisis particular.

Cuadro 8.1 Métodos de enlace en el análisis aglomerante jerárquico

Tipo de enlace	Otro nombre	Nombre en R
Completo	Vecino más lejano	complete
Único	Vecino más cercano	single
Promedio	Vinculación intergrupos	average

8.4 Comentarios finales

Este capítulo introdujo el análisis aglomerante jerárquico como un método de análisis de agrupamientos que permite clasificar un universo de casos diversos en grupos con casos homogéneos dentro de sí. Existen, sin embargo, otros algoritmos más allá del aglomerante jerárquico para clasificar casos de manera automatizada, algunos muy complejos en el mundo de *big data* y *machine learning*. Por ejemplo, el algoritmo de recomendaciones que utiliza la plataforma digital Netflix incluye métodos avanzados de agrupamiento, entre otras herramientas (Gómez-Uribe y Hunt, 2015).

Podrá notarse que el método de la aglomeración jerárquica, por medio del dendrograma, pierde utilidad cuando se amplía el número de casos, ya que la interpretación gráfica se complica cuantos más objetos existan. Con muchas observaciones, como al analizar personas en una encuesta, es preferible recurrir a otros métodos de aglomeración, como

la técnica de k medias. Además, para analizar combinaciones de variables métricas y categóricas debería recurrirse a otros métodos y algoritmos de clasificación más sofisticados. Sin embargo, en problemas de pocos casos y muchas variables, como en política comparada, la aglomeración jerárquica es una herramienta con potencial descriptivo, tipológico y analítico.

Para cerrar, debe tenerse presente que el método aglomerante jerárquico no estima parámetros ni calcula errores. Es un algoritmo matemático, no un modelo estadístico. Por esta razón, se aleja teóricamente de los métodos vistos en los capítulos anteriores.

8.5 Ejercicios

1. Calcule la distancia euclidiana para los dos objetos con las siguientes cinco variables:

	x1	x2	x3	x4	x5
Objeto 1	5.6	2.0	9.1	10.2	5.3
Objeto 2	7.8	4.3	6.4	1.8	4.9

2. Utilice la base de datos "eleccionesCentroamerica.xlsx" para clasificar las elecciones centroamericanas con las variables índice de democracia electoral (`demelectoral`) y número efectivo de partidos presidenciales (`NEPpres`). Compare las soluciones con los tres tipos de enlace y concluya cuál brinda una mejor clasificación.

Capítulo 9

Análisis exploratorio de factores

9.1 Introducción

El psicólogo Charles Spearman (1863-1945) ideó el análisis de factores como una técnica para medir la inteligencia (Bartholomew, 1995). En el proceso de creación del método, introdujo el concepto de *variable latente*, para referirse a constructos teóricamente estimables, pero no directamente observados o medibles. Desarrollos posteriores avanzaron el trabajo inicial de Spearman en el análisis de factores y sus extensiones hacia métodos más avanzados, como los modelos de ecuaciones estructurales (sobre la evolución de métodos psicométricos, ver Poole, 2008).

El objetivo principal del análisis de factores es describir una serie de datos con muchas variables mediante un número menor de variables, denominadas factores. Es, por lo tanto, una técnica de reducción de datos basada en la identificación de la variancia común de las observaciones. Por ejemplo, las personas lectoras de este libro posiblemente tengan distintos orígenes, motivaciones y personalidades; pero tienen en común que estudian (formal o informalmente) ciencia política o carreras afines. Así que podríamos, con una sola variable, como el interés en ciencias políticas y sociales, describir una porción de la diversidad de personas lectoras. Esta es la lógica del análisis de factores: distinguir matemáticamente lo común de lo individual y utilizar lo primero para ofrecer una descripción parsimoniosa de la totalidad.

Existen dos enfoques en el análisis de factores. El primero se denomina *análisis factorial exploratorio*. Este busca un resultado que se ajuste mejor a los datos sin tener demasiadas expectativas teóricas. El segundo se llama *análisis factorial confirmatorio* y permite imponer un modelo teórico a los datos. En términos metodológicos, ambos persiguen objetivos distintos. El exploratorio permite construir teoría cuando el conocimiento previo es escaso y no se tienen hipótesis sobre cuáles o cuántos son los factores. El confirmatorio se utiliza para probar teorías previamente desarrolladas, ya que existe conocimiento sobre cuáles y cuántos son los factores, es decir, hay hipótesis sobre la estructura latente (Finch, 2020).

Aparte de explorar y probar teoría, puede recurrirse al análisis de factores con otros fines. Primero, permite construir índices en los que la estructura de datos genera una ponderación más refinada de cada variable, en comparación con un simple promedio. Segundo, ayuda a resolver problemas de regresión como la multicolinealidad. Como se vio en el capítulo 6, existe multicolinealidad cuando las variables independientes están muy correlacionadas entre sí. Con el análisis de factores, se producen nuevas variables no correlacionadas que pueden incluirse como variables independientes en un modelo de regresión, con lo cual se evita la multicolinealidad. Tercero, al reducir el número de variables, se puede corregir el exceso de variables respecto al número de casos, problemático en el análisis de regresión.

Aunque los orígenes del análisis de factores están en psicología, existen numerosas aplicaciones en ciencia política. Veamos algunas. En el estudio de la participación, los análisis de factores simplifican la multiplicidad de acciones políticas. En esta línea, Theocharis y Van Deth (2018) identificaron 17 formas distintas de participación política entre la población, desde el voto hasta el boicot de productos de consumo por razones políticas. Con un análisis exploratorio de factores, logran sintetizar la diversidad de formas de participación en cinco tipos: participación institucionalizada, participación digital, protesta, voluntariado y participación de consumo. Por su parte, Altman *et al.* (2009) parten de una caracterización de partidos políticos en América Latina con base en 24 posiciones ideológicas. Con un análisis de factores reducen esta variabilidad a tres dimensiones de competencia partidaria: Estado-mercado, tradición y democracia-autoritarismo. Como última ilustración, Akkerman *et al.* (2014) identifican un factor latente de populismo (no observado) en la opinión pública, a través de siete ítems (observados) medidos en una encuesta. Con el análisis de factores, el populismo se distingue empíricamente de las dimensiones pluralista y elitista de la democracia.

Estos tres ejemplos evidencian que el análisis de factores, con su objetivo primario de reducir la variabilidad para obtener una representación parsimoniosa de los datos, se puede aplicar en múltiples temas de investigación.

9.2 Conceptos

El análisis de factores plantea un modelo estadístico para explicar la variabilidad total de los datos. El método estima variables latentes o no observadas a partir de variables medidas u observadas. Hay una semejanza con los modelos de regresión, ya que las variables observadas están en función de las variables latentes y de errores aleatorios. Sin embargo, a diferencia de los modelos de regresión, no se establece una relación entre variables independientes y variables dependientes.

En el análisis de factores, se descompone la variabilidad total de cada variable x_j, con $j = 1, ..., k$, entre variancia compartida y variancia específica. La variancia compartida, denominada comunalidad, para una variable x_1, es la variabilidad que comparte con las otras variables $x_2, x_3, ..., x_k$ a través de factores comunes. La variancia específica es exclusiva de cada x_j y se llama también unicidad. En resumen:

$$Variabilidad\ total\ de\ x_j = comunalidad_j + unicidad_j$$

El punto clave del análisis es determinar un número de factores que contengan una proporción importante de comunalidad, es decir, encontrar una amplia variancia compartida entre variables para expresarlas en unos pocos factores y que lo que no se pueda explicar por los factores (la unicidad) sea menor. Por ejemplo, si se tienen cuatro variables, se desearía encontrar un factor que explique el 70 % de la variabilidad de las cuatro. Con ello se gana eficiencia y parsimonia: no hace falta utilizar cuatro variables cuando un único factor explica el 70 % de lo que ellas originalmente contenían, sacrificando un 30 % en aras de la simplicidad.

El análisis de factores calcula la correlación entre cada variable observada y cada factor latente. Estas correlaciones se llaman cargas factoriales. Al sumarse los cuadrados de las cargas factoriales, se obtiene la comunalidad:

$$Comunalidad\ de\ x_j = (carga_{1j})^2 + (carga_{2j})^2 + (carga_{3j})^2 + ...$$

Por lo tanto, si la variabilidad total es la suma de comunalidad y unicidad, entonces:

$$Variabilidad\ de\ x_j = (carga_{1j})^2 + (carga_{2j})^2 + (carga_{3j})^2 + ... + unicidad_j = 1$$

La suma de comunalidad y unicidad es igual a uno porque las variables se estandarizan en el análisis de factores.

La teoría detrás del análisis de factores es más compleja que esta breve síntesis. En este método, no solo existe la diferencia entre exploratorio y confirmatorio, sino que además la estimación varía según el método de extracción. Además, para casos de dos factores o más se pueden aplicar rotaciones, las cuales permiten mejorar la distribución de las cargas factoriales para que estas se concentren solo en algunos factores y se facilite la interpretación. Asimismo, no hay un único método de rotación.

A continuación, se exponen dos ejemplos de análisis de factores, no con la intención de abarcar exhaustivamente el método, sino de mostrar su utilidad en dos casos prácticos dentro de la ciencia política: uno en opinión pública y otro en política comparada.

9.3 Ejemplo: valoración de instituciones

Es común que las encuestas incluyan amplias baterías de valoración sobre distintas instituciones públicas y privadas. En ocasiones, nos importa la valoración puntual a una de ellas; pero, en otras, nos interesa más el sentimiento general de las personas hacia un conjunto de objetos y temas políticos (Stimson, 2015). Así, en una encuesta del Centro de Investigación y Estudios Políticos (CIEP, 2020) se recogen valoraciones (en escala 0 a 10) hacia la Asamblea Legislativa (`nota_al`), el gobierno (`nota_gob`), los partidos políticos (`nota_pp`), el Poder Judicial (`nota_pj`), la Sala Constitucional (`nota_siv`), el Organismo de Investigación Judicial (`nota_oij`), la Universidad de Costa Rica (`nota_ucr`) y las otras universidades públicas (`nota_uni_pub`). La pregunta es si habrá una o más actitudes latentes hacia estas instituciones, en la opinión pública costarricense, para lo cual realizaremos un análisis de factores exploratorio.

En primer lugar, se carga la base de datos de la encuesta, "CIEPnoviembre2020.dta":

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
```

Para el análisis de factores requerimos la matriz de correlaciones, es decir, el conjunto de correlaciones de Pearson entre las variables observadas. La función es la misma de la correlación estudiada antes (capítulo 4), `cor()`, pero la aplicamos únicamente entre las variables que se incluyen en el análisis de factores y no en toda la base de datos. Asigno esta matriz al objeto `inst_cor`.

```
inst_cor<-cor(ciep[,c("nota_al", "nota_gob", "nota_pp",
                      "nota_pj", "nota_siv","nota_oij",
                      "nota_ucr", "nota_uni_pub")],
              use="complete.obs")
round(inst_cor, 3)
```

```
##              nota_al nota_gob nota_pp nota_pj nota_siv nota_oij nota_ucr
## nota_al        1.000    0.487   0.555   0.311    0.396    0.286    0.125
## nota_gob       0.487    1.000   0.515   0.338    0.392    0.307    0.208
## nota_pp        0.555    0.515   1.000   0.425    0.430    0.320    0.178
## nota_pj        0.311    0.338   0.425   1.000    0.720    0.679    0.429
## nota_siv       0.396    0.392   0.430   0.720    1.000    0.516    0.389
## nota_oij       0.286    0.307   0.320   0.679    0.516    1.000    0.392
## nota_ucr       0.125    0.208   0.178   0.429    0.389    0.392    1.000
## nota_uni_pub   0.142    0.218   0.251   0.445    0.420    0.445    0.720
##              nota_uni_pub
## nota_al             0.142
## nota_gob            0.218
## nota_pp             0.251
## nota_pj             0.445
## nota_siv            0.420
## nota_oij            0.445
## nota_ucr            0.720
## nota_uni_pub        1.000
```

Las correlaciones de Pearson evidencian cuáles variables están más relacionadas entre sí y entre cuáles se podrían generar los factores. Por ejemplo, la nota de la Asamblea Legislativa mantiene mayor correlación con las notas del gobierno y de los partidos políticos que con el resto. Podríamos esperar que estas tres coincidan en un factor.

En este capítulo, utilizo el paquete `psych` para el análisis de factores, aunque hay otros disponibles en R. El paquete `psych` incluye la función `KMO()` para calcular el coeficiente

Kaiser-Meyer-Olkin (KMO). Cuanto mayor es el KMO, más adecuadas son las variables en el análisis de factores. Algunos indican que un KMO mayor a 0.7 supone que el análisis de factores es pertinente (Hernández Rodríguez, 2013, p. 131), aunque los criterios pueden variar. Por ello, este valor debe tomarse como un indicador sugestivo y no como un criterio absoluto.

```
library(psych)
KMO(inst_cor)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = inst_cor)
## Overall MSA =  0.79
## MSA for each item =
##      nota_al     nota_gob      nota_pp      nota_pj     nota_siv     nota_oij
##         0.79         0.86         0.81         0.76         0.83         0.84
##     nota_ucr nota_uni_pub
##         0.72         0.73
```

El KMO global, denotado como `Overall MSA` (*measure of sampling adecuacy* o medida de adecuación muestral) es 0.79, un buen indicio. Además, para cada variable se calcula un MSO mayor a 0.7, lo que muestra que ninguna variable debe descartarse.

Con este resultado, se puede realizar con confianza el análisis de factores con la función `fa()`, pero antes habrá que tomar ciertas decisiones. Lo primero es que R requiere predeterminar un número de factores por estimar. Aunque el análisis sea exploratorio, se debería tener una idea de cuáles y cuántas serían las variables latentes.

En el ejemplo, podría pensarse que las notas del gobierno, Asamblea Legislativa y partidos políticos estén relacionadas en un único factor (como se vio en la matriz de correlaciones). Las notas del Poder Judicial, la Sala Constitucional y el Organismo de Investigación Judicial podrían también asociarse, ya que institucionalmente están vinculadas (la segunda y la tercera dependen del Poder Judicial). Por último, es previsible que las notas de la Universidad de Costa Rica y las universidades públicas se relacionen más entre sí que frente a otras instituciones. Por lo tanto, podríamos suponer la existencia de tres factores, por lo que indicamos `nfactors=3`. En el análisis se verá si este número es apropiado y se corregirá de ser necesario. Esta iteración entre presupuestos conceptuales y resultados empíricos es parte del análisis de factores exploratorio.

La segunda decisión importante se refiere al método de estimación. El método de máxima verosimilitud (*maximum likelihood*, en inglés) es considerado preferible desde el punto

de vista estadístico (Everitt y Hothorn, 2011, p. 142), aunque hay alternativas. Una muy popular es la extracción por componentes principales –la cual no debe confundirse con el análisis de componentes principales, que es un método similar, pero no igual, al análisis de factores–. Otros métodos de extracción disponibles en R son ejes principales y mínimos cuadrados ordinarios. Puede compararse los resultados de varios métodos de estimación para examinar la estabilidad de los resultados, aunque cuando estos difieren surgen dilemas sobre cuál escoger. En el ejemplo utilizamos máxima verosimilitud porque es confiable, sobre todo para muestras grandes (Finch, 2020). Especificamos este método en la función `fa()` con `fm="ml"`.

Por último, indicamos un método de rotación. Las rotaciones ayudan a interpretar mejor los resultados. Nuevamente, no hay una única forma de rotación. Algunas generan factores no correlacionados entre sí (llamadas rotaciones ortogonales); mientras que otras permiten la correlación de factores (rotaciones oblicuas). La rotación ortogonal varimax es muy común en los análisis; la especificamos como `rotate="varimax"`.

```
inst_f1<-fa(r=inst_cor, nfactors=3, fm="ml", rotate="varimax")
```

Empezamos analizando las cargas factoriales (*loadings*):

```
inst_f1$loadings
```

```
##
## Loadings:
##               ML1   ML3   ML2
## nota_al       0.146 0.736
## nota_gob      0.176 0.640 0.119
## nota_pp       0.255 0.696 0.116
## nota_pj       0.944 0.226 0.228
## nota_siv      0.607 0.385 0.264
## nota_oij      0.587 0.235 0.312
## nota_ucr      0.262       0.713
## nota_uni_pub 0.221 0.118 0.914
##
##                  ML1   ML3   ML2
## SS loadings    1.839 1.711 1.592
## Proportion Var 0.230 0.214 0.199
## Cumulative Var 0.230 0.444 0.643
```

Las columnas muestran las variables latentes (los factores), las cuales se nombran `ML1`, `ML3` y `ML2` por el método escogido (*maximum likelihood*). En cada factor se muestran las cargas factoriales estandarizadas, es decir, la correlación entre la variable observada y el factor; se omiten las de menor magnitud. Cuanto mayor es una carga factorial, más relación tiene esta variable con el factor. El primer factor, `ML1`, presenta las mayores cargas en las variables nota al Poder Judicial, nota la Sala Constitucional y nota al Organismo de Investigación Judicial. `ML3` está más correlacionada con las notas a la Asamblea Legislativa, al gobierno y a los partidos políticos. `ML2` carga en las notas de la Universidad de Costa Rica y de las otras universidades públicas.

Podemos también obtener las comunalidades y las unicidades del análisis:

```
inst_f1$communalities
```

```
##      nota_al     nota_gob      nota_pp      nota_pj     nota_siv     nota_oij
##    0.5645311    0.4548449    0.5626171    0.9950000    0.5856900    0.4968936
##     nota_ucr nota_uni_pub
##    0.5839318    0.8989599
```

```
inst_f1$uniquenesses
```

```
##      nota_al     nota_gob      nota_pp      nota_pj     nota_siv     nota_oij
##  0.435470652  0.545156625  0.437381293  0.004998389  0.414309507  0.503106282
##     nota_ucr nota_uni_pub
##  0.416067301  0.101040171
```

Como vimos en la teoría, la suma de comunalidad y unicidad es uno (aproximadamente).

```
inst_f1$communalities+inst_f1$uniquenesses
```

```
##      nota_al     nota_gob      nota_pp      nota_pj     nota_siv     nota_oij
##    1.0000018    1.0000015    0.9999984    0.9999984    0.9999995    0.9999999
##     nota_ucr nota_uni_pub
##    0.9999991    1.0000000
```

Podemos además constatar que la suma de las cargas al cuadrado es igual a la comunalidad. Para ejemplificar, con la nota al gobierno (`nota_gob`):

$$\textit{Comunalidad de nota al gobierno} = 0.176^2 + 0.640^2 + 0.119^2 = 0.455$$

Al observar las comunalidades y las unicidades, vemos contrastes entre variables. Por ejemplo, la nota al Poder Judicial tiene mucha variancia compartida (alta comunalidad) y poca variancia específica (baja unicidad); mientras que la nota al gobierno tiene la mayor variancia específica (alta unicidad) y la menor variancia compartida (baja comunalidad). Preferiblemente, las variables deben tener baja unicidad y, en análisis exploratorio, una muy alta unicidad sugiere que la variable debería excluirse del modelo.

Decíamos antes que en R hay que indicar de antemano cuántos factores se estiman. Uno de los criterios clásicos para evaluar el modelo es la presencia de autovalores propios (*eigenvalues*) mayores a uno. Estos se presentan donde dice `SS loadings`. En el ejemplo, los tres factores tienen autovalores mayores a uno: `ML1` = 1.839, `ML3` = 1.711 y `ML2` = 1.592. Tres factores son adecuados según esta medida.

La siguiente fila, `Proportion Var`, indica la proporción de la variancia original que explica cada factor. El primero contiene 0.230 (o 23 %) de la variancia, el segundo 0.214 (o 21.4 %) y el tercero 0.199 (o 19.9 %). Sumando estos tres porcentajes, sabemos que el modelo en conjunto explica el 64.3 % de la variancia original, como se muestra en `Cumulative Var`. En otras palabras, en lugar de las ocho variables observadas, podemos utilizar tres factores para representar el 64.3 % de la variabilidad en las actitudes políticas hacia instituciones públicas.

¿Qué pasaría si en lugar de tres factores estimamos cuatro? Comparemos con un nuevo análisis:

```
inst_f2<-fa(r=inst_cor, nfactors=4, fm="ml", rotate="varimax")
inst_f2$loadings
```

```
##
## Loadings:
##              ML4   ML3   ML1   ML2
## nota_al      0.733       0.106
## nota_gob     0.641 0.123 0.131 0.104
## nota_pp      0.710 0.120 0.223
## nota_pj      0.249 0.263 0.883 0.288
## nota_siv     0.400 0.289 0.556 0.185
## nota_oij     0.216 0.268 0.343 0.871
## nota_ucr           0.747 0.199 0.120
## nota_uni_pub 0.128 0.884 0.156 0.145
```

```
##
##                    ML4   ML3   ML1   ML2
## SS loadings    1.746 1.594 1.349 0.937
## Proportion Var 0.218 0.199 0.169 0.117
## Cumulative Var 0.218 0.417 0.586 0.703
```

En este análisis con cuatro factores, hay uno que presenta cargas altas solamente en la nota al Organismo de Investigación Judicial. Este último factor (`ML2`) tiene un autovalor de 0.937, menor a uno, que no es recomendable, y una variancia explicada baja de 0.117 (11.7 %). En total, el análisis de factores explica el 70.3 % de la variabilidad. Puesto que al estimar una variable latente adicional se gana poca variancia explicada (de 64.3 % a 70.3 %), el análisis con tres factores resulta preferible.

En el análisis de factores, es importante que, más allá del ajuste estadístico según el coeficiente KMO, los autovalores y la variancia explicada, los factores resultantes sean interpretables, es decir, tengan sentido teórico o conceptual. Volviendo al análisis inicial de tres factores, el primer factor, correlacionado con las notas al Poder Judicial, la Sala Constitucional y el Organismo de Investigación Judicial, contiene valoraciones de las *instituciones judiciales*. El segundo, donde las cargas mayores están en Asamblea Legislativa, gobierno y partidos políticos, se refiere a *instituciones de representación política*. El tercero claramente se refiere a *universidades públicas*. Este paso de interpretación de los factores es fundamental para evaluar la pertinencia del resultado.

9.4 Ejemplo: tipos de democracias

Arend Lijphart es un politólogo que dedicó su carrera académica al estudio de las democracias en perspectiva comparada. Uno de sus más valiosos argumentos sostiene que las democracias varían según distintas características institucionales y políticas entre dos polos: uno mayoritario y otro consensual (Lijphart, 1999). A grandes rasgos, en el modelo mayoritario el poder se concentra en una mayoría que toma las decisiones, mientras que en el modelo consensual el poder se reparte entre las instituciones y grupos en el poder. Estos dos tipos varían en grado en diez aspectos. Así, en la realidad vemos que la mayoría de los países no son ni completamente mayoritarios ni completamente consensuales, sino que sus configuraciones institucionales los ubican cercanos a uno de los dos tipos ideales.

Las diez características que permiten clasificar las democracias teóricamente se dividen en dos dimensiones (Lijphart, 1999, pp. 3-4). La primera dimensión contiene:

- número efectivo de partidos;
- concentración del Poder Ejecutivo en gabinetes mayoritarios;
- dominio del Poder Ejecutivo sobre el Poder Legislativo;
- desproporcionalidad electoral;
- organización de intereses en un sistema pluralista versus un sistema corporativista.

La segunda dimensión abarca:

- gobierno centralizado unitario versus federal y descentralizado;
- unicameralismo versus bicameralismo simétrico;
- flexibilidad para cambiar la constitución;
- revisión judicial de las constituciones por medio de las cortes;
- independencia de los bancos centrales.

Explicar cada variable escapa a los propósitos de este libro y el objetivo del ejemplo es más modesto: replicar el análisis de factores que realiza Lijphart para comprobar si efectivamente las diez variables institucionales y políticas se pueden representar en dos variables latentes. Para ello cargamos "democraciasLijphart.xlsx" que reproduce la base de datos de la segunda edición del libro (Lijphart, 2012). Los datos consideran 36 democracias en el periodo de 1945 (o primer año democrático de cada país) hasta 2010. Utilizamos además la función `column_to_rownames()`, que vimos en el capítulo 8, para que el código de país enumere las filas.

```
library(readxl)
democracias<-read_excel("democraciasLijphart.xlsx")
democracias<-column_to_rownames(democracias, var="pais")
```

Iniciamos el análisis de factores al calcular la matriz de correlaciones de Pearson y el coeficiente KMO como un indicador inicial de la relevancia de las variables. Vemos que el indicador global resulta aceptable (0.73), aunque para algunos ítems particulares es algo bajo (federalismo e independencia del banco central). Sin embargo, hay una teoría robusta detrás de las variables seleccionadas, por lo que ninguna debería eliminarse bajo puros criterios estadísticos. Por lo tanto, estimamos dos factores por medio del método de máxima verosimilitud y con rotación varimax.

```
dem_cor<-cor(democracias, use="all.obs")
KMO(dem_cor)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = dem_cor)
## Overall MSA =  0.73
## MSA for each item =
##           numpartidos    gabineteunipartido      dominioexecutivo
##                  0.70                  0.76                  0.81
##   desproporcionalidad            pluralismo           federalismo
##                  0.84                  0.80                  0.61
##         bicameralismo rigidezconstitucional      revisionjudicial
##                  0.72                  0.79                  0.75
##     indepbancocentral
##                  0.58
```

```
dem_f<-fa(r=dem_cor, nfactors=2, fm="ml", rotate="varimax")
dem_f$loadings
```

```
##
## Loadings:
##                       ML2    ML1
## numpartidos          -0.915
## gabineteunipartido    0.928
## dominioexecutivo      0.835  0.132
## desproporcionalidad   0.658
## pluralismo            0.720
## federalismo          -0.249  0.966
## bicameralismo                0.717
## rigidezconstitucional        0.600
## revisionjudicial      0.247  0.547
## indepbancocentral            0.607
##
##                  ML2   ML1
## SS loadings    3.478 2.498
## Proportion Var 0.348 0.250
## Cumulative Var 0.348 0.598
```

El primer factor presenta cargas altas (se puede ignorar el signo negativo e interpretarlas con valor absoluto) en número de partidos, gabinete unipartidario, dominio del Poder Ejecutivo, desproporcionalidad y pluralismo-corporativismo. En el segundo factor, las cargas más altas se estiman para federalismo, bicameralismo, rigidez constitucional, revisión judicial e independencia del banco central. Es importante que cada variable cargue predominantemente en un factor, como ocurre en este ejemplo. Cuando una variable presenta cargas altas en más de un factor, entonces hay problemas de ajuste. Los dos factores presentan autovalores mayores a uno (ver `SS loadings`), lo cual sugiere que ambos se deben retener y, en conjunto, explican 59.8 % de la variabilidad total.

De acuerdo con la teoría descrita, la estructura resultante refleja las previsiones teóricas. `ML2` corresponde a la dimensión que Lijphart denomina ejecutivos-partidos, mientras `ML1` es la dimensión federal-unitaria. Aunque este no es un ejemplo de análisis de factores confirmatorio, el cual requiere más especificaciones para definirse como tal, las variables observadas se relacionan correctamente con las dimensiones latentes.

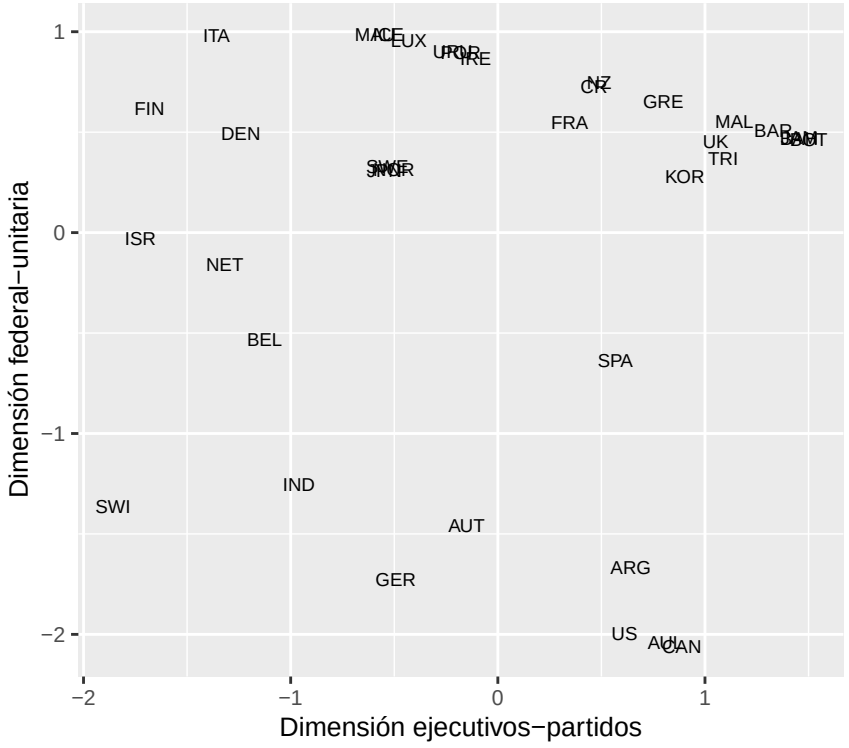


Figura 9.1 Las democracias según las dimensiones teóricas de Lijphart (1999)

La figura 9.1 utiliza puntajes factoriales, es decir, los valores que se obtienen de las variables latentes para cada caso, con el fin de graficar los países analizados en las dos dimensiones. En el eje horizontal se encuentra la dimensión ejecutivos-partidos; en el eje vertical, la dimensión federal-unitaria. En ambos, un mayor puntaje refleja diseños mayoritarios, mientras que un menor puntaje evidencia una configuración más consensual de la democracia (la escala de la segunda dimensión se revirtió para que ambos factores tengan la misma orientación conceptual). Esta reducción de diez variables en dos dimensiones refleja el objetivo principal del análisis de factores.

9.5 Comentarios finales

Este capítulo introdujo el análisis de factores y se centró en el análisis de factores de tipo exploratorio. Dicho método estima variables latentes o inobservadas a partir de variables observadas para encontrar una estructura más simple que describa adecuadamente un conjunto de datos multivariables. Sin embargo, como se mencionó, existe el análisis de factores confirmatorio, más riguroso y sofisticado. Además, el análisis de factores se conecta con los modelos de regresión al constituir los modelos de ecuaciones estructurales. Estos definen variables dependientes e independientes (como en regresión), en conjunto con variables observadas y latentes (como en análisis de factores).

Un punto importante es que el análisis de factores conlleva tomar una serie de decisiones: cuántos factores estimar, cuál método de extracción utilizar y cuál rotación incluir. Por ello es conveniente evaluar la estabilidad del modelo ante cambios en estas elecciones. Un buen modelo debería ser robusto ante variantes.

Por último, notemos que en los ejemplos se incluyen únicamente variables métricas. Las variables categóricas requieren otro tratamiento, como calcular correlaciones tetracóricas entre variables dicotómicas 0 y 1. Por ello, los modelos de ecuaciones estructurales son comparativamente ventajosos al permitir un tratamiento más flexible de variables categóricas junto con métricas.

9.6 Ejercicios

1. En el libro *Respuestas ciudadanas ante el malestar con la política* (Raventós Vorst *et al.*, 2012, p. 89), se presenta un análisis de factores de ocho ítems referidos a la disposición de participar en política a través de distintas acciones. El siguiente cuadro muestra las cargas factoriales de dicho análisis. Con base en estos datos, interprete los factores, para eso identifique cuáles variables tienen mayores cargas en cada uno y nómbrelos conceptualmente.

	Factor 1	Factor 2	Factor 3
Ayudar en la campaña de un político	0.79	-0.03	0.10
Bloquear carreteras en protesta	0.08	0.38	0.87
Denuncia ante la Defensoría de los Habitantes	0.15	0.83	0.06
Firmar una carta a políticos planteando un problema	0.67	0.21	0.08
Participar en manifestaciones	0.17	0.15	0.81
Presentar un recurso ante la Sala IV	0.15	0.80	0.13
Reunirse con un político	0.80	0.16	0.08
Reunirse con una autoridad del gobierno	0.58	0.38	0.19

2. Calcule las comunalidades y las unicidades en el punto 1.

3. En la base de datos "CIEPnoviembre2020.dta" se incluyen cinco variables referidas a apoyo al sistema político: `b1` "¿Hasta qué punto cree usted que los Tribunales de Justicia de Costa Rica garantizan un juicio justo?"; `b2` "¿Hasta qué punto tiene usted respeto por las instituciones políticas de Costa Rica?"; `b3` "¿Hasta qué punto cree usted que los derechos básicos del ciudadano están bien protegidos por el sistema político costarricense?"; `b4` "¿Hasta qué punto se siente usted orgulloso(a) de vivir bajo el sistema político costarricense?"; `b6` "¿Hasta qué punto piensa usted que se debe apoyar al sistema político costarricense?". Teóricamente se espera que los cinco ítems midan una variable latente de legitimidad política (Seligson, 2002). Estime un análisis de factores con un único factor, utilizando máxima verosimilitud. Nota: cuando se estima un factor único, no hay rotación, por lo tanto, en la función `fa()` se indica `rotation ="none"`.

4. Repita el análisis del punto 3, pero estime ahora dos factores. Compare ambos resultados y concluya cuál es preferible.

Apéndice A: Breve introducción a R

R es un lenguaje estadístico de código abierto y un programa gratuito. Se puede descargar desde https://www.r-project.org/ para distintos sistemas operativos: Windows, macOS y Linux. Es conveniente también descargar RStudio, el cual provee un ambiente fácil y amigable para trabajar en R; una versión gratuita de RStudio, también disponible para varios sistemas operativos, se encuentra en https://posit.co/download/rstudio-desktop/#download. En RStudio se escriben instrucciones y funciones en un *script* ubicado en la parte superior de la pantalla. Para ejecutar estas instrucciones se selecciona el texto deseado y se clica *Run* o se utilizan las teclas Ctrl+Enter (en Windows y Linux) o Cmd+Return (en Mac). Los resultados aparecen en la consola inferior. La captura de pantalla (figura A1) muestra la ejecución de la simple operación 1+1.

```
1+1
```

```
## [1] 2
```

R se puede utilizar para todo tipo de operaciones matemáticas; es decir, R funciona como calculadora. Puede probarse ejecutando el siguiente conjunto de operaciones:

```
10*(3+45)
```

```
## [1] 480
```

En R el signo numeral (#) indica un comentario que el programa ignora en la ejecución:

```
500/2 #la barra inclinada se usa para dividir
```

```
## [1] 250
```

```
2**3 #el doble asterisco indica potencia, es decir, 2 a la 3
```

```
## [1] 8
```

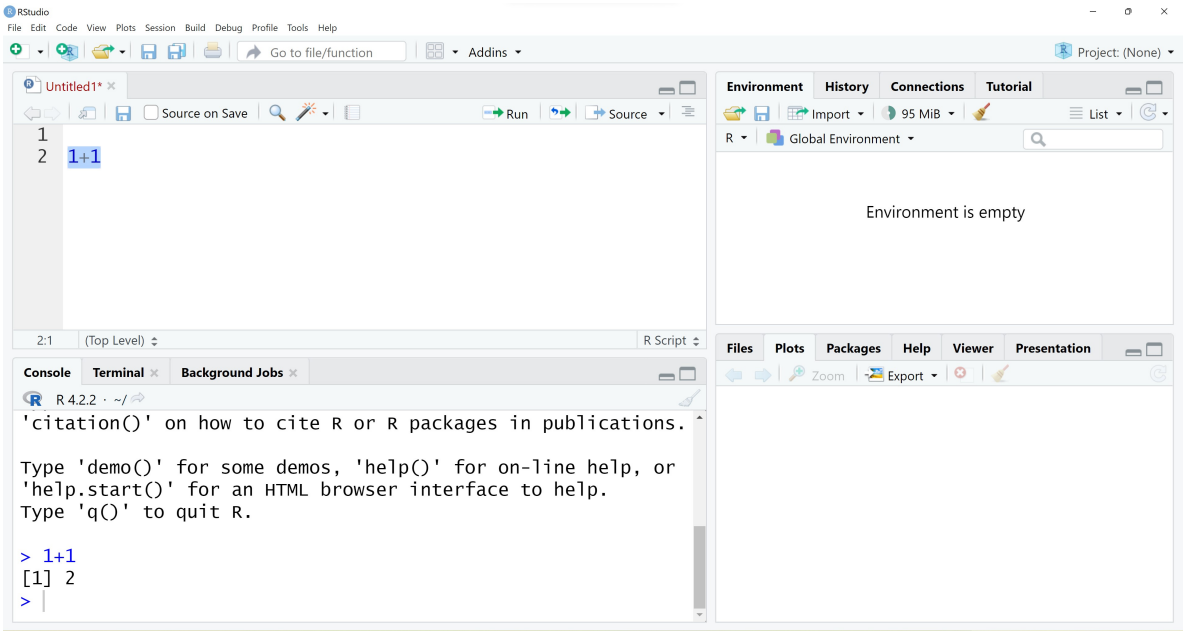


Figura A1 Imagen del entorno RStudio

```
2^3 #también corresponde a una potencia
```

```
## [1] 8
```

Existen funciones ya programas para operaciones matemáticas:

```
sqrt(64) #raíz cuadrada
```

```
## [1] 8
```

```
log(97) #logaritmo base e
```

```
## [1] 4.574711
```

```
log10(783) #logaritmo base 10
```

```
## [1] 2.893762
```

Un segundo paso en la programación de R es la creación de un objeto. Un objeto puede ser un valor, un vector, una base de datos, un gráfico, etc. Por ejemplo, podemos asignar, mediante la combinación de símbolos `<-`, el resultado de la operación 1+1 en un objeto que llamo, arbitrariamente, `ejemplo1`:

```
ejemplo1 <- 1+1
```

Para visualizar el resultado simplemente ejecutamos `ejemplo1`:

```
ejemplo1
```

```
## [1] 2
```

Lo interesante es que ahora podemos hacer operaciones con este objeto:

```
log(ejemplo1^3)
```

```
## [1] 2.079442
```

Si creamos un segundo objeto, las operaciones se pueden realizar entre objetos:

```
ejemplo2 <- 5
ejemplo3 <- ejemplo1/ejemplo2
ejemplo3
```

```
## [1] 0.4
```

Debe atenderse que R es sensible a mayúsculas. Si intentamos ejecutar `Ejemplo3`, el resultado será error porque, hasta ahora, el objeto `Ejemplo3` no existe, solamente `ejemplo3`. En general, cualquier cambio de un carácter generaría un error.

```
Ejemplo3
```

```
## Error in eval(expr, envir, enclos): object 'Ejemplo3' not found
```

```
ejmplo3
```

```
## Error in eval(expr, envir, enclos): object 'ejmplo3' not found
```

El segundo tipo de objeto es el vector. Un vector puede ser una columna de números que se establece con `c()`, donde los valores se separan con comas:

```
nacimiento <- c(1915, 1947, 1953)
```

Si un dato se desconoce o se perdió, se reporta como `NA`. Por ejemplo:

```
nacimiento <- c(1915, 1947, NA)
```

Un vector también puede ser una columna de textos que incluimos con comillas:

```
nombres <- c("Robert Dahl", "Theda Skocpol", "Pippa Norris")
```

Por último, podemos combinar los vectores en una base de datos con la función `data.frame()`, siempre y cuando estos vectores tengan la misma longitud:

```
db <- data.frame(nacimiento, nombres)
db
```

```
##   nacimiento       nombres
## 1       1915   Robert Dahl
## 2       1947 Theda Skocpol
## 3         NA  Pippa Norris
```

Lo explicado hasta ahora son funciones básicas de programación en R. Para mostrar cómo trabajar estadística descriptiva importaremos una base de datos, en lugar de construirla manualmente como en el ejemplo anterior.

Una forma sencilla de trabajar en R es importar bases de datos construidas en Excel. Para ello es importante seguir reglas básicas de construcción de bases de datos: los casos se ordenan en filas, las variables en columnas, la primera fila se destina al nombre de las variables, entre otras (sobre construcción de bases de datos en hojas de cálculo como Excel, consúltese Broman y Woo, 2018).

En el siguiente ejemplo se utiliza la base de datos "JohanSkytte.xlsx" que recopila las personas ganadoras del premio Johan Skytte –el más prestigioso galardón en ciencia política– desde 1995 hasta 2022. Para importar el archivo Excel debemos utilizar un paquete adicional, llamado `readxl` que se instala en R. Los paquetes nuevos se deben instalar una única vez, pero luego, cada vez que se utilizan, se deben cargar por medio de la función `library()`.

```
install.packages("readxl") #para la instalación la primera vez
library(readxl) #en cada sesión que se utiliza
```

Ahora podemos importar los datos. Para ello debemos conocer dónde se alojan en nuestra computadora. Una estrategia útil es contenerlos en una carpeta de trabajo donde incluimos no solo datos, sino que también alojaremos los gráficos que producimos en R. Por ejemplo, podemos definir como directorio de trabajo una carpeta existente llamada "manual":

```
setwd("C:/manual/")
```

Si en la carpeta "manual" guardamos la base de datos "JohanSkytte.xlsx", cargarla es una acción directa que realizamos con la función `read_excel()` y que asignamos en un objeto que denomino `datosjs`.

```
datosjs <- read_excel("JohanSkytte.xlsx", sheet="premios")
```

Notemos que en la función anterior especifiqué la hoja (*sheet*) llamada "premios", ya que un mismo libro de Excel puede contener múltiples hojas y hay que indicarle al programa cuál de todas queremos importar. Si no se indica la hoja, R asume que la primera del libro es la que nos interesa.

Con las funciones `names()`, `View()` y `head()` podemos explorar la base de datos. La primera nos da los nombres de las variables.

```
names(datosjs)
```

```
## [1] "premio"     "nombre"     "sexo"       "nacimiento" "muerte"
## [6] "origen"
```

La segunda, `View()`, permite visualizar la base de datos completa como si estuviéramos en Excel. Con `head()` examinamos solo las primeras filas de la base de datos:

```
head(datosjs)
```

```
## # A tibble: 6 x 6
##   premio nombre           sexo   nacimiento muerte origen
##    <dbl> <chr>            <chr>       <dbl>  <dbl> <chr>
## 1   1995 Robert Dahl      hombre       1915   2014 EEUU
## 2   1996 Juan Linz        hombre       1926   2013 Alemania/España
## 3   1997 Arend Lijphart   hombre       1936     NA Holanda
## 4   1998 Alexander George hombre       1920   2006 EEUU
## 5   1999 Elinor Ostrom    mujer        1933   2012 EEUU
## 6   2000 Fritz Scharpf    hombre       1935     NA Alemania
```

Los `NA` que se presentan en la columna `muerte` indican valores faltantes; en este caso, se presentan en las personas premiadas que no han fallecido a la fecha en que se recopiló la información (2022). Los `NA` aparecen en R porque el Excel los contenía como celdas vacías. Es decir, R asume que una celda vacía es un valor `NA`.

Para examinar cada variable, es necesario indicar a R que queremos una variable de una base de datos específica, ya que es posible tener una o más bases de datos a la vez (esta, de hecho, es una de las grandes ventajas de R). Por ejemplo, para examinar la primera variable, `premio`, escribimos:

```
datosjs$premio
```

```
## [1] 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006
2007 2008 2009
## [16] 2010 2011 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
2021 2022
```

O sea, examinamos `premio` que está contenida en `datosjs` –el signo de dólar ($) determina esta relación–. Ejecutando el código anterior, obtenemos todos los valores de la variable. Ahora bien, en otros casos es preferible resumir la variable. Para ello se utiliza la función `summary()`, con la cual se extrae el valor mínimo, el primer cuartil, la mediana, la media, el tercer cuartil y el valor máximo.

```
summary(datosjs$premio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1995    2002    2009    2009    2015    2022
```

Siguiendo la lógica de los objetos vista anteriormente, podemos crear nuevas variables a partir de las existentes. Por ejemplo, con base en `premio`, que se refiere al año en que cada persona obtuvo el Johan Skytte, y `nacimiento`, que contiene la edad de nacimiento de la persona, podemos calcular la edad que tenían el año en que recibieron el galardón al restar ambas cantidades. Queremos, además, que esta variable sea parte de la base de datos existente. Por lo tanto:

```
datosjs$edad.premio <- datosjs$premio - datosjs$nacimiento
```

Con esta nueva variable podemos calcular la edad promedio que tienen las personas al ganar el premio Johan Skytte. Este promedio se puede calcular con la aplicación de la fórmula del promedio aritmético, al sumar todas las edades con `sum()` y al dividir entre el número total de personas con `length()`; otra opción es la función `mean()` que calcula directamente el promedio. De ambas formas se obtiene el mismo resultado.

```
sum(datosjs$edad.premio)/length(datosjs$edad.premio)
```

```
## [1] 70.86207
```

```
mean(datosjs$edad.premio)
```

```
## [1] 70.86207
```

Para obtener la media redondeada a un decimal, se combina `mean()` con la función `round()`; se anida la primera dentro de la segunda y se especifica el número de decimales deseados, en este ejemplo, uno:

```
round(mean(datosjs$edad.premio), 1)
```

```
## [1] 70.9
```

Es importante que, al haber valores faltantes codificados `NA`, se especifique en la función `mean()` que estos existen, indicando `mean(..., na.rm=TRUE)`. De otra forma, el programa no computa la media. Por ejemplo, la media del año de muerte, que incluye valores `NA`, debería calcularse con `na.rm=TRUE`:

```
mean(datosjs$muerte)
```

```
## [1] NA
```

```
mean(datosjs$muerte, na.rm=TRUE)
```

```
## [1] 2015.444
```

También se puede cambiar un valor en una variable original. Si quisiéramos que, en el país de origen, en lugar de "Reino Unido" dijera la sigla "UK", utilizamos esta sintaxis:

```
datosjs$origen[datosjs$origen=="Reino Unido"]<-"UK"
```

La línea anterior se puede leer así: si `origen` es igual a "Reino Unido", cambiar por "UK" en la misma variable `origen`. El código entre paréntesis cuadrados es un argumento condicional que puede modificarse, según los fines, con los operadores lógicos igual (==), menor que (<), mayor que (>) y diferente a (!=). Por ejemplo, si deseáramos crear una nueva variable categórica, que identifique si las personas son de una primera generación, nacida antes de 1935, o si son de una segunda generación, nacida en 1935 o luego, entonces creamos una nueva variable `generacion` con base en `nacimiento`:

```
datosjs$generacion <- NA #variable vacía
datosjs$generacion[datosjs$nacimiento<1935] <- "Primera generación"
datosjs$generacion[datosjs$nacimiento>=1935] <- "Segunda generación"
```

Con `table()` obtenemos el número de casos por categoría para la variable creada:

```
table(datosjs$generacion)
```

```
##
## Primera generación Segunda generación
##                 10                 19
```

La variable `generacion` que se creó es de texto. R, sin embargo, trabaja mejor con variables tipo factor, que se codifican con números, pero se complementan con etiquetas textuales para denotar las categorías. Para construir `generacion` como factor se procede, primero, con una asignación numérica (arbitraria); segundo, con la función `factor()`, para convertir la variable creada en un factor; dentro de esta función se indican las etiquetas según el orden numérico de forma creciente, es decir, del valor más bajo al más alto. Debe cuidarse que la correspondencia entre valores y etiquetas sea la correcta, o sea, la que se pretende. Variar la relación es un error común.

```
datosjs$generacion.factor <- NA
datosjs$generacion.factor[datosjs$nacimiento<1935] <- 0
datosjs$generacion.factor[datosjs$nacimiento>=1935] <- 1
datosjs$generacion.factor <- factor(datosjs$generacion.factor,
                                    labels=c("Primera generación",
                                             "Segunda generación"))
```

El resultado es, en apariencia, igual a la variable de tipo textual; pero, en la práctica, facilita el uso.

```
table(datosjs$generacion.factor)
```

```
##
## Primera generación Segunda generación
##                 10                 19
```

Estas funciones de manipulación de datos y estadística descriptiva son tan solo una pizca en el universo del lenguaje de programación R. El objetivo era dar un primer acercamiento, no exhaustivo, que permita trabajar los métodos y modelos del texto.

Apéndice B: Temas y fuentes para profundizar

Como se señaló a lo largo de este libro, hay métodos y modelos estadísticos que no se abarcaron y otros contenidos que se podrían profundizar. A continuación, sugiero bibliografía accesible y útil para estos temas.

Primero, sobre la historia de la estadística –de la cual se ofrecen pinceladas en el capítulo 1 y en otras partes– puede leerse la muy amena obra de David Salsburg (2001), *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Sobre el desarrollo de métodos cuantitativos en la ciencia política, King (1990) ilustra las principales etapas en el análisis de datos, al distinguir el paso crucial de la recolección de datos existentes (por ejemplo, electorales) a la creación de bases de datos originales. Además, recomiendo el libro de Nate Silver (2012), *The Signal and the Noise*, un éxito de ventas que ilustra el potencial de la estadística para realizar pronósticos en campos variados, desde la política hasta los deportes y los fenómenos naturales.

La teoría de la inferencia clásica (capítulo 2) es tema de presencia obligatoria en cualquier texto introductorio a la estadística. Considero que Agresti y Franklin (2013), *Statistics: The Art and Science of Learning from Data*, es riguroso y a la vez amigable, pues ofrece explicaciones detalladas y numerosos ejemplos. Hernández Rodríguez (2015), en *Elementos de probabilidades e inferencia estadística para Ciencias Sociales*, expone las principales pruebas de hipótesis, incluidas la comparación de medias estudiada en el capítulo 3 y la prueba *chi* cuadrado del capítulo 4. Para demostraciones y mayor formalidad matemática, puede consultarse Wackerly *et al.* (2002), *Estadística Matemática con Aplicaciones*. Para explorar la amplia galería de coeficientes de asociación, más allá de los estudiados en el capítulo 4, el libro *The Measurement of Association* (Berry *et al.*, 2018) es bastante completo.

Sobre el paradigma bayesiano de la inferencia, puede leerse en Jackman (2004) una breve introducción teórica, con aplicaciones en ciencia política, y en Vallverdú (2016) un abordaje histórico y filosófico que compara ambos enfoques, bayesiano y frecuentista.

Los modelos de regresión (capítulos 5, 6 y 7) constituyen un tema amplio y, en consecuencia, la literatura es abundante. Una introducción al modelo gaussiano –similar a la expuesta en este libro, pero con mayor énfasis en los supuestos– se encuentra en Lewis-Beck y Lewis-Beck (2016). En sus diversas ediciones, *Econometría* de Gujarati y Porter (2010) es un tratamiento muy completo de los modelos de regresión (no solo el gaussiano) sin exigir niveles altos de matemática, a diferencia de otros valiosos, pero más densos, como *Econometric Analysis of Cross Section and Panel Data*, de Jeffrey Wooldridge (2010). En *Regression and Other Stories*, de Gelman *et al.* (2020), el abordaje es fresco, riguroso y aplicado; en el sitio https://avehtari.github.io/ROS-Examples/index.html se puede encontrar una versión gratuita del libro, junto con ejemplos y código para R. Sobre las interacciones, efectos compuestos de dos o más variables que no se trataron en el capítulo 6, puede consultarse el ya mencionado libro de Gujarati y Porter (2010), o bien, el artículo de Brambor *et al.* (2006).

Para profundizar en regresión logística (capítulo 7), incluyendo no solo el modelo binario sino también los ordinales y el multinomial, puede verse Hosmer *et al.* (2013), *Applied Logistic Regression*. Otros modelos lineales generalizados (ver el apéndice C) se pueden estudiar en Dobson y Barnett (2018), *An Introduction to Generalized Linear Models*, incluyendo la regresión Poisson, los modelos logísticos y los modelos de análisis de sobrevivencia. Estos últimos –también llamados modelos de duración– permiten incorporar, además de un resultado categórico (por ejemplo, aprobación/rechazo de un proyecto de ley), el tiempo hasta este resultado (días desde la presentación del proyecto hasta la aprobación). El artículo de Box-Steffensmeier y Jones (1997) es una introducción al análisis de sobrevivencia para ciencia política, mientras que el libro de Kleinbaum y Klein (2012), *Survival Analysis: A Self-Learning Text*, tiene un enfoque multidisciplinar muy completo.

Hay modelos de regresión apropiados para estructuras de datos diferentes a las abarcadas en este libro. Cuando las unidades de análisis están agrupadas, como votantes en países o partidos políticos en elecciones, la independencia entre errores que asume el modelo lineal estimado por mínimos cuadrados ordinarios se incumple, ya que las observaciones de un mismo grupo se asemejan más entre sí que con las de otros grupos.

Por lo tanto, es preferible utilizar modelos multinivel o jerárquicos. Los textos de Luke (2004) y Steenbergen y Jones (2002) son introducciones concisas al modelaje multinivel, mientras que el muy conocido Gelman y Hill (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, brinda un tratamiento más amplio, pues comprende inferencia clásica y bayesiana. Al respecto del modelaje multinivel, existe cierta controversia sobre cuántos grupos son necesarios –por ejemplo, países– para estimar modelos multinivel de efectos aleatorios. Stegmueller (2013) sugiere, para modelos estimados por máxima verosimilitud, más de 20, pero este es un debate abierto.

Como se expone en el capítulo 1, los datos agrupados con una dimensión temporal se denominan de panel o longitudinales. Su análisis requiere modelos que, como en el abordaje multinivel, consideren el agrupamiento y la interdependencia de las observaciones. En política comparada, por ejemplo, es importante considerar las características propias de cada país (institucionales, políticas, etc.), lo cual en estadística se denomina heterogeneidad entre observaciones. Con datos de panel y longitudinales hay que determinar, además, si existe una dependencia temporal. Pignataro (2018) y Wawro (2002) son artículos introductorios a los principales modelos para datos de panel, mientras que Frees (2004), *Longitudinal and Panel Data*, y nuevamente Wooldridge (2010) son libros con un amplio tratamiento del tema. Clark y Linzer (2015) ofrecen criterios para escoger entre modelos de efectos fijos y aleatorios, un dilema recurrente con datos de panel. En ciencia política es muy citado el artículo de Beck y Katz (1995); debe tenerse presente, sin embargo, que la solución que este propone aplica para datos donde la dimensión temporal es mayor a la espacial (por ejemplo, conjuntos de datos en los que hay más años que países), no el caso contrario (más países que años).

Es posible que exista dependencia espacial, tanto con datos transversales como con datos de panel o longitudinales. En estas situaciones, la cercanía geográfica es relevante, como ocurre, por ejemplo, con los procesos de difusión (*spillover effects*) en economía, política económica y políticas públicas. Para una sólida introducción a los distintos modelos de rezagos espaciales, puede consultarse el libro de Darmofal (2015), *Spatial Analysis for the Social Sciences*.

Además de las estructuras de datos transversales y longitudinales, existen series de tiempo, descritas también en el capítulo 1. Dos textos muy prácticos para el análisis de series de tiempo son *Introducción a las Series Cronológicas* (Hernández Rodríguez, 2011) y *Time Series Analysis for the Social Sciences* (Box-Steffensmeier *et al.*, 2014). El

primero contiene métodos simples de descomposición aditivos y multiplicativos, modelos ARIMA y análisis de regresión. El segundo enfatiza en modelos de regresión dinámicos (recurrentes en ciencia política), modelos no estacionarios y cointegración.

El análisis de datos multivariados incluye una enorme diversidad de métodos. Algunos textos se destinan a un método particular, como el análisis de conglomerados (Aldenderfer y Blashfield, 1984), el análisis de correspondencias (Clausen, 1998) y el análisis de factores exploratorio (Finch, 2020). Otros comprenden, en un volumen, múltiples métodos: el análisis de componentes principales, el escalamiento multidimensional, el análisis de factores confirmatorio y otros, así como los ya mencionados análisis de conglomerados y el análisis de factores exploratorio (por ejemplo, Everitt y Hothorn, 2011; Hernández Rodríguez, 2013). Puesto que la variedad puede abrumar, cuando se escoge el método es importante tener presente cuál es el objetivo del análisis multivariado: si se quiere clasificar, describir, explicar o reducir dimensiones de un fenómeno.

Entre los métodos multivariados, el análisis de factores, cuando se combina con regresión en los modelos de ecuaciones estructurales, constituye un abordaje poderoso para estimar efectos directos e indirectos con variables tanto observadas como latentes. *Generalized Latent Variable Modeling* (Skrondal y Rabe-Hesketh, 2004) es un texto referente en cuanto al modelaje con variables latentes en distintas estructuras de datos.

Por último, debe recordarse que el desarrollo de métodos de análisis cuantitativos es permanente. Ya no solo se cuenta con las técnicas clásicas, que fueron creadas para problemas propios de la agronomía, demografía y ciencias sociales y pensadas para muestras pequeñas, en comparación con la generación masiva de datos que existe actualmente. Ahora está en apogeo el análisis de datos masivos (*big data*), el análisis de texto, imagen y video como datos y muchas otras herramientas más. Sobre las implicaciones y los desafíos para la ciencia política, producto de esta explosión en volumen, variedad y velocidad de datos, puede leerse el ensayo de Brady (2019).

Apéndice C: Los modelos lineales generalizados

Este apéndice sintetiza, en un único marco teórico, la mayoría de los métodos y los modelos vistos en el libro. La perspectiva que se adopta es la de los modelos lineales generalizados (MLG), introducidos por Nelder y Wedderburn (1972), precisamente con la intención de construir un enfoque unificado a una diversidad de modelos estadísticos. Para comprender el enfoque de los MLG, se debe tener en mente que los datos se comportan según distintas distribuciones de probabilidad. Para variables métricas, la más conocida de estas es la distribución normal o gaussiana. Con datos categóricos, pueden tenerse distribuciones como la binomial y la Poisson. Una manera de distinguir las distribuciones es por la forma que adoptan al graficarse en un histograma (recuérdese la distribución normal estándar del capítulo 2).

El punto clave de los MLG es que las distribuciones –normal, binomial, Poisson y otras– forman parte de una familia llamada *exponencial*. Bajo este principio, los diversos modelos, que podrían parecer inconexos, en realidad no lo son, pues el modelo lineal generalizado los engloba. En este modelo lineal generalizado, los modelos específicos varían en la distribución de los errores y la función de enlace entre la variable dependiente y las variables independientes.

Entre los MLG con distribución normal, se tiene el modelo de regresión lineal estimado por mínimos cuadrados ordinarios (capítulos 5 y 6) y el análisis de variancia (ANOVA) (capítulo 3). Asimismo, se puede vincular la prueba t con los MLG, en tanto esta prueba es un caso particular del ANOVA con dos medias. Por su parte, el coeficiente de correlación de Pearson (capítulo 4) está íntimamente ligado con la regresión lineal (recuérdese que el coeficiente de determinación en regresión es igual al coeficiente de Pearson al cuadrado).

La regresión logística (capítulo 7) forma parte de los MLG porque se puede expresar como un modelo lineal con una función logito que enlaza las variables independientes con la dependiente; su variable dependiente sigue una distribución binomial de 0 y 1. La prueba *chi* cuadrado (capítulo 4) tampoco es lejana de la perspectiva de los MLG, ya que puede asumirse que una tabla cruzada está compuesta por distribuciones Poisson independientes, que pertenecen a la familia exponencial.

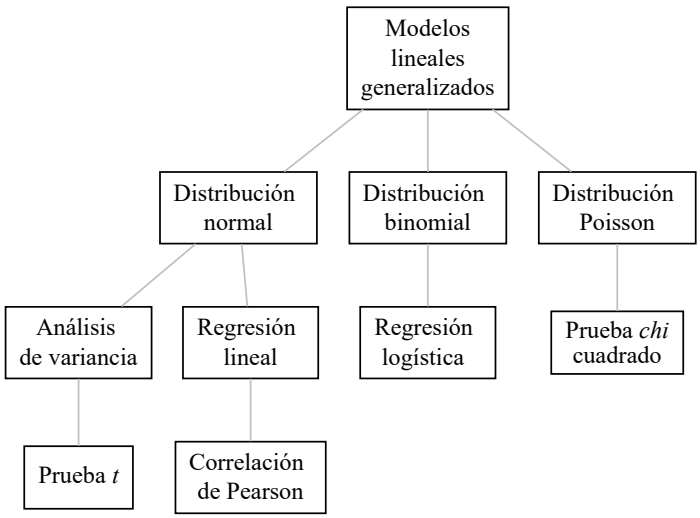


Figura A2 Esquema conceptual de los modelos lineales generalizados

El análisis factorial (capítulo 9), aunque no se considera dentro de los MLG, presenta una estructura similar al modelo de regresión normal, puesto que es posible entender las variables observadas como dependientes de las variables explicativas latentes que corresponden a los factores (Finch, 2020). Además, cuando se estima por medio de máxima verosimilitud, se asume una distribución multinormal, que es parte de la familia exponencial. Sin embargo, la técnica de aglomeración jerárquica (capítulo 8) queda fuera del esquema de los MLG, pues, como se explicó en su oportunidad, se define como un algoritmo matemático, no como un modelo estadístico.

Respuestas a los ejercicios

Capítulo 2

1. El intervalo [315060, 337906] contiene el valor real del ingreso per cápita promedio de julio de 2020, en Costa Rica, con una confianza del 95 %.

2. El intervalo [25.0 %, 27.4 %] contiene el valor real del porcentaje de los hogares pobres en Costa Rica en 2020, con una confianza del 95 %.

3. El margen de error es ±0.16. El intervalo [6.4, 6.8] contiene el valor real de la nota promedio que las personas le dan al Tribunal Supremo de Elecciones, con una confianza del 95 %.

4. El margen de error en la estimación de la intención de voto para el Partido Restauración Nacional es ±2.6; y para el Partido Acción Ciudadana es ±2.2. El intervalo [14.4 %, 19.6 %] contiene la intención de voto en enero de 2018 para el Partido Restauración Nacional con una confianza del 95 %. El intervalo [8.8 %, 13.2 %] contiene la intención de voto en enero de 2018 para el Partido Acción Ciudadana con una confianza del 95 %.

5. Los intervalos no se traslapan, lo cual sugiere que los apoyos electorales son diferentes.

6. En el censo no se calculan márgenes de error porque no se realiza inferencia estadística. Los datos del censo son poblacionales, con errores no estimables.

7. a) El estadístico calculado es: $\frac{6.2-6.3}{\frac{1.2}{\sqrt{600}}} = -2.04$. Luego se obtiene el valor p para la cola izquierda con R. La evidencia es baja para la hipótesis nula y favorece la hipótesis alternativa de que el promedio de la valoración de las carreteras es menor a 6.3.

```
pnorm(-2.04, lower.tail=TRUE)
```

```
## [1] 0.02067516
```

b) El estadístico calculado se obtiene así: $\frac{6.2-6.4}{\frac{1.2}{\sqrt{600}}} = -4.08$. Se obtiene el valor p para la cola derecha. Hay bastante evidencia a favor de la hipótesis nula de que el promedio es igual a 6.4 en contra de la hipótesis alternativa de que es mayor a 6.4.

```
pnorm(-4.08, lower.tail=FALSE)
```

```
## [1] 0.9999775
```

Capítulo 3

1. Para calcular el intervalo de confianza de la diferencia primero se obtienen las medias, las desviaciones estándar y el tamaño de muestra para cada grupo.

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
library(tidyverse)
nota_al_sexo<-ciep%>%
  group_by(sexo)%>%
  summarize(Media=mean(nota_al, na.rm=TRUE),
            DevEst=sd(nota_al, na.rm=TRUE),
            Muestra=sum(!is.na(nota_dh)))
nota_al_sexo
```

```
## # A tibble: 3 x 4
##   sexo        Media DevEst Muestra
##   <dbl+lbl>   <dbl>  <dbl>   <int>
## 1  0 [Mujer]   4.40   2.61     459
## 2  1 [Hombre]  4.55   2.29     436
## 3 NA           3.67   3.21       3
```

$$4.40 - 4.55 = \pm 1.96 * \sqrt{\frac{2.61^2}{459} + \frac{2.29^2}{436}} = -0.15 \pm 0.32$$

El intervalo [-0.47, 0.17] contiene el valor real de la diferencia de los promedios de valoración de la Asamblea Legislativa entre mujeres y hombres. Puesto que el intervalo contiene el cero, no puede decirse que las medias sean estadísticamente diferentes.

2. El valor p de la prueba t es 0.351. Por lo tanto, no puede rechazarse la hipótesis nula de que las medias son iguales. Mujeres y hombres no difieren significativamente en su valoración de la Asamblea Legislativa.

```
t.test(nota_al~sexo, data=ciep)
```

```
##
## Welch Two Sample t-test
##
## data: nota_al by sexo
## t = -0.93285, df = 913.7, p-value = 0.3511
## alternative hypothesis: true difference in means between group 0
and group 1 is not equal to 0
## 95 percent confidence interval:
## -0.4682374 0.1665213
## sample estimates:
## mean in group 0 mean in group 1
## 4.402954 4.553812
```

3. El valor p del análisis de variancia es 0.0001. Se puede rechazar la hipótesis nula de que las medias son iguales. Por lo tanto, la valoración promedio de la Defensoría de los Habitantes varía según el nivel educativo de las personas.

```
summary(aov(nota_dh~factor(educarec), data=ciep))
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## factor(educarec)   2    127   63.75   9.211 0.00011 ***
## Residuals        895   6194    6.92
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 71 observations deleted due to missingness
```

4. Si se utiliza el análisis de variancia, el valor p resultante es muy similar al de la prueba t y la conclusión es la misma: no existe una diferencia estadísticamente significativa.

Capítulo 4

1. a) Entre las personas con educación universitaria, 41.5 % aprueba positivamente la gestión del gobierno.

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")

tabla2<-table(ciep$educarec, ciep$gestionrec)
round(prop.table(tabla2, 1)*100, 1)
```

```
##
##        0    1
##   1 71.7 28.3
##   2 67.7 32.3
##   3 58.5 41.5
```

b) Entre las personas que valoran negativamente la gestión del gobierno, 26.4 % tiene educación primaria o menos.

```
round(prop.table(tabla2, 2)*100, 1)
```

```
##
##        0    1
##   1 26.4 19.7
##   2 41.7 37.6
##   3 31.9 42.7
```

c) La prueba *chi* cuadrado arroja un valor $p = 0.002$. Por lo tanto, se rechaza la hipótesis nula de independencia y se concluye que las variables nivel educativo y valoración del gobierno están asociadas.

```
chisq.test(tabla2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla2
## X-squared = 12.083, df = 2, p-value = 0.002378
```

d) Al ser la V de Cramer 0.1125, la relación se puede considerar débil.

```
cramerV(tabla2)
```

```
## Cramer V
##   0.1125
```

2. a) La correlación entre la participación en febrero y la participación en abril es 0.99, prácticamente perfecta y positiva, es decir, hay alta estabilidad en la participación por cantón.

```
library(readxl)
eleccion18<-read_excel("eleccionesCR2018.xlsx")
cor(eleccion18$febparticipacion, eleccion18$abrilparticipacion,
    use="all.obs")
```

```
## [1] 0.9902254
```

b) Según el coeficiente de correlación de 0.23, la relación entre el apoyo al PLN en febrero y el PAC en abril es positiva, pero baja. Esto indica que cuanto mayor fue el voto por el PLN en febrero, mayor fue el voto para el PAC en abril, aunque la relación no es fuerte.

```
cor(eleccion18$febPLNporcentaje, eleccion18$abrilPACporcentaje,
    use="all.obs")
```

```
## [1] 0.2318382
```

Capítulo 5

1. El coeficiente de la pendiente es 0.542 y su valor p menor a 0.001. Al aumentar un punto porcentual el voto al partido que obtiene la presidencia, la bancada legislativa de este partido se incrementa en 0.5 escaños en promedio; este coeficiente es estadísticamente distinto de cero ($p < 0.001$). El modelo explica el 71 % de la variación en el número de legisladores del partido de gobierno en Costa Rica. Se puede concluir que la elección presidencial afecta en buena medida el resultado legislativo: cuanto mayor es el apoyo al partido ganador de la presidencia, mayor resulta la bancada legislativa que lo respalda. En cambio, cuanto menor sea el voto para el partido que obtiene la presidencia, hay más probabilidad de un gobierno dividido, es decir, una presidencia sin mayoría en el parlamento.

```
votospres<-c(64.7,46.4,50.3,50.5,54.8,43.4,50.5,58.8,52.3,
             51.5,49.6,47.0,38.6,40.9,46.9,30.6,21.6,16.8)
diputadosgob<-c(30,10,29,26,32,27,27,33,29,29,28,27,19,25,24,13,10,10)
presidenciales<-data.frame(diputadosgob, votospres)

modelodiputados<-lm(diputadosgob~votospres, data=presidenciales)
summary(modelodiputados)
```

```
##
## Call:
## lm(formula = diputadosgob ~ votospres, data = presidenciales)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3798  -0.8699   1.5406   2.1968   4.2456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.75997    4.07883  -0.186    0.855
## votospres    0.54181    0.08716   6.216 1.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.357 on 16 degrees of freedom
## Multiple R-squared:  0.7072, Adjusted R-squared:  0.6889
## F-statistic: 38.64 on 1 and 16 DF,  p-value: 1.235e-05
```

Capítulo 6

1. Se estima el siguiente modelo de regresión múltiple:

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
modelopp<-lm(nota_pp~nota_gob+nota_al+edad+sexo+provinciarec, data=ciep)
summary(modelopp)
```

```
##
## Call:
## lm(formula = nota_pp ~ nota_gob + nota_al + edad + sexo + provinciarec,
##     data = ciep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7260 -1.2064 -0.0123  1.1694  6.6283
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.440503   0.261569   5.507 4.80e-08 ***
## nota_gob       0.310063   0.027859  11.130  < 2e-16 ***
## nota_al        0.400300   0.030829  12.984  < 2e-16 ***
## edad          -0.027270   0.004475  -6.094 1.65e-09 ***
## sexo           0.158179   0.134093   1.180    0.238
## provinciarec -0.016640   0.147645  -0.113    0.910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.972 on 871 degrees of freedom
##   (92 observations deleted due to missingness)
## Multiple R-squared:  0.4126, Adjusted R-squared:  0.4092
## F-statistic: 122.4 on 5 and 871 DF,  p-value: < 2.2e-16
```

2. Por cada punto que aumenta la nota del gobierno, la nota promedio a los partidos aumenta 0.31 en promedio, con las demás variables constantes. Por cada punto que incrementa la nota de la Asamblea Legislativa, la nota promedio de los partidos aumenta 0.4, manteniendo constantes las demás variables. Con cada año de edad cumplido, la nota de los partidos políticos disminuye 0.03 en promedio (constantes las otras variables). Entre los hombres, la nota promedio de los partidos es 0.16 mayor que entre las mujeres. Entre personas que residen en provincias centrales, la nota de los partidos políticos es 0.02 menor en promedio que entre personas que viven en provincias costeras. Los coeficientes significativamente distintos de cero son nota del gobierno, nota de la Asamblea Legislativa y edad. Según el R^2 ajustado, el modelo explica el 41 % de la variabilidad en la nota a los partidos.

3. Podría existir un sesgo de variable omitida al no incluir la simpatía partidaria, pues personas simpatizantes valorarían mejor a los partidos políticos y, a la vez, la simpatía partidaria estaría asociada con la nota al gobierno y la nota a la Asamblea Legislativa. La multicolinealidad se podría presentar debido a la inclusión de la nota del gobierno y la nota de la Asamblea Legislativa como variables independientes correlacionadas entre sí. Es posible que la nota del gobierno y la nota de la Asamblea Legislativa sean variables endógenas, ya que la valoración de los partidos políticos podría influir recíprocamente en estas dos.

Capítulo 7

1. Se estima el siguiente modelo de regresión logística:

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
ciep$educarec<-as_factor(ciep$educarec, levels="labels")

modelogestion<-glm(gestionrec~sit_economica+desempleado+edad+sexo+
                   provinciarec+educarec, data=ciep,
                   family=binomial, na.action=na.exclude)
summary(modelogestion)
```

```
##
## Call:
## glm(formula = gestionrec ~ sit_economica + desempleado + edad +
##     sexo + provinciarec + educarec, family = binomial, data = ciep,
##     na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9260  -0.8788  -0.6570   1.1397   2.0486
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.869313   0.374971  -7.652 1.98e-14 ***
## sit_economica          0.861703   0.098035   8.790  < 2e-16 ***
## desempleado           -0.171415   0.193538  -0.886  0.37578
## edad                   0.006731   0.005148   1.308  0.19101
## sexo                  -0.370459   0.150678  -2.459  0.01395 *
## provinciarec           0.412547   0.170078   2.426  0.01528 *
## educarecSecundaria     0.265985   0.206580   1.288  0.19790
## educarecUniversitaria  0.634168   0.207233   3.060  0.00221 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1197.7  on 928  degrees of freedom
## Residual deviance: 1084.6  on 921  degrees of freedom
##   (40 observations deleted due to missingness)
## AIC: 1100.6
##
## Number of Fisher Scoring iterations: 4
```

2. Al mejorar la valoración de la situación económica nacional, aumenta la probabilidad de valorar positivamente la gestión del gobierno, con significancia estadística de 0.001. Entre personas desempleadas, la probabilidad de valorar positivamente la gestión gubernamental es menor, pero el coeficiente no es significativamente distinto de cero. A mayor edad, mayor probabilidad de valorar de forma positiva la gestión del gobierno, pero el coeficiente no es diferente de cero. Entre los hombres, la probabilidad de valorar positivamente la gestión es menor que entre mujeres, con un nivel de significancia de 0.05. Entre habitantes de provincias centrales, la probabilidad de valorar positivamente la gestión del gobierno es mayor que entre habitantes de provincias costeras, con un nivel de significancia de 0.05. Entre personas con educación secundaria, la probabilidad de aprobar la gestión del gobierno es mayor que entre personas con educación primaria o menos, pero sin significancia estadística. Entre personas con educación universitaria, la probabilidad de valorar positivamente la gestión del gobierno es mayor que entre personas con educación primaria o menos; el coeficiente es significativamente distinto de cero al 0.01.

3. El porcentaje de predicción correcta es 70 % y el porcentaje de error es 30 %.

```
ciep$probpred<-predict(modelogestion, type="response")
ciep$gestionrecpredicha<-NA
ciep$gestionrecpredicha[ciep$probpred >=.5]<-"Predice gestión positiva"
ciep$gestionrecpredicha[ciep$probpred <.5]<-"Predice gestión negativa"
round(prop.table(table(ciep$gestionrec, ciep$gestionrecpredicha))*100, 1)
```

```
##
##     Predice gestión negativa Predice gestión positiva
##   0                     59.2                      6.2
##   1                     23.8                     10.8
```

4. La probabilidad de que una persona con estas características valore de forma positiva la gestión del gobierno es 0.236 o 23.6 %.

Capítulo 8

1. La distancia euclidiana es 9.39.

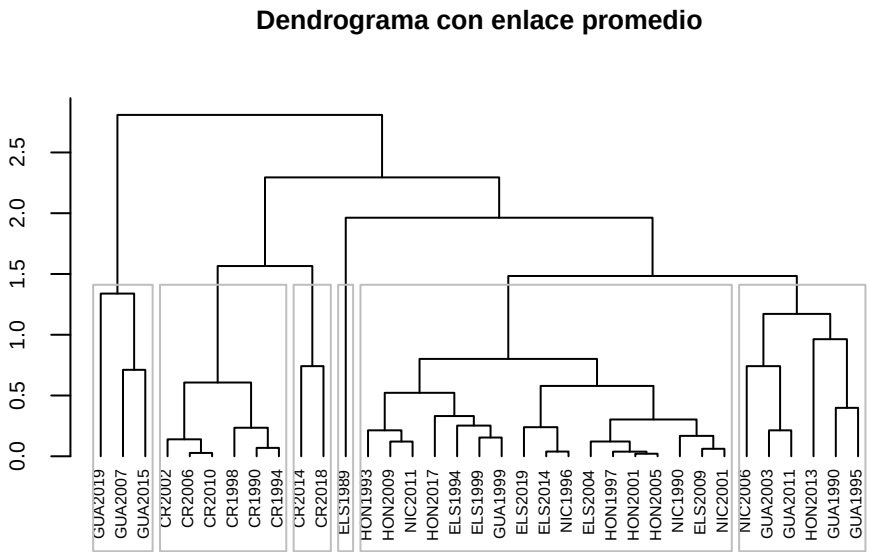
2. La comparación de los tres dendrogramas muestra que el enlace promedio es el que mejor agrupa las elecciones. El enlace único genera grupos muy pequeños y de caso único, junto con un grupo grande de elecciones. El enlace completo genera grupos más balanceados, pero deja a Guatemala 2019 y El Salvador 1989 sin agrupar. El enlace promedio, a una altura de 1.4, genera seis grupos más claramente identificables: elecciones de Guatemala de democracia intermedia y alta participación; elecciones de Costa Rica con menor número efectivo de partidos; elecciones de Costa Rica con mayor número efectivo de partidos; El Salvador 1989 como una elección atípica por el bajo nivel de democracia; varias elecciones con democracia electoral intermedia y bajo número de partidos; y varias elecciones de democracia intermedia y mayor número de partidos políticos.

```
library(readxl)
ca<-read_excel("eleccionesCentroamerica.xlsx")

library(tidyverse)
ca<-column_to_rownames(ca, var="etiqueta")

caEST<-scale(ca[, c("demelectoral", "NEPpres")])
distancias<-dist(caEST, method="euclidean")
conglomerados<-hclust(distancias, method="average")

plot(conglomerados, cex=.5, cex.axis=.7, cex.main=.8, hang=-1,
     main="Dendrograma con enlace promedio", sub="", xlab="", ylab="")
rect.hclust(conglomerados, h=1.4, border="grey")
```



Capítulo 9

1. El primer factor presenta cargas factoriales altas en las variables "ayudar en la campaña de un político", "firmar una carta a políticos planteando un problema", "reunirse con un político" y "reunirse con una autoridad del gobierno"; puede interpretarse como *contacto político*. El segundo factor se correlaciona con "denuncia ante la Defensoría de los Habitantes" y "presentar un recurso ante la Sala IV"; puede denominarse *denuncia institucional*. El tercer factor carga en "bloquear carreteras en protesta" y "participar en manifestaciones"; se puede definir como *protesta política*.

2. Comunalidades: 0.64, 0.91, 0.72, 0.5, 0.71, 0.68, 0.67, 0.52. Unicidades: 0.36, 0.09, 0.28, 0.5, 0.29, 0.32, 0.33, 0.48.

3. El coeficiente KMO general y los individuales para cada ítem son todos iguales o mayores a 0.76, lo cual se considera apropiado para el análisis de factores. El factor tiene un autovalor mayor a uno y explica 45.6 % de la variabilidad original. Todas las variables presentan cargas altas en este factor.

```
library(haven)
ciep<-read_dta("CIEPnoviembre2020.dta")
sispol_cor<-cor(ciep[,c("b1","b2","b3","b4","b6")],
                use="complete.obs")
```

```
library(psych)
KMO(sispol_cor)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = sispol_cor)
## Overall MSA =  0.8
## MSA for each item =
##   b1   b2   b3   b4   b6
## 0.83 0.87 0.79 0.76 0.78
```

```
sispol_fa1<-fa(r=sispol_cor, nfactors=1, fm="ml", rotate="none")
sispol_fa1$loadings
```

```
##
## Loadings:
##    ML1
## b1 0.580
## b2 0.536
## b3 0.743
## b4 0.757
## b6 0.727
##
##                  ML1
## SS loadings    2.279
## Proportion Var 0.456
```

4. Si se estiman dos factores, con el método de máxima verosimilitud y la rotación varimax, ambos presentan autovalores mayores a uno y la variancia explicada aumenta a 57 %. El primer factor carga en `b1`, `b2` y `b3`; el segundo en `b6`; ambos tienen correlaciones similares con `b4`. Sin embargo, al no haber una interpretación conceptualmente clara de los dos factores, la estimación de un factor resulta preferible.

```
sispol_fa2<-fa(r=sispol_cor, nfactors=2, fm="ml", rotate="varimax")
sispol_fa2$loadings
```

```
##
## Loadings:
##    ML2   ML1
## b1 0.546 0.241
## b2 0.458 0.264
## b3 0.840 0.256
## b4 0.488 0.518
## b6 0.299 0.921
##
##                  ML2   ML1
## SS loadings    1.540 1.311
## Proportion Var 0.308 0.262
## Cumulative Var 0.308 0.570
```

Referencias

Achen, Christopher H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, *22*(4), 327-339.

Agresti, Alan y Franklin, Christine. (2013). *Statistics: The Art and Science of Learning from Data*. Pearson.

Akkerman, Agnes, Mudde, Cas y Zaslove, Andrej. (2014). How populist are the people? Measuring populist attitudes in voters. *Comparative Political Studies*, *47*(9), 1324-1353.

Aldenderfer, Mark S. y Blashfield, Roger K. (1984). *Cluster Analysis*. SAGE.

Aldrich, John. (2005). Fisher and regression. *Statistical Science*, *20*(4), 401-417.

Almond, Gabriel A. (1999). *Una disciplina segmentada: Escuelas y corrientes en las ciencias políticas*. Fondo de Cultura Económica.

Altman, David, Luna, Juan Pablo, Piñeiro, Rafael y Toro, Sergio. (2009). Partidos y sistemas de partidos en América Latina: Aproximaciones desde la encuesta a expertos 2009. *Revista de Ciencia Política*, *29*(3), 775-798.

Anderson, William D., Box-Steffensmeier, Janet M. y Sinclair-Chapman, Valeria. (2003). The keys to legislative success in the US House of Representatives. *Legislative Studies Quarterly*, *28*(3), 357-386.

Aruguete, Natalia y Calvo, Ernesto. (2018). Time to #Protest: Selective exposure, Cascading Activation, and Framing in Social Media. *Journal of Communication*, *68*(3), 480-502.

Barberá, Pablo, Casas, Andreu, Nagler, Jonathan, Egan, Patrick J., Bonneau, Richard, Jost, John T. y Tucker, Joshua A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, *113*(4), 883-901.

Bartholomew, David J. (1995). Spearman and the origin and development of factor analysis. *British Journal of Mathematical and Statistical Psychology*, *48*(2), 211-220.

Beach, Derek y Pedersen, Rasmus Brun. (2013). *Process-Tracing Methods: Foundations and Guidelines*. University of Michigan Press.

Beck, Nathaniel y Katz, Jonathan N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, *89*(3), 634-647.

Bellucci, Paolo. (2006). Tracing the cognitive and affective roots of party competence: Italy and Britain, 2001. *Electoral Studies*, *25*(3), 548-569.

Bellucci, Paolo y Lewis-Beck, Michael. (2011). A stable popularity function? Cross-national analysis. *European Journal of Political Research*, *50*(2), 190-211.

Berinsky, Adam J. (2009). *In Time of War: Understanding American Public Opinion from World War II to Iraq*. University of Chicago Press.

Berinsky, Adam J. (2017). Measuring public opinion with surveys. *Annual Review of Political Science*, *20*, 309-329.

Berry, Kenneth J., Johnston, Janis E. y Mielke, Paul W. (2018). *The Measurement of Association: A Permutation Statistical Approach*. Springer.

Berry, William D. (1984). *Nonrecursive Causal Models*. SAGE.

Bhattacharya, Ananyo. (2021). *The Man from the Future: The Visionary Life of John von Neumann*. W.W. Norton & Company.

Box-Steffensmeier, Janet M., Brady, Henry E. y Collier, David (Eds.). (2008). *The Oxford Handbook of Political Methodology*. Oxford University Press.

Box-Steffensmeier, Janet M., Freeman, John R., Hitt, Matthew P. y Pevehouse, Jon. (2014). *Time Series Analysis for the Social Sciences*. Cambridge University Press.

Box-Steffensmeier, Janet M. y Jones, Bradford S. (1997). Time is of the essence: Event history models in political science. *American Journal of Political Science*, *41*(4), 1414-1461.

Brady, Henry E. (2008). Causation and Explanation in Social Science. En Janet M. Box-Steffensmeier, Henry E. Brady y David Collier (Eds.), *The Oxford Handbook of Political Methodology* (pp. 217-270). Oxford University Press.

Brady, Henry E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, *22*, 297-323.

Brady, Henry E. y Collier, David (Eds.). (2010). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield Publishers.

Brambor, Thomas, Clark, William Roberts y Golder, Matt. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, *14*(1), 63-82.

Broman, Karl W. y Woo, Kara H. (2018). Data organization in spreadsheets. *The American Statistician*, *72*(1), 2-10.

Bunge, Mario. (1999). *Buscar la filosofía en las ciencias sociales*. Siglo XXI.

Cameron, Charles M. (2000). *Veto Bargaining*. Cambridge University Press.

Centro de Investigación y Estudios Políticos. (2018). *Informe de Resultados del Estudio de Opinión Sociopolítica. 31 de enero de 2018*. Universidad de Costa Rica.

Centro de Investigación y Estudios Políticos. (2020). *Base de datos de la encuesta de noviembre 2020 [archivo Stata]*. Universidad de Costa Rica.

Chavarría-Mora, Elías y Angell, Katie. (2023). Shifting Positions: Party Positions and Political Manifestos in Costa Rica. *Latin American Politics and Society*, *65*(1), 1-21.

Chiozza, Giacomo. (2002). Is there a clash of civilizations? Evidence from patterns of international conflict involvement, 1946-97. *Journal of Peace Research*, *39*(6), 711-734.

Clark, Tom S. y Linzer, Drew A. (2015). Should I use fixed or random effects? *Political Science Research and Methods*, *3*(2), 399-408.

Clausen, Sten Erik. (1998). *Applied Correspondence Analysis*. SAGE.

Collier, David, Laporte, Jody y Seawright, Jason. (2008). Typologies: Forming Concepts and Creating Categorical Variables. En Janet M. Box-Steffensmeier, Henry E. Brady y David Collier (Eds.), *The Oxford Handbook of Political Methodology* (pp. 152-173). Oxford University Press.

Colomer, Josep. (2003). Son los partidos los que eligen los sistemas electorales (o las leyes de Duverger cabeza abajo). *Revista Española de Ciencia Política*, *9*(39–63).

Comisión Económica para América Latina y el Caribe. (2022). *Bases de Datos y Publicaciones Estadísticas. Tasa de desocupación, Costa Rica*. https://statistics.cepal.org/portal/cepalstat/index.html.

Cox, David R. (1982). Statistical significance tests. *British Journal of Clinical Pharmacology*, *14*(3), 325-331.

Creswell, John W. (2009). *Research Design*. SAGE.

Darmofal, David. (2015). *Spatial Analysis for the Social Sciences*. Cambridge University Press.

Davidian, Marie y Louis, Thomas A. (2012). Why statistics? *Science*, *336*(6077), 12-12.

De Mesquita, Bruce Bueno y Lalman, David. (1988). Empirical support for systemic and dyadic explanations of international conflict. *World Politics*, *41*(1), 1-20.

Dietrich, Bryce J. (2021). Using motion detection to measure social polarization in the US House of Representatives. *Political Analysis*, *29*(2), 250-259.

Dobson, Annette J. y Barnett, Adrian G. (2018). *An Introduction to Generalized Linear Models*. CRC Press.

Dolan, Kathleen. (2010). The impact of gender stereotyped evaluations on support for women candidates. *Political Behavior*, *32*, 69-88.

Duverger, Maurice. (1957). *Los partidos políticos*. Fondo de Cultura Económica.

Enders, Walter, Sandler, Todd y Gaibulloev, Khusrav. (2011). Domestic versus transnational terrorism: Data, decomposition, and dynamics. *Journal of Peace Research*, *48*(3), 319-337.

Ennser, Laurenz. (2012). The homogeneity of West European party families: The radical right in comparative perspective. *Party Politics*, *18*(2), 151-171.

Everitt, Brian y Hothorn, Torsten. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer.

Finch, W. Holmes. (2020). *Exploratory Factor Analysis*. SAGE.

Fisher, Ronald A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London*, *222*, 309-368.

Fournier Facio, Marco Vinicio. (2002). Una tipología de los electores. *Revista de Ciencias Sociales*, *98*, 9-18.

Franzese, Robert J. (2007). Multicausality, context-conditionality, and endogeneity. En Carles Boix y Susan C. Stokes (Eds.), *The Oxford Handbook of Comparative Politics* (pp. 152-173). Oxford University Press.

Frees, Edward W. (2004). *Longitudinal and Panel Data*. Cambridge University Press.

Fuhrmann, Matthew. (2009). Spreading Temptation: Proliferation and Peaceful Nuclear Cooperation Agreements. *International Security*, *34*(1), 7-41.

Geddes, Barbara. (2003). *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. University of Michigan Press.

Gelman, Andrew, Carlin, John B., Stern, Hal S. y Rubin, Donald B. (2004). *Bayesian Data Analysis*. CRC Press.

Gelman, Andrew y Hill, Jennifer. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, Andrew, Hill, Jennifer y Vehtari, Aki. (2020). *Regression and Other Stories*. Cambridge University Press.

Gelman, Andrew y Stern, Hal. (2006). The difference between significant and not significant is not itself statistically significant. *The American Statistician*, *60*(4), 328-331.

Gerber, Alan S. y Malhotra, Neil. (2008). Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research*, *37*(1), 3-30.

Gill, Jeff. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*(3), 647-674.

Glasgow, Garrett y Alvarez, R. Michael. (2008). Discrete Choice Methods. En Janet M. Box-Steffensmeier, Henry E. Brady y David Collier (Eds.), *The Oxford Handbook of Political Methodology* (pp. 513-529). Oxford University Press.

Goldthorpe, John H. (2001). Causation, statistics, and sociology. *European Sociological Review*, *17*(1), 1-20.

Gómez-Uribe, Carlos A. y Hunt, Neil. (2015). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, *6*(4), 1-19.

Green, Donald P. y Gerber, Alan S. (2008). *Get Out the Vote: How to Increase Voter Turnout*. Brookings Institution Press.

Gujarati, Dadomar N. y Porter, Dawn C. (2010). *Econometría*. McGraw-Hill.

Guzmán Castillo, Jesús. (2021). Participación ciudadana y democracia en Costa Rica 2018: entre activismo y apatía. En Ronald Alfaro Redondo (Ed.), *Participación y política electoral: nuevas miradas a las elecciones 2018 en Costa Rica* (pp. 71-86). Tribunal Supremo de Elecciones.

Head, Megan L., Holman, Luke, Lanfear, Rob, Kahn, Andrew T. y Jennions, Michael D. (2015). The Extent and Consequences of P-Hacking in Science. *PLOS Biology*, *13*(3), e1002106.

Hendershot, Marcus E., Hurwitz, Mark S., Lanier, Drew Noble y Pacelle Jr, Richard L. (2013). Dissensual decision making: Revisiting the demise of consensual norms within the US Supreme Court. *Political Research Quarterly*, *66*(2), 467-481.

Hernández Rodríguez, Óscar. (2004). Costa Rica. En John G. Geer (Ed.), *Public Opinion and Polling Around the World: A Historical Encyclopedia* (pp. 559-564). ABC-CLIO.

Hernández Rodríguez, Óscar. (2011). *Introducción a las Series Cronológicas*. Editorial UCR.

Hernández Rodríguez, Óscar. (2012). *Estadística elemental para Ciencias Sociales*. Editorial UCR.

Hernández Rodríguez, Óscar. (2013). *Temas de Análisis Estadístico Multivariante*. Editorial UCR.

Hernández Rodríguez, Óscar. (2015). *Elementos de probabilidades e inferencia estadística para Ciencias Sociales*. Editorial UCR.

Holland, Paul W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Horrace, William C. y Oaxaca, Ronald L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, *90*(3), 321-327.

Hosmer, David W., Lemeshow, Stanley y Sturdivant, Rodney X. (2013). *Applied Logistic Regression*. Wiley.

Huntington, Samuel. (1996). *The Clash of Civilizations and the Remaking of World Order*. Simon & Schuster.

Instituto Nacional de Estadística y Censos. (2012). *X Censo Nacional de Población y VI de Vivienda 2011: Características Sociales y Demográficas Tomo I*. Instituto Nacional de Estadísticas y Censos.

Instituto Nacional de Estadística y Censos. (2020). *Encuesta Nacional de Hogares Julio 2020. Resultados Generales*. Instituto Nacional de Estadísticas y Censos.

Jackman, Simon. (2004). Bayesian Analysis for Political Research. *Annual Review of Political Science*, *7*, 483-505.

Jenne, Erin K., Hawkins, Kirk A. y Silva, Bruno Castanho. (2021). Mapping populism and nationalism in leader rhetoric across North America and Europe. *Studies in Comparative International Development*, *56*(2), 170-196.

Kahn, Kim Fridkin y Geer, John G. (1994). Creating impressions: An experimental investigation of political advertising on television. *Political Behavior*, *16*(1), 93-116.

Kahneman, Daniel. (2012). *Pensar rápido, pensar despacio*. Random House Mondadori.

King, Gary. (1990). On Political Methodology. *Political Analysis*, *2*, 1-29.

King, Gary, Keohane, Robert O. y Verba, Sidney. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press.

Kish, Leslie. (1965). *Survey Sampling*. John Wiley & Sons.

Kleinbaum, David G. y Klein, Mitchel. (2012). *Survival Analysis: A Self-Learning Text*. Springer.

Krosnick, Jon A. (1999). Survey Research. *Annual Review of Psychology*, *50*(1), 537-567.

Kuhn, Thomas S. (1996). *The Structure of Scientific Revolutions*. University of Chicago Press.

Lazarsfeld, Paul y Fiske, Marjorie. (1938). The panel as a new tool for measuring opinion. *Public Opinion Quarterly*, *2*(4), 596-612.

Lektzian, David y Souva, Mark. (2009). A comparative theory test of democratic peace arguments, 1946-2000. *Journal of Peace Research*, *46*(1), 17-37.

Lewis-Beck, Colin y Lewis-Beck, Michael. (2016). *Applied Regression: An Introduction*. SAGE.

Lewis-Beck, Michael y Stegmaier, Mary. (2013). The VP-function revisited: a survey of the literature on vote and popularity functions after over 40 years. *Public Choice*, *157*, 367-385.

Lieberman, Evan S. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, *99*(3), 435-452.

Lijphart, Arend. (1971). Comparative Politics and the Comparative Method. *American Political Science Review*, *65*(3), 682-693.

Lijphart, Arend. (1999). *Patterns of Democracy*. Yale University Press.

Lijphart, Arend. (2012). *Data from the Appendix of Patterns of Democracy*. https://polisci.ucsd.edu/people/faculty/faculty-directory/emeriti-faculty/lijphart-profile.html.

Lindley, Dennis V. (2000). The Philosophy of Statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 293-337.

Lindstädt, René, Proksch, Sven-Oliver y Slapin, Jonathan B. (2020). When experts disagree: Response aggregation and its consequences in expert surveys. *Political Science Research and Methods*, *8*(3), 580-588.

Luke, Douglas A. (2004). *Multilevel Modeling*. SAGE.

Mahoney, James y Goertz, Gary. (2006). A tale of two cultures: Contrasting quantitative and qualitative research. *Political Analysis*, *14*(3), 227-249.

Mainwaring, Scott y Pérez-Liñán, Aníbal. (2013). *Democracies and Dictatorships in Latin America: Emergence, Survival, and Fall*. Cambridge University Press.

Martínez Franzoni, Juliana. (2008). Welfare regimes in Latin America: Capturing constellations of markets, families, and policies. *Latin American Politics and Society*, *50*(2), 67-100.

Massey, Douglas S. (1987). The ethnosurvey in theory and practice. *International Migration Review*, *21*(4), 1498-1522.

McCarty, Nolan y Razaghian, Rose. (1999). Advice and Consent: Senate Responses to Executive Branch Nominations 1885-1996. *American Journal of Political Science*, *43*(4), 1122-1143.

McDermott, Rose. (2002). Experimental methods in political science. *Annual Review of Political Science*, *5*(1), 31-61.

McLaughlin Mitchell, Sara y Hensel, Paul R. (2007). International institutions and compliance with agreements. *American Journal of Political Science*, *51*(4), 721-737.

Monroe, Kristin Renwick. (2007). The Perestroika Movement, its Methodological Concerns, and the Professional Implications of These Methodological Issues. *Qualitative Methods*, *5*(1), 2-6.

Mooney, Christopher Z. (1997). *Monte Carlo Simulation*. SAGE.

Moore, Barrington. (1966). *Social Origins of Dictatorship and Democracy*. Beacon Press.

Morton, Rebecca B. y Williams, Kenneth C. (2008). Experimentation in Political Science. En Janet M. Box-Steffensmeier, Henry E. Brady y David Collier (Eds.), *The Oxford Handbook of Political Methodology* (pp. 339-356). Oxford University Press.

Muñoz-Portillo, Juan. (2021). Effects of ballot type and district magnitude on local public goods bill-initiation behavior: Evidence from Honduras. *Political Research Quarterly*, *74*(2), 388-402.

Nelder, John Ashworth y Wedderburn, Robert. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370-384.

Noggle, Gary y Kaid, Lynda Lee. (2000). The effects of visual images in political ads: Experimental testing of distortions and visual literacy. *Social Science Quarterly*, *81*(4), 913-927.

Pearl, Judea y Mackenzie, Dana. (2018). *The Book of Why*. Basic Books.

Persson, Torsten y Tabellini, Guido. (2003). *The Economic Effects of Constitutions*. MIT Press.

Pignataro, Adrián. (2018). Análisis de datos de panel en ciencia política: ventajas y aplicaciones en estudios electorales. *Revista Española de Ciencia Política*, *46*, 259-283.

Pignataro, Adrián y Cascante, María José. (2018). *Los electorados de la democracia costarricense*. Tribunal Supremo de Elecciones.

Pignataro, Adrián y Cascante Segura, Carlos Humberto. (2017). Una sensibilidad focalizada: opinión pública y política exterior de costa rica hacia Nicaragua. *América Latina Hoy*, *77*, 93-114.

Pignataro, Adrián y Treminio, Ilka. (2019). Reto económico, valores y religión en las elecciones nacionales de Costa Rica 2018. *Revista de Ciencia Política*, *39*(2), 239-263.

Piovani, Juan Ignacio. (2007). Los orígenes de la estadística: de investigación socio-política empírica a conjunto de técnicas para el análisis de datos. *Revista de Ciencia Política y Relaciones Internacionales de la Universidad de Palermo*, *1*(1), 25-44.

Poole, Keith T. (2008). The Evolving Influence of Psychometrics in Political Science. En Janet M. Box-Steffensmeier, Henry E. Brady y David Collier (Eds.), *The Oxford Handbook of Political Methodology* (pp. 199-213). Oxford University Press.

Porta, Donatella della y Keating, Michael. (2008). How many approaches in the social sciences? An epistemological introduction. En Donatella della Porta y Keating Michael (Eds.), *Approaches and Methodologies in the Social Sciences: A Pluralist Perspective* (pp. 19-39). Cambridge University Press.

Przeworski, Adam, Álvarez, Michael E., Cheibub, José Antonio y Limongi, Fernando. (2000). *Democracy and Development*. Cambridge University Press.

Putnam, Robert D. (1993). *Making Democracy Work*. Princeton University Press.

Ragin, Charles C. (1987). *The Comparative Method*. University of California Press.

Rainey, Carlisle y McCaskey, Kelly. (2021). Estimating logit models with small samples. *Political Science Research and Methods*, *9*(3), 549-564.

Raventós Vorst, Ciska. (2008). Lo que fue ya no es y lo nuevo aún no toma forma: Elecciones 2006 en perspectiva histórica. *América Latina Hoy*, *49*.

Raventós Vorst, Ciska, Fournier Facio, Marco Vinicio, Fernández Montero, Diego y Alfaro Redondo, Ronald. (2012). *Respuestas ciudadanas ante el malestar con la política*. Tribunal Supremo de Elecciones.

Russett, Bruce M., Oneal, John R. y Cox, Michaelene. (2000). Clash of civilizations, or realism and liberalism déjà vu? Some evidence. *Journal of Peace Research*, *37*(5), 583-608.

Salsburg, David. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Henry Holt; Company.

Samuels, David J. y Shugart, Matthew S. (2010). *Presidents, Parties, and Prime Ministers*. Cambridge University Press.

Schmitter, Philippe C. (2008). The design of social and political research. En Donatella della Porta y Keating Michael (Eds.), *Approaches and Methodologies in the Social Sciences: A Pluralist Perspective* (pp. 263-295). Cambridge University Press.

Schmitter, Philippe C. (2009). The nature and future of comparative politics. *European Political Science Review*, *1*(1), 33-61.

Seligson, Mitchell A. (2002). Trouble in Paradise? The Erosion of System Support in Costa Rica, 1978-1999. *Latin American Research Review*, *37*(1), 160-185.

Shugart, Matthew S. y Taagepera, Rein. (2017). *Votes from Seats: Logical Models of Electoral Systems*. Cambridge University Press.

Silver, Nate. (2012). *The Signal and the Noise*. The Penguin Press.

Skrondal, Anders y Rabe-Hesketh, Sophia. (2004). *Generalized Latent Variable Modeling. Multilevel, Longitudinal, and Structural Equation Models*. CRC Press.

Steenbergen, Marco R. y Jones, Bradford S. (2002). Modeling multilevel data structures. *American Journal of Political Science*, *46*(1), 218-237.

Stegmueller, Daniel. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*(3), 748-761.

Stigler, Stephen M. (1978). Mathematical Statistics in the Early States. *The Annals of Statistics*, *6*(2), 239-265.

Stigler, Stephen M. (2010). Darwin, Galton and the statistical enlightenment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(3), 469-482.

Stimson, James A. (2015). *Tides of Consent: How Public Opinion Shapes American Politics*. Cambridge University Press.

Taylor-Robinson, Michelle M. y Geva, Nehemia (Eds.). (2023). *The Image of Gender and Political Leadership*. Oxford University Press.

Theocharis, Yannis y Van Deth, Jan W. (2018). The continuous expansion of citizen participation: A new taxonomy. *European Political Science Review*, *10*(1), 139-163.

Tribunal Supremo de Elecciones. (2018). *Cómputo de Votos Elecciones 2018 [archivo Excel]*. https://www.tse.go.cr/estadisticas_elecciones.htm.

Tribunal Supremo de Elecciones. (2022). *Elecciones presidenciales en cifras 1953-2022 [archivo Excel]*. https://www.tse.go.cr/estadisticas_elecciones.htm.

Vallverdú, Jordi. (2016). *Bayesians Versus Frequentists: A Philosophical Debate on Statistical Reasoning*. Springer.

Velleman, Paul F. y Wilkinson, Leland. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65-72.

Verba, Sidney y Nie, Norman H. (1972). *Participation in America: Political Democracy and Social Equality*. University of Chicago Press.

Wackerly, Dennis D., Mendenhall, William y Scheaffer, Richard L. (2002). *Estadística Matemática con Aplicaciones*. Thomson.

Wasserstein, Ronald L., Schirm, Allen L. y Lazar, Nicole A. (2019). Moving to a world beyond p< 0.05. *The American Statistician*, *73*(sup1), 1-19.

Wawro, Gregory. (2002). Estimating Dynamic Panel Data Models in Political Science. *Political Analysis*, *10*(1), 25-48.

Wooldridge, Jeffery M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wooldridge, Jeffrey M. (2023). What is a standard error?(And how should we compute it?). *Journal of Econometrics*, *237*(2), 105517.

Acerca del autor

Adrián Pignataro es doctor en Ciencia Política, Política Europea y Relaciones Internacionales por el programa conjunto de la Scuola Superiore Sant'Anna, la Università di Siena, la Università di Firenze y la Università di Pisa. Tiene una Maestría en Estadística y una Licenciatura en Ciencias Políticas, ambas de la Universidad de Costa Rica. Actualmente es profesor asociado en la Escuela de Ciencias Políticas de la Universidad de Costa Rica y forma parte del Centro de Investigación y Estudios Políticos (CIEP) de la misma institución. Investiga temas de comportamiento político, política comparada y opinión pública. Correo electrónico: adrian.pignataro@ucr.ac.cr

Corrección filológica: *Ariana Alpízar L.* • Revisión de pruebas: *Pamela Bolaños A.*
Diseño y diagramación de contenido y realización del libro digital: *El autor* • Diseño de portada: *Grettel Calderón A.*
Control de calidad de la versión digital: *Grettel Calderón A. y Hazel Aguilar B.*

Editorial UCR es miembro del Sistema Editorial Universitario Centroamericano (SEDUCA),
perteneciente al Consejo Superior Universitario Centroamericano (CSUCA).

Edición digital de la Editorial Universidad de Costa Rica. Fecha de creación: abril, 2024.

CIENCIAS SOCIALES | Ciencias Políticas

Este libro ofrece una introducción a los métodos y los modelos fundamentales de la estadística para aplicarse en la investigación empírica en ciencia política. A partir de un nivel de conocimiento previo básico en matemática y estadística, la meta es exponer de forma clara los conceptos fundamentales de la estadística clásica o frecuentista, con un número reducido de fórmulas, sin demostraciones formales y mediante el uso de ejemplos propios de la disciplina de la ciencia política. Esta segunda edición utiliza el programa gratuito de código abierto R como la herramienta de análisis.

EDITORIAL UCR