

MANUAL PARA EL ANÁLISIS POLÍTICO CUANTITATIVO

ADRIÁN PIGNATARO


EDITORIAL
UCR
2016



320.072

P632m Pignataro, Adrián
Manual para el análisis político cuantitativo / Adrián
Pignataro. –1. ed.– [San José], C. R.: Editorial UCR, 2016
1 recurso en línea (x, 171 p.) : il., digital, archivo PDF;
4.2 MB

Forma de acceso: World Wide Web

ISBN 978-9968-46-602-8

1. CIENCIA POLÍTICA – INVESTIGACIONES.
2. ESTADÍSTICA MATEMÁTICA. 3. ESTADÍSTICA
POLÍTICA. 4. SPSS (SISTEMA DE COMPUTACIÓN
ELECTRÓNICA). I. Título.

CIP/3038
CC/SIBDI.UCR

Edición aprobada por la Comisión Editorial de la Universidad de Costa Rica.

Primera edición: 2016.

La EUCR es miembro del Sistema de Editoriales Universitarias de Centroamérica (SEDUCA),
perteneciente al Consejo Superior Universitario Centroamericano (CSUCA).

Corrección filológica: *Marta Benavides G.* • Revisión de pruebas, diseño y diagramación: *El autor.*
Diseño de portada y control de calidad: *Wendy Aguilar G.*

© Editorial Universidad de Costa Rica, Ciudad Universitaria Rodrigo Facio, Costa Rica
Apto. 11501-2060 • Tel: 2511-5310 • Fax: 2511-5257 • administracion.siedin@ucr.ac.cr
www.editorial.ucr.ac.cr

Prohibida la reproducción total o parcial. Todos los derechos reservados.
Hecho el depósito de ley.

Edición digital de la Editorial Universidad de Costa Rica. Fecha de creación: julio, 2016.
Universidad de Costa Rica. Ciudad Universitaria Rodrigo Facio.

CONTENIDOS

Prefacio	vi
Capítulo 1. Metodología de la investigación cuantitativa	1
Introducción	1
El diseño metodológico cuantitativo	4
Alcances y limitaciones.....	6
Algunos conceptos básicos.....	10
Tipos de datos	10
Tipos de variables	13
Comentarios finales	15
Capítulo 2. Nociones generales de inferencia estadística	16
Introducción	16
Conceptos básicos	16
El cálculo de los errores	19
Intervalos de confianza	20
Caso de un promedio.....	20
Caso de un porcentaje.....	23
Pruebas de hipótesis	24
Comentarios finales	27
Ejercicios	28
Capítulo 3. Comparación de dos medias	30
Introducción	30
Procedimiento	31
Nota sobre igualdad de variancias	36
Comentarios finales	37
Ejercicios	37
Capítulo 4. Análisis de variancia de un factor.....	39
Introducción	39
Un ejemplo experimental.....	40
Procedimiento	44

Comentarios finales	48
Ejercicios	49
Capítulo 5. Medidas de asociación	50
Introducción.....	50
Prueba <i>chi</i> cuadrado para tablas de contingencia.....	51
Procedimiento para prueba <i>chi</i> cuadrado	56
Correlación bivariada: coeficiente de correlación de Pearson.....	61
Procedimiento para correlación de Pearson.....	64
Comentarios finales	66
Ejercicios	67
Capítulo 6. Regresión lineal simple por mínimos cuadrados ordinarios	68
Introducción.....	68
Conceptos.....	69
Etapas.....	70
Modelo	70
Ejemplo.....	72
Especificación del modelo.....	74
Estimación	74
Evaluación.....	78
Comentarios finales	81
Ejercicios	82
Capítulo 7. Regresión lineal múltiple por mínimos cuadrados ordinarios	84
Introducción.....	84
Modelo.....	85
Ejemplo.....	85
Especificación del modelo.....	87
Estimación	88
Evaluación.....	90
Problemas comunes en regresión múltiple.....	93
¿Cuántas variables independientes incluir?	93
Sesgo de variable omitida.....	94
Multicolinealidad	95
Endogeneidad.....	97
Comentarios finales	98

Ejercicios	99
Capítulo 8. Regresión logística	101
Introducción	101
Modelo.....	102
Ejemplo	104
Especificación	105
Estimación	105
Evaluación	113
Comentarios finales	115
Ejercicios	116
Capítulo 9. Análisis de conglomerados	118
Introducción	118
Ejemplo 1: clasificación de elecciones	120
Ejemplo 2: clasificación de parlamentos	127
Comentarios finales	132
Ejercicios	133
Capítulo 10. Análisis de factores	134
Introducción	134
Conceptos básicos	136
Ejemplo 1: satisfacción con los servicios públicos	137
Ejemplo 2: las democracias según Lijphart.....	147
Comentarios finales	151
Ejercicios	152
Apéndice A. Fuentes para profundizar.....	154
Apéndice B. Los modelos lineales generalizados	156
Respuestas a los ejercicios.....	158
Bibliografía	163
Acerca del autor.....	172

PREFACIO

El siguiente material busca introducir, a estudiantes de ciencia política, en los métodos, modelos y técnicas de la estadística para su uso en la investigación empírica. En ese sentido, no pretende reemplazar los intereses sustantivos con herramientas analíticas, sino complementarlos, para lo cual es fundamental partir de preguntas relevantes y motivadoras, teoría clara y coherente y de un conocimiento amplio de los contextos históricos y sociales de cada objeto de investigación.

En la amplia vertiente cuantitativa, concretamente, el texto presenta métodos estadísticos clásicos o frecuentistas, excluyendo otras metodologías cuantitativas como la modelación formal, la simulación computacional y la teoría de juegos, más bien orientadas hacia la generación de teoría mediante la formalización matemática.

Se promueve un enfoque predominantemente aplicado por medio del paquete SPSS, por lo que se dejan a un lado las demostraciones matemáticas, se minimiza el número de fórmulas y simbología y se recurre continuamente a ejemplos, la mayor parte basados en datos reales de encuestas o cifras electorales e institucionales de diversos países.¹ Debido a que se da un tratamiento superficial a la teoría estadística, los requisitos de conocimiento previo son escasos. Sin embargo, sí se supone un manejo básico de SPSS y de conceptos de estadística descriptiva.

El texto se organiza en tres grandes bloques. El primero introduce la investigación con orientación cuantitativa desde el punto de vista metodológico (capítulo 1) y de la teoría clásica de la inferencia estadística (capítulo 2); con base en estos conceptos, se estudian pruebas de hipótesis para dos medias (capítulo 3), para tres o más medias por medio del análisis de variancia (capítulo 4) y medidas de asociación para datos categóricos, por un lado, y continuos, por otro (capítulo 5).

¹ Los datos de los ejemplos y ejercicios se pueden descargar en la página de la Editorial Universidad de Costa Rica: <http://www.editorial.ucr.ac.cr/index.php/librosdigitales>.

Una segunda parte se dedica a los modelos de regresión, se examina el caso del modelo lineal gaussiano con un predictor (capítulo 6) o varios (capítulo 7). Para analizar variables categóricas binarias se recurre al modelo logístico (capítulo 8).

Finalmente, el tercer bloque corresponde al análisis multivariado. Entre las numerosas técnicas existentes, se presenta el análisis de conglomerados con el método aglomerante jerárquico (capítulo 9) y el análisis de factores exploratorio (capítulo 10).

Se cierra el libro con dos apéndices. El primero sugiere bibliografía útil y accesible para profundizar en los distintos temas que se introdujeron o para ahondar en otros que no se abarcaron del todo y el segundo hace una referencia a los modelos lineales generalizados como marco teórico que unifica la mayoría de los contenidos estudiados.

Se agradece a las distintas generaciones de estudiantes del curso CP-3414 Análisis e Interpretación de Datos Políticos por las observaciones y correcciones hechas a los borradores del presente libro. La publicación de este material es un resultado directo de haber impartido este curso junto con la motivación de numerosos colegas: Alberto Cortés, Felipe Alpízar, Fernando Ramírez, Fernando Zeledón, Gina Sibaja, Ilka Treminio, Luis Vives, María José Cascante, Sergio Moya, Shirley Rojas y Luz Marina Vanegas.

Se aprecian las recomendaciones de publicación extendidas por la Comisión Editorial de la Escuela de Ciencias y el Comité Científico del Centro de Investigación y Estudios Políticos (CIEP) y se agradecen las sugerencias de las personas evaluadoras del texto.

A las compañeras y compañeros de la Editorial de la Universidad de Costa Rica, en particular Aída Cascante, Cherryl Corrioso, María Benavides y Wendy Aguilar, se les destaca su apoyo y dedicación, pues gracias ellas se mejoró incuestionablemente el material.

CAPÍTULO I

METODOLOGÍA DE LA INVESTIGACIÓN CUANTITATIVA

Introducción

Desde una perspectiva histórica, es posible distinguir una relación estrecha entre estadística como disciplina y el estudio de fenómenos políticos y sociales. En la actualidad, se puede entender la estadística –entre las muchas formas posibles– como la ciencia para aprender de los datos y medir, controlar y comunicar la incertidumbre (Davidian y Louis, 2012). Pero llegar a esta definición exigió de bastantes años de evolución.

El vínculo entre ambas áreas del conocimiento se remonta al siglo XVII, cuando se desarrolló en Inglaterra la denominada “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, que gradualmente incorporó datos cuantitativos y dotó de nombre a la disciplina. Estas tradiciones investigativas –concentradas en la recolección, pero carentes de métodos analíticos– convergen con la teoría de la probabilidad francesa (más abstracta y matemática) para dar origen a la primitiva ciencia estadística (Piovani, 2007).

A finales del siglo XIX y principios del XX, la teoría estadística, sus métodos y técnicas de análisis avanzaron gracias a aplicaciones en diversos campos como genética, agronomía, sociología, demografía y a los trabajos de figuras como Francis Galton, Karl Pearson, Ronald A. Fisher y muchos otros (cfr. Salsburg, 2001). El impacto de la metodología estadística es tal que diversas ciencias empiezan a adquirir un razonamiento probabilístico en lugar del determinista previo.

La ciencia política no fue la excepción y, por ende, evidencia estas dos lógicas de la causalidad. Por ejemplo, la famosa frase de Barrington Moore sentenció de forma determinística que “no burguesía, no democracia” (1966, p. 418, traducción propia) refiriéndose a la transición de una sociedad hacia un régimen democrático y entendiendo la presencia de la burguesía como una condición necesaria cuya presencia conducía inevitablemente al establecimiento de la democracia. En la misma línea de investigación –pero con una argumentación probabilística– el estudio comparativo sobre desarrollo y democracia de Przeworski *et al.* (2000) sostiene que:

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 monótonicamente conforme el ingreso per cápita es mayor (p. 273, traducción propia).

Aunque es evidente una diferencia temporal de ambos textos, esto no significa que existió una “conversión” total de lo determinístico a lo probabilístico. Por el contrario, ambas perspectivas dependen más de las premisas científicas y de los enfoques metodológicos, de modo que ambos conviven en la actualidad (Beach y Pedersen, 2013).

Lo que es cierto es que el aporte de la estadística a la ciencia política ha sido notable, tanto en la introducción de ese razonamiento probabilístico como en cuanto a la aplicación de técnicas y metodologías para el análisis político, el cual se desencadenó dramáticamente con la revolución conductual de la ciencia política durante la década de 1960 (King, 1990).

Así, por ejemplo, el estudio de la cultura política surgió, en buena medida, gracias a la “tecnología de la investigación mediante encuestas”, basadas en el desarrollo de las técnicas de muestreo, de entrevistas con mayor confiabilidad, de escalas y

técnicas de medición y de métodos de análisis y de inferencia estadísticos (Almond, 1999, p. 201).²

Pero el aporte de los métodos cuantitativos no se ha limitado a métodos de recolección y análisis de encuestas. El análisis de datos agregados en la política comparada emergió prácticamente al mismo tiempo que el análisis conductual (Schmitter, 2009) y buscaba relacionar variables medidas para unidades políticas (comúnmente Estados, pero también unidades subnacionales como regiones, cantones, etc.) como tipo de régimen, crecimiento económico, desigualdad, desarrollo humano, indicadores demográficos (estructura de la población, tasas de mortalidad, esperanza de vida), componentes institucionales (como el sistema electoral), configuración de actores (sistemas de partidos, estructuras de agregación de intereses) y otros.

En menor medida, se han aplicado diseños experimentales debido a la dificultad para asignar tratamientos aleatoriamente y manipular variables de interés y por los altos costos operativos e imposibilidades técnicas. A pesar de ello, existen interesantes aplicaciones como la comparación de métodos para movilizar a los electores para que asistan a las urnas (Green y Gerber, 2003) y, en general, la experimentación en ciencia política ha registrado un crecimiento en los últimos años (Morton y Williams, 2008).

De manera tal que, en la actualidad, los métodos estadísticos para datos observacionales y experimentales forman parte del currículo formativo global en ciencia política. Sin embargo, esto no quiere decir que apliquen inexorablemente en todos los problemas de estudio. La investigación cuantitativa corresponde a un diseño metodológico específico, con sus alcances y desventajas, como se verá a continuación.

² 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. Los sondeos políticos fueron iniciados, a su vez, por George Gallup y Louis Bean poco tiempo después (Salsburg, 2001, pp. 172-175; ver también Crespi, 2000, capítulo 7). En Costa Rica, los sondeos iniciaron a mediados de la década de 1960 como actividades privadas de partidos políticos, pero el estudio sistemático y periódico de la opinión pública empieza en 1974, desde la Oficina de Información del Ministerio de la Presidencia. Posteriormente, se establecen las primeras empresas privadas de encuestas: CID-Gallup en 1977 y UNIMER en 1986 (Hernández, 2004, pp. 559-564).

El diseño metodológico cuantitativo

Una vez definida una pregunta de investigación —es decir, un fenómeno político que interese describir, explicar o predecir— es posible argumentar que la mejor forma de responderla sea mediante los métodos cuantitativos.

Un diseño cuantitativo de investigación se destina a probar teorías objetivas examinando relaciones entre variables, usualmente por medio de procedimientos estadísticos (Creswell, 2009, p. 4). Esta metodología, tradicionalmente, asume una perspectiva filosófica o epistemológica denominada pospositivista (pues constituye una revisión del positivismo clásico). El pospositivismo se caracteriza por sostener la existencia de relaciones causales, presentar simplicidad en las ideas que se quieren probar (también denominada parsimonia³), desarrollar el conocimiento basado en la observación y la medición de la realidad objetiva y buscar leyes o teorías que gobiernen el mundo (Creswell, 2009, p. 7; ver además della Porta y Keating, 2008).⁴

Al asumir el diseño metodológico cuantitativo, es necesario contar con un cuerpo teórico robusto o de conocimiento previo (hallazgos de investigaciones anteriores) que provean de hipótesis que son respuestas conjeturales a las preguntas. Las hipótesis, comúnmente, relacionan variables o características de las observaciones o casos de estudio; para analizar empíricamente dichas relaciones entre variables se debe disponer de datos, es decir, operacionalizar los conceptos en indicadores e índices.

Los indicadores son aproximaciones a las categorías teóricas que interesan en el fondo, pero que no se pueden medir directamente; piénsese en conceptos como clase social, apoyo al sistema político, poder presidencial, influencia internacional, los cuales no conllevan una cuantificación directa. Sin embargo, es posible proponer formas empíricas indirectas para medir estos conceptos con datos

³ Sin embargo, no todos recomiendan buscar la parsimonia como un “bien esencial”, para otros más bien las teorías deben ser tan complicadas como la evidencia lo sugiera (King, Keohane y Verba, 1994, p. 20).

⁴ Esto no significa que todos los pospositivistas incorporen métodos cuantitativos; los análisis cualitativos de condiciones necesarias y suficientes, por ejemplo, tienden a ser pospositivistas en tanto es clara la premisa de causalidad, la búsqueda de leyes y la referencia a un conocimiento objetivo.

disponibles; por ejemplo, Przeworski *et al.* (2000) utilizan el ingreso per cápita como indicador del desarrollo económico, reconociendo sus limitaciones en cuanto no dice todo lo que implica el desarrollo, pero su disponibilidad es una ventaja y se prefiere frente a medidas alternativas como consumo energético, alfabetismo e industrialización.

Otro ejemplo lo constituye la medición de la participación electoral, pues aunque no se sabe cuál es el número de personas que asistieron a las urnas (usualmente no se realiza este conteo en cada mesa electoral), se tiene el número total de votos depositados; con ello, se puede calcular el porcentaje de participación dividiendo el total de votos entre el total de personas empadronadas. No obstante, incluso un indicador que parece sencillo como este dista mucho de ser perfecto, ya que sobreestimaría la participación en países que tengan muchas personas sin empadronar por diversas razones. Por ello, es más común (Geys, 2006) que se calcule la participación como el total de votos entre la población en edad de votar, que igualmente conlleva desventajas (incluye extranjeros sin derecho al voto, por ejemplo). Por lo tanto, en realidad no hay indicadores “perfectos”.⁵

Con base en estos indicadores escogidos, para los cuales existen datos, es posible probar las relaciones entre ellos, es decir las hipótesis. Este diseño investigativo esbozado sigue lo que se conoce como modelo deductivo-hipotético (Bunge, 1999), en tanto se parte de teorías, de las cuales se derivan hipótesis y sus respectivas variables, medidas a través de indicadores (ver figura 1.1).

⁵ La medición de conceptos es un tema más amplio de lo expuesto y debe considerar necesariamente los criterios de validez (que el concepto mida lo que propone medir) y la confiabilidad (que provean el mismo resultado si las mediciones se repiten muchas veces). Ver Shively (2011, pp. 45-55). Además, algunos conceptos simplemente no se pueden capturar con uno o varios indicadores, por lo tanto se tratan como “constructos” o variables latentes.

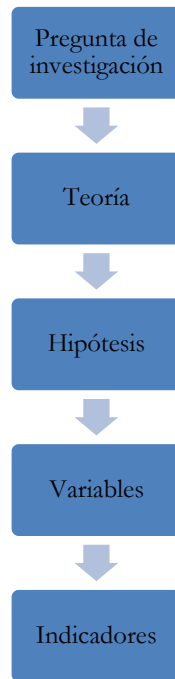


Figura 1.1. Esquema de la investigación orientada cuantitativamente.

Fuente: elaboración propia.

Alcances y limitaciones

Este estilo de investigación y la aplicación de métodos estadísticos implican ciertas ventajas y desventajas de las cuales es preferible ser consciente. Algunos de sus alcances son los siguientes:

- *Permite estudiar una cantidad grande de observaciones o casos: países, electores, legisladores, leyes, conflictos internacionales, etc.* Por ejemplo, los estudios de opinión por encuestas recopilan información de cientos o incluso miles de personas en contraste con entrevistas cualitativas e historias de vida que se centran alrededor de unos pocos informantes de relevancia. En política comparada, mientras que los estudios cualitativos se concentran, por lo general, en unos dos o tres países, los estudios cuantitativos incorporan muchos más casos (20 regiones en Putnam, 1993; 36 países

en Lijphart, 1999; 93 partidos políticos en Altman *et al.*, 2009; 4730 años-régimen en Przeworski *et al.*, 2000).

- *Analiza muchas variables.* Los métodos clásicos de comparación inferen utilizando casos que sean diferentes en todo excepto en una circunstancia común o semejantes en todo excepto en una variable (ver Ragin, 1987), por lo que se escogen según unas pocas variables que se puedan controlar. Con los modelos estadísticos, por el contrario, los análisis multivariados permiten controlar mayor número de variables, o bien examinar fenómenos multidimensionales o multicausales.
- *Generaliza.* La inferencia estadística permite concluir para una población extensa con base en una muestra pequeña en relación con el universo original, pero solamente si la muestra se realiza por medio de selección aleatoria (punto que se abarcará con mayor amplitud en el capítulo 2).
- *Prueba teorías explicativas o predictivas.* Aunque los estudios cuantitativos pueden ser descriptivos y de asociación, por lo general el fin último de una investigación en el marco positivista consiste en examinar hipótesis que relacionan variables independientes con una variable dependiente. Los modelos de regresión (capítulos 6, 7 y 8) buscan contrastar este tipo de relaciones.
- *Es capaz de calcular el error en la explicación o predicción.* Aunque todo trabajo científico es susceptible de equivocarse en sus conclusiones, los métodos estadísticos estiman errores en la inferencia, generalización y explicación de fenómenos.

Por otra parte, estas son ciertas limitaciones frecuentes en estudios cuantitativos:

- *Al estudiar relaciones causales, lo que se puede determinar son efectos de causas* (ver Mahoney y Goertz, 2006). Con esto se quiere decir que se analiza el resultado de determinados factores sobre un fenómeno, pero no se puede establecer de antemano cuáles variables constituyen las causas necesarias o las explicaciones reales; son las teorías previamente desarrolladas —y no los métodos— las que indican cuáles variables se deben considerar.
- *Los métodos estadísticos no están orientados hacia la generación de teoría.* Siguiendo el punto anterior, las hipótesis y teorías que se prueban deben

desarrollarse en trabajos previos o construirse por otros métodos como estudios de caso, teoría fundamentada, teoría de juegos u otros.

- *Se generaliza, pero no se especifica ni se detalla en casos particulares.* Lo usual es trabajar con efectos promedio, pues explicar puntualmente por qué ocurre un fenómeno en determinado contexto (¿por qué se dio la transición en España?, ¿cuál secuencia histórica llevó a desarrollarse la revolución en Rusia?) requiere de herramientas cualitativas como el rastreo del proceso (*process tracing*) que revele los mecanismos causales detrás del fenómeno (ver Beach y Pedersen, 2013).
- *Se exige una codificación numérica de los datos.* Los conceptos y variables deben ser trasladados a un lenguaje numérico para poder tratarlos y analizarlos estadísticamente.
- *Comúnmente no se admite mayor número de variables respecto al número de observaciones* (esto se profundizará más adelante).
- *Se encuentran limitados por el desarrollo teórico y computacional.* Es decir, los modelos, métodos y técnicas que se aplican comúnmente son aquellos que la teoría estadística ya desarrolló y que los paquetes computacionales permiten implementar. De hecho, muchos cálculos serían prácticamente imposibles de no existir las computadoras actuales.

Tanto por las limitaciones metodológicas como por la tendencia hegemónica en la comunidad científica (algunos incluso equiparan los métodos estadísticos con metodología política; por ejemplo, King, 1990), la investigación con orientación cuantitativa ha recibido fuertes críticas, sobre todo en la academia estadounidense donde el predominio de lo cuantitativo es claro (para una perspectiva del debate, ver Brady y Collier, 2010). Los comentarios de autores como Sartori (2004) y del movimiento Perestroika⁶ van en dicha dirección.

⁶ El movimiento Perestroika nace a partir de un correo electrónico (firmado bajo el seudónimo Mr. Perestroika), enviado en el año 2000, que se disemina masivamente, en el cual se atacaba el énfasis cuantitativo, conductista y de elección racional imperante en la ciencia política estadounidense. Se decía que la estadística alcanzaba niveles técnicos elevados que oscurecían la importancia sustantiva de los fenómenos e ignoraban aspectos básicos referidos a la definición de los conceptos y la calidad de los datos (ver Monroe, 2007).

Un aspecto crucial es que los métodos estadísticos, por más potentes que sean, no sustituyen una buena teoría. En otras palabras, no se debe confundir correlación con causalidad⁷ y gran cantidad de situaciones muestran las absurdas conclusiones a las que se llegaría si se ignorara esta advertencia. Existe, por ejemplo, una relación positiva entre el número de cigüeñas y el número de nacimientos en Europa (Gómez, 1998); el aumento de las ventas de helado se correlaciona con el incremento de los incendios forestales; incluso se ha llegado a predecir la dirección del mercado de acciones según la liga de procedencia del equipo ganador del Supertazón (Silver, 2012).⁸

En efecto, se pueden encontrar relaciones estadísticamente significativas, como algunos observaron para los anteriores ejemplos, pero la ausencia de teoría y de lógica obliga a repensar la relación. Como elegantemente expresaron Stepan y Skach (1994, p. 128), una proposición probabilística en la política es más que una aseveración estadística: conlleva la identificación y explicación de un proceso político específico que tiende a producir resultados probabilísticos. Es decir, debe existir un mecanismo causal (un “por qué ocurre”) razonable y teóricamente fundamentado que explique los hallazgos.

Una forma de sobrellevar las limitaciones de la estadística en la ciencia política (pero también aprovechando sus alcances) consiste en combinar métodos cualitativos y cuantitativos. Algunos denominan a dicha estrategia como “métodos mixtos”.

El diseño mixto puede adoptar muchas formas, pero en general implica la construcción de una lógica investigativa que combine lo cuantitativo y lo cualitativo, aprovechando las fortalezas de cada uno para obtener resultados más profundos de lo que se ganaría con cada método por aparte (Creswell, 2009, p. 203). Para ejemplificar, una posible investigación mixta puede iniciar con un análisis cualitativo que permite construir un instrumento (como un cuestionario),

⁷ El concepto de causalidad implica una amplia discusión filosófica y metodológica que se abarcará en este libro. Para ello puede consultarse Brady (2008).

⁸ En el sitio web <http://www.tylervigen.com/spurious-correlations> se expone un mayor número de ejemplos de correlaciones espurias.

el cual se aplica luego en una muestra representativa, siguiendo los procedimientos cuantitativos.

También es posible partir de un estudio cuantitativo comparado que incluya muchas observaciones. Con los resultados estadísticos, se pueden realizar posteriores análisis cualitativos con un menor número de casos escogidos según criterios deliberados, como seleccionar el caso más alejado de la línea de ajuste en regresión (es decir, el excepcional o desviado que contradice la tendencia promedio). Esta estrategia, denominada análisis anidado (*nested analysis*), puede ser especialmente enriquecedora en política comparada, donde los casos son usualmente países o unidades políticas, pero también en el análisis de actitudes o comportamientos individuales (Lieberman, 2005).

Otro ejemplo de estrategia mixta es la denominada etnoencuesta (*ethnosurvey*), con la que se combina —de manera sistemática y ordenada— la aplicación de cuestionarios semiestructurados (susceptibles a análisis estadístico) con técnicas etnográficas como observación e historias de vida, donde una fructífera aplicación ha sido el estudio de procesos migratorios (Massey, 1987).

Algunos conceptos básicos

Antes de proceder con los métodos, modelos y técnicas particulares, hay que tener claros algunos conceptos referidos a los tipos de datos y de variables, es decir, la materia prima con la cual se aplicarán las diversas herramientas. Conocer los datos es fundamental pues, como indica Charles Wheelan (2013, p. 111), al igual que una receta de cocina requiere de buenos ingredientes, en estadística no importa cuán sofisticado sea el análisis, no se puede compensar el hecho de tener datos de mala calidad.

Tipos de datos. Se pueden distinguir tres tipos de datos, según el diseño o la forma en que se recopilan. El primero corresponde a los datos de series de tiempo o cronológicos, aquellos para una misma unidad de análisis que se recogen en distintos puntos en el tiempo. Así, por ejemplo, las cifras de participación electoral (medida como porcentaje de votos totales entre personas empadronadas) en Costa Rica constituirían una serie de tiempo (figura 1.2).

El segundo tipo corresponde a los datos transversales, obtenidos para distintas unidades de análisis en un único punto en el tiempo. Una encuesta en la que las personas sean entrevistadas una única vez durante un periodo determinado corresponde a un diseño transversal, al igual que una comparación de los porcentajes de participación electoral en los países centroamericanos en determinadas elecciones (una por país) (figura 1.3). Por último, los datos de panel o longitudinales combinan datos transversales con series de tiempo. Pueden ser encuestas en las que existen distintas rondas de entrevistas y se repiten a las personas encuestadas. Asimismo, puede tenerse un diseño de este tipo para un análisis sobre participación electoral en distintos países centroamericanos a lo largo de varias elecciones para cada país (figura 1.4).

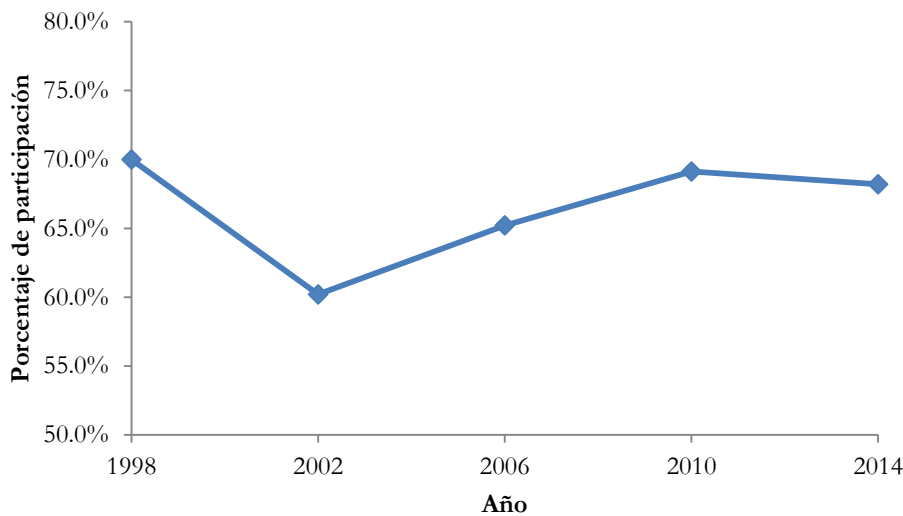


Figura 1.2. Participación electoral en Costa Rica (elecciones presidenciales de primera ronda).

Fuente: IDEA (2014).

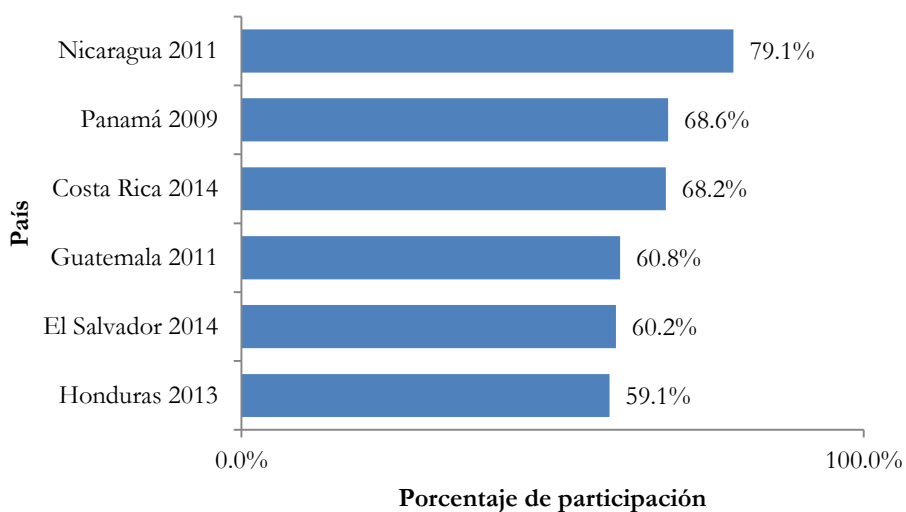


Figura 1.3. Participación electoral en Centroamérica (elecciones presidenciales de primera ronda).

Fuente: IDEA (2014).

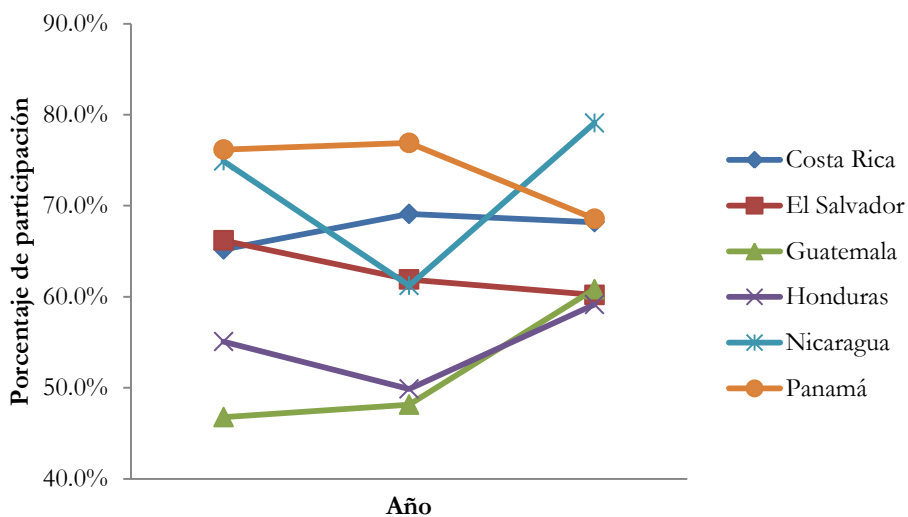


Figura 1.4. Participación electoral en Centroamérica para las últimas tres elecciones (elecciones presidenciales de primera ronda).

Fuente: IDEA (2014).

En este texto se abordarán métodos estadísticos únicamente para datos transversales. El análisis de series de tiempo y de datos de panel implica métodos específicos –y algo más complejos– para tomar en cuenta aspectos como la heterogeneidad entre observaciones (cada individuo o unidad política tiene características propias que se deben controlar) y las relaciones dinámicas como la dependencia temporal (los valores en un punto en el tiempo están asociados con los valores temporalmente previos) (Frees, 2004).

Por otra parte, también se pueden clasificar los datos según su naturaleza, sea experimental u observacional. El primer tipo está caracterizado por manipulación o asignación –aleatoria– de variables o tratamientos, de modo que los datos se generan en la investigación. En el segundo, la información existe ya en el mundo social, la historia o la naturaleza, no es creada por los investigadores.

Téngase presente que es posible combinar los distintos tipos de datos según el diseño o estructura y su naturaleza, de manera que pueden existir datos experimentales, tanto transversales como de panel (las personas participantes del experimento son sometidas a tratamientos de manera repetida) y los observacionales también pueden recopilarse de las tres formas descritas.

Tipos de variables. Las variables pueden ser categóricas o cualitativas cuando se refieren a atributos, mientras que las métricas o cuantitativas se expresan naturalmente en cantidades. Las variables categóricas se miden en un nivel nominal cuando no existe un orden en sus categorías (por ejemplo, mujer y hombre para la variable sexo), mientras que el nivel ordinal aplica cuando sí se pueden ordenar (como en el caso de estratos sociales alto, medio y bajo o los niveles educativos primario, secundario y universitario).

Los niveles de medición de intervalo y de razón corresponden a variables métricas o cuantitativas. La diferencia entre ambas está en el significado del cero: para las escalas de intervalo, el cero es un valor arbitrario, mientras que para las de razón significa la ausencia de la cantidad. Generalmente, los textos (como Hernández, 2012) ejemplifican una de intervalo con la temperatura en centígrados donde el cero no es la ausencia de temperatura y la de razón con una longitud como la estatura o el número de hijos. En ciencia política, podría ilustrarse una medición de intervalo por el índice *Polity IV* que varía entre –10

(autocracia) y 10 (democracia) y el cero no se interpreta como ausencia de un régimen político, sino como un régimen intermedio (ver Marshall y Cole, 2011).

Las de razón son más comunes y se observan en número de partidos políticos por legislatura, porcentajes de votos hacia un candidato o niveles de participación electoral (tiene sentido hablar de cero partidos, cero votos o de 0% de participación cuando nadie asiste a las urnas). Lo anterior se resume en la figura 1.2.

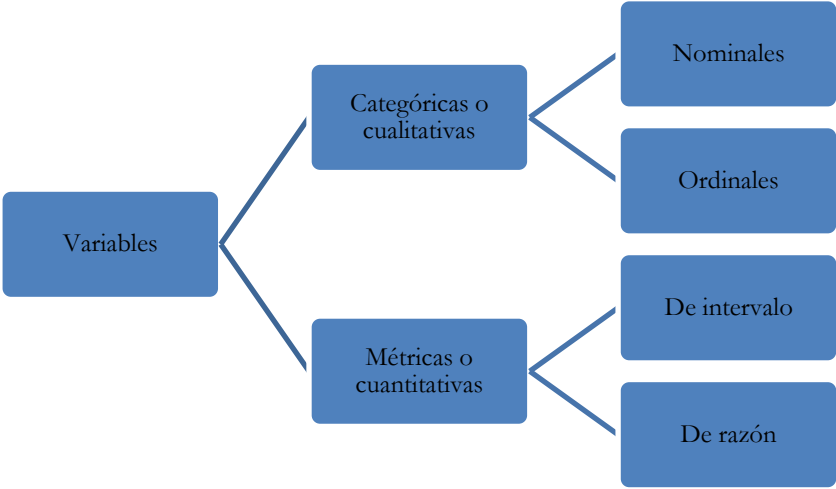


Figura 1.5. Tipos de variables y niveles de medición.

Fuente: elaboración propia.

Aunque algunos apelan hacia una utilización práctica y flexible de los niveles de medición de las variables (Velleman y Wilkinson, 1993), la distinción entre métricas y categóricas es importante puesto que hay restricciones entre lo que se puede hacer o no según el tipo de variable.

En el nivel elemental, nadie podría obtener un promedio simple de una variable categórica como la provincia de procedencia, pero sí sus porcentajes en la distribución; lo mismo sucede con los métodos estadísticos más sofisticados y en el cuadro 1.1 se exponen las técnicas que se verán en esta obra con los correspondientes tipos de variables para las que aplican.

Cuadro 1.1. Técnicas estadísticas según los tipos de variables

Técnicas	Variables para las que aplican
Prueba <i>t</i> de comparación de medias	Una categórica (2 categorías) y una métrica
Análisis de variancia de un factor	Una categórica (3 o más categorías) y una métrica
Prueba <i>chi</i> cuadrado	Dos categóricas
Correlación de Pearson	Dos métricas
Regresión por mínimos cuadrados ordinarios	Variable dependiente métrica
Regresión logística	Variable dependiente categórica
Análisis de conglomerados	Métricas
Análisis de factores	Métricas

Fuente: elaboración propia.

Comentarios finales

La investigación cuantitativa cuenta con una larga y consolidada tradición en la ciencia política que ha permitido abarcar múltiples preguntas de investigación en diversas áreas (comportamiento político, política comparada, relaciones internacionales, etc.). Sin embargo, el seguimiento cabal de la metodología cuantitativa implica el conocimiento de sus supuestos epistemológicos, de sus alcances y de sus limitaciones. Además, otros aspectos metodológicos como la formulación de las teorías e hipótesis, la operacionalización de variables y la escogencia de indicadores apropiados se vinculan indisolublemente con una investigación cuantitativa de calidad.

Antes de recurrir directamente a la estadística, hay que pensar cuál es el mejor método para contestar la pregunta de investigación. Si interesa generalizar, examinar un gran número de casos, controlar por muchas variables, determinar relaciones entre variables, estimar efectos promedio y obtener un error medible, entonces los modelos, métodos y técnicas estadísticos resultan apropiados. Asimismo, debe seleccionarse la herramienta estadística apropiada entre la multiplicidad disponible, prestando atención a la naturaleza de los datos, los tipos de variables y los niveles de medición.

CAPÍTULO 2

NOCIONES GENERALES DE INFERENCIA ESTADÍSTICA

Introducción

La inferencia estadística busca obtener conclusiones sobre cantidades desconocidas como pueden ser las características de una población o las relaciones hipotéticas entre estas características (Gelman *et al.*, 2004).

En el caso de la inferencia para una población, se generalizan los resultados obtenidos en una muestra aleatoria hacia la población de la cual se extrajo. Para esta operación, se puede medir la precisión de la generalización realizada. Seguidamente, se verán algunos conceptos fundamentales de la inferencia estadística clásica (también llamada frecuentista) enfocados en el ámbito de la inferencia para la población.

Conceptos básicos

Un principio general de la investigación cuantitativa es que, para resolver una pregunta, debe incluirse en el análisis todo el universo o población y no una selección intencional de casos; por ejemplo, si interesa analizar el desempeño económico de las democracias en desarrollo, deben considerarse todos los países disponibles, no únicamente aquellos de mayor relevancia internacional o aquellos más conocidos (Geddes, 2003).

La población depende de cómo se defina en una investigación. Puede ser la población de un país, el total de electores en un padrón, los diputados de un Congreso, el conjunto de partidos políticos de una contienda electoral...

Sin embargo, si esta población es muy amplia para ser medida en su totalidad – por ejemplo, una población de millones de habitantes en un país o un registro de miles de votos de los diputados de un parlamento– resulta preferible extraer una muestra aleatoria de la población; por el contrario, si la población es pequeña, como los 57 diputados de la Asamblea Legislativa costarricense, conviene medir a toda la población en lugar de obtener una muestra. El hecho de que la selección de la muestra sea al azar permitirá realizar inferencias para cada población particular.⁹

Por ejemplo, una muestra aleatoria de la población de docentes de una universidad permite inferencias para la población de docentes de la universidad, no sobre la población total de la universidad que incluiría estudiantes, administrativos, personal de mantenimiento, etc. Asimismo, una encuesta telefónica permite inferir solamente a la población que tiene teléfono y no a todos los habitantes de un territorio. Por ello, hay que tener muy claro cuál es la población hacia la cual se quiere inferir con base en una muestra.

Antes que nada, en el tema de la inferencia estadística, hay dos conceptos importantes por aprender:

- *Parámetro*: es un valor desconocido que se quiere estimar mediante la inferencia estadística. Puede corresponder a una característica de la población o a un proceso o relación hipotética (como un coeficiente de regresión, ver capítulos 6, 7 y 8). Usualmente se denotan con una letra griega (como μ , σ , β).
- *Estimador*: corresponde a un cálculo o fórmula que se utiliza para aproximar o estimar el valor desconocido del parámetro con base en un conjunto de valores extraídos de la población y que conforman una muestra aleatoria (por ejemplo: la media, la desviación estándar, la mediana, el máximo, el mínimo, etc.).

Para ejemplificar, supóngase que la población de interés está constituida por los habitantes del cantón de Montes de Oca en el año 2015. Con base en un

⁹ En todo el libro se asume que el procedimiento de selección de la muestra es irrestricto al azar, pues es el más simple de todos. Para una explicación de otras técnicas de muestreo puede consultarse Hernández (2012, pp. 9-23).

procedimiento aleatorio se extrae una muestra de esta población. Es de interés conocer el promedio de edad denotado μ en esta población de Montes de Oca; este valor corresponde al parámetro y es desconocido. Basado en la muestra, se puede estimar un valor del promedio de edad e inferirlo para la población.

No obstante, para el cálculo del parámetro hay muchas fórmulas posibles, muchos estimadores. Entonces, ¿cómo saber que el cálculo de un promedio muestral permite inferir con seguridad el valor del promedio poblacional?

Esto se conoce por la teoría estadística matemática, la cual determina cuáles son los mejores estimadores para los valores poblacionales. Actualmente, el procedimiento que se utiliza para ello se denomina “máxima verosimilitud” y fue propuesto por uno de los padres de la estadística: Sir Ronald A. Fisher (1890-1962).¹⁰

Usualmente, se espera que los estimadores posean dos propiedades fundamentales: (1) que sean insesgados; (2) que sean eficientes.¹¹

Ser insesgado significa que el estimador en promedio es igual al parámetro. A este “en promedio” se le llama el valor esperado o esperanza matemática. Formalmente se expresa así: $E(\hat{\theta}) = \theta$.

Es decir, si θ es un parámetro y $\hat{\theta}$ su estimador, entonces el valor esperado de $\hat{\theta}$ o el promedio de $\hat{\theta}$ es igual al parámetro θ cuando es insesgado. Si el promedio de estimadores y el parámetro son diferentes, entonces se dice que es sesgado. Por ejemplo, el promedio simple, denotado como \bar{X} se calcula de la siguiente forma:

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

Este \bar{X} es un estimador insesgado de la media desconocida o parámetro poblacional μ , ya que si se obtuvieran muchas muestras, y a cada una se le calcula un promedio, la media de estos promedios (es decir, el valor esperado de \bar{X}) sería igual al parámetro: $E(\bar{X}) = \mu$.

¹⁰ Para una reseña de la biografía de Fisher y sus aportes en la estadística y en otras ciencias como la genética, consúltase Yates y Mather (1963), así como el libro de Salsburg (2001).

¹¹ Existen otras como la consistencia y la suficiencia.

Por otro lado, es demostrable que un estimador de la desviación estándar calculado como

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

es sesgado respecto al parámetro σ (la desviación estándar en la población). En términos formales, $E(s') \neq \sigma$.

Sin embargo, con un poco de matemática se encuentra que si la desviación estándar de la muestra se calcula con un $n - 1$ en el denominador, este estimador resulta insesgado (para la demostración, ver Wackerly, Mendenhall y Scheaffer, 2002, pp. 372-373); por ello se habla de desviación estándar muestral –distinta de la poblacional– y se escribe de la siguiente manera:

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

De modo que el valor esperado de dicho estimador es igual al parámetro. En otras palabras, $E(s) = \sigma$.

La otra característica principal de un “buen” estimador es su eficiencia. Un estimador eficiente es aquel que permite calcular el valor poblacional o del parámetro con poco error, es decir, su variabilidad es pequeña. Cuanto mayor sea su eficiencia, menor el error y más precisa es la estimación o inferencia. Esto conlleva a hablar del error en la estadística.

El cálculo de los errores

Cuando en estadística se habla de error, no siempre significa equivocaciones al realizar cálculos u omisiones cometidas en el trabajo de campo (como escribir mal una respuesta al completar un cuestionario). El error se refiere a la imprecisión en la estimación de los parámetros.

Esta imprecisión de la estimación está directamente relacionada con cuán variable es el fenómeno. Véase el siguiente ejemplo: un laboratorista extrae sangre de una

paciente para realizar ciertas pruebas clínicas, con unos cuantos miligramos tiene certidumbre de obtener una muestra precisa del total de los 5 litros de sangre presentes en el cuerpo humano. ¿Por qué basta con esta pequeña muestra? Porque se sabe que la sangre tiene escasa variabilidad en el cuerpo, es uniforme, y con un poco se puede inferir con confianza hacia la población total de sangre en dicha paciente.

Otra forma de ilustrarlo: si un bosque está compuesto por 100 árboles absolutamente iguales (la misma especie, el mismo grosor, la misma altura, la misma edad), basta con tomar mediciones de un solo árbol para inferir con precisión todo el bosque, ya que la variabilidad es prácticamente nula.

En contraste, piénsese en las poblaciones humanas. Las personas pueden variar en sexo, edad, origen étnico, características fisiológicas, nivel educativo, nivel socioeconómico, lugar de residencia, contexto social, actitudes políticas y un sinnúmero de características más. Esta variabilidad es la que hace que una encuesta deba construir muestras mayores a las que recoge el laboratorista para examinar la sangre: el investigador social debe tomar en cuenta toda la variabilidad social, por lo tanto su muestra tendrá que ser mayor si quiere estimar con precisión los parámetros.

Intervalos de confianza

La precisión de las estimaciones muchas veces se entiende por medio de los llamados intervalos de confianza, un aporte del estadístico Jerzy Neyman (1894-1981). La construcción de estos intervalos se explicará para dos casos: el de un promedio y el de un porcentaje.

Caso de un promedio. Siguiendo con el ejemplo ya referido de la encuesta en Montes de Oca, se quería inferir el promedio μ de las edades de los habitantes del cantón; para ello, se extrajo una muestra aleatoria de unas 600 personas. Como la teoría estadística indica que el promedio \bar{X} es un buen estimador de la media μ (\bar{X} es insesgado y eficiente), se procede a calcular este promedio simple con base en los datos de la muestra. El cálculo indica que el promedio muestral es de 55 años, y se denomina una estimación puntual. Además, se sabe que la desviación estándar en la población es 12.5 años.

Ahora, ¿cómo saber cuán precisa es la estimación de 55 años respecto a la media μ que es desconocida? Para ello se puede calcular el margen de error y construir intervalos de confianza.

Primero, se calcula el error estándar, definido como la desviación estándar σ entre la raíz cuadrada del tamaño de la muestra n , es decir $\frac{\sigma}{\sqrt{n}}$.

Luego se define un nivel de confianza, como por ejemplo 95%; con este se asume que el promedio muestral caerá en un rango de aproximadamente dos errores estándar con un probabilidad de 95% (siguiendo la llamada distribución normal). En específico, con 1.96 errores estándar se puede obtener una intervalo de 95%. Por lo tanto, el margen de error, al 95% de confianza, es:

$$\pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Obsérvese que conforme σ se incrementa, el margen de error aumenta; mientras que si n se incrementa, disminuye. Por esa razón, se espera tener muestras de suficiente tamaño que permitan disminuir el error. Sin embargo, nótese que en esta fórmula no se especifica el tamaño de la población; así, se pueden recopilar muestras de 600 personas tanto en Costa Rica como en Estados Unidos, donde los tamaños de la población son distintos (aproximadamente 4.8 millones vs. 319 millones de personas), pero si la variabilidad es idéntica (determinada por medio de la desviación estándar), los errores serán los mismos.

Con los datos del ejemplo, el error estándar sería $\frac{12.5}{\sqrt{600}} = 0.51$. Con una confianza del 95%, el margen de error es:

$$\pm 1.96 * 0.51 = \pm 1.0.$$

Los intervalos se calculan sumando, por un lado, y restando, por otro, el margen de error a la estimación puntual:

$$\text{Límite inferior: } 55.0 - 1.0 = 54.0.$$

$$\text{Límite superior: } 55.0 + 1.0 = 56.0.$$

En resumen, la fórmula de los intervalos de confianza al 95% para un promedio es:

$$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

El intervalo se interpreta de la siguiente manera: con una confianza del 95%, el intervalo [54.0 , 56.0] contiene el valor real (el parámetro) del promedio de edad de los habitantes de Montes de Oca. La confianza, por su parte, dice que si se extraen 100 muestras de idéntico tamaño, 95 de los 100 intervalos (es decir, el 95%) contendrían el valor real (parámetro) del promedio.¹²

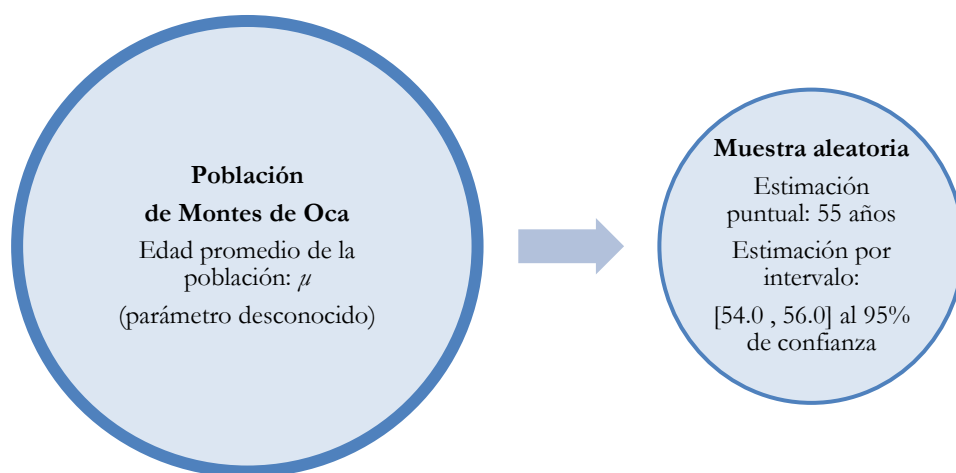


Figura 2.1. Representación gráfica de una población y su muestra.

Fuente: elaboración propia.

Puede advertirse que el margen de error corresponde a la estimación de una variable particular, en este caso de un promedio de edad; pero el error no sería igual, por ejemplo, para estimar el promedio de calificaciones otorgadas al gobierno, que cuenta con una desviación estándar distinta. Por ello, aunque popularmente se hable del margen de error de la “encuesta”, cada variable tiene su margen de error proveniente de su propia variabilidad, como se verá en el ejemplo de un porcentaje a continuación.

¹² Si se quisiera una mayor confianza, por ejemplo de 99%, se utiliza 2.58 en lugar de 1.96 pues con ± 2.58 veces el error estándar la probabilidad de obtener el resultado es 0.99 o 99%.

Caso de un porcentaje. Si la variable de interés en lugar de ser edad corresponde, por ejemplo, al porcentaje de apoyo al gobierno, se puede suponer que existe 56.0% de apoyo en la muestra de 600 personas de la encuesta de Montes de Oca. El margen de error se estima de una forma similar a la vista para el caso de la media. Para el caso de un porcentaje, la desviación estándar se calcula como $\sqrt{\hat{p} * (100 - \hat{p})}$, donde \hat{p} es el porcentaje muestral; o sea, $\sqrt{56.0 * (100 - 56.0)} = 49.6$. De modo que el margen de error del porcentaje desconocido p , con un 95% de confianza, es:

$$\pm 1.96 * \frac{49.6}{\sqrt{600}} = \pm 4.0 \text{ puntos porcentuales.}$$

Ahora se puede calcular un intervalo de confianza sumando y restando:

$$\text{Límite inferior: } 56.0\% - 4.0 = 52.0\%$$

$$\text{Límite superior: } 56.0\% + 4.0 = 60.0\%$$

Por lo tanto, el intervalo [52.0% , 60.0%] contiene el valor real de apoyo al gobierno con una confianza del 95% (o bien, que de 100 muestras del mismo tamaño, 95 intervalos contendrían el valor real del porcentaje).

En síntesis, el intervalo de confianza para el porcentaje se calcula así:

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

Finalmente, hay que realizar cierta aclaración respecto al tamaño de la muestra. Como se observa en la fórmula de los márgenes de error, cuanto mayor sea la muestra, el error disminuye, manteniendo constante la variabilidad. Sin embargo, la relación no es proporcional pues hay una raíz cuadrada para el tamaño de la muestra. Esto significa que si se aumenta demasiado la muestra, la reducción del error es cada vez más pequeña. Ya que la mayor muestra implica un costo operativo y económico más grande, puede caerse en un desperdicio de recursos que no se traduce en una mejoría real en la precisión (ver figura 2.2).

De modo que pasar de una muestra de 300 personas a una de 1200 es aconsejable en tanto reduce el error de ± 5.6 a ± 2.8 (casi ± 3 puntos porcentuales). Por el

contrario, una encuesta de 3000 personas solo disminuiría el margen ± 1 punto porcentual respecto a la muestra de 1200. La utilidad de muestras tan grandes dependería del grado de precisión requerido para ciertos objetos de estudio (donde no se admitieran errores mayores a ± 1.8 puntos porcentuales, en el caso de $n = 3000$) o de la necesidad de contar con submuestras considerables (por ejemplo, una submuestra de mujeres, entre el total de la muestra, que conlleve un margen de error aceptable).

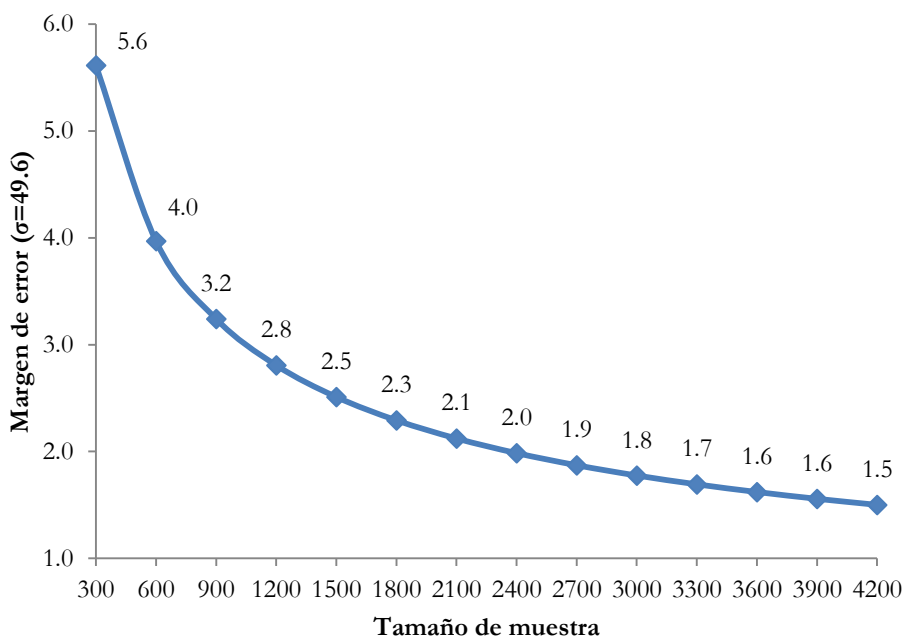


Figura 2.2. Relación entre el tamaño de muestra y el margen de error.

Fuente: elaboración propia.

Pruebas de hipótesis

Con base en los aspectos que se han tratado, es posible avanzar hacia métodos para probar las inferencias de los estimadores respecto a los parámetros. Estos se denominan pruebas o contrastes de hipótesis y algunos se verán en los próximos capítulos; por ahora, se examinan ciertos conceptos básicos.

Cuando se plantea una hipótesis estadística, es común formalizarla. Se establece una hipótesis nula que se pone a prueba y luego una hipótesis alternativa que contradice la anterior. Con las pruebas de hipótesis, se evalúa la evidencia contra la hipótesis nula; cuando la evidencia es significativa estadísticamente, ello implica un rechazo de la hipótesis nula.

Por ejemplo, siguiendo el caso de la encuesta en Montes de Oca, una hipótesis nula puede ser que la edad promedio es 55 años mientras la alternativa dice que la edad promedio no es 55 años; es decir:

- Hipótesis nula: $\mu = 55$ años.
- Hipótesis alternativa: $\mu \neq 55$ años.

Puesto que se trabaja en un marco probabilístico, necesariamente se asumen errores. Las pruebas de hipótesis postulan dos:

- *Error tipo 1*: rechazar la hipótesis nula cuando en realidad es verdadera (v. g. concluir que la edad promedio no es 55 años cuando en realidad sí lo es). La probabilidad de cometer este error se denota con α , llamado nivel de significancia.
- *Error tipo 2*: aceptar la hipótesis nula cuando en realidad es falsa (v. g. afirmar que 55 años es la edad promedio cuando en la población no lo es). Su probabilidad se denota con β .

En el cuadro 2.1, se resumen los escenarios de decisión y las consecuencias al aplicar las pruebas de hipótesis.

Cuadro 2.1. Decisiones y errores en las pruebas de hipótesis

Decisión	Situación real (desconocida)	
	Hipótesis nula es verdadera	Hipótesis nula es falsa
Rechazar la hipótesis nula	Error tipo 1	Decisión correcta
Aceptar la hipótesis nula	Decisión correcta	Error tipo 2

Fuente: Hernández (2010).

En el trabajo científico, lo común es fijar un nivel de significancia (es decir, una probabilidad aceptable de equivocarse al rechazar la hipótesis nula), mientras que la probabilidad de cometer un error tipo 2 no siempre se calcula.

Este nivel de significancia definido *a priori* (y con cierta arbitrariedad, según sus críticos) depende del campo de investigación en el que se trabaje; por ejemplo, en el clínico o el farmacéutico el equivocarse puede generar consecuencias fatales en la población, por lo que las pruebas deben llevarse a cabo con niveles muy estrictos. En ciencia política es habitual asignarse un nivel de significancia α de 0.05 (o 5%), aunque no tiene que ser necesariamente siempre así, pues el investigador es capaz de decidir cuánto riesgo asume.¹³

Por otro lado, es posible obtener manualmente o por medios de paquetes estadísticos los valores de significancia alcanzados o valores p . Los valores p , que oscilan entre 0 y 1, indican la evidencia a favor de la hipótesis nula para determinada prueba, donde un mayor número se interpreta como más apoyo para la hipótesis nula. Si es un valor bajo (menor al nivel de significancia α), entonces se tenderá a rechazar la hipótesis nula, pues no existe evidencia estadística suficiente para aceptarla y se declara el resultado como “estadísticamente significativo”.

Siguiendo el caso de Montes de Oca, imagínese que una socióloga ha escrito que la edad promedio de los habitantes del anterior cantón no puede ser 55, pues la población envejecida ha fallecido y se ha reemplazado por cohortes más jóvenes, o bien los adultos mayores han migrado a otros cantones por el alza en el valor de la tierra.

La analista plantea las hipótesis de la forma señalada (hipótesis nula indica que 55 es la edad promedio) y se propone un nivel de significancia del 0.05. Según la prueba, el valor p es de 0.01, por lo que la probabilidad de equivocarse al rechazar la hipótesis nula es más baja que el nivel aceptable, entonces se puede rechazar con seguridad y afirmar que la edad promedio no es igual a 55 (el sociólogo está en lo correcto entonces). Si, por el contrario, el valor p hubiese sido de 0.25, entonces el riesgo de equivocarse al rechazar la hipótesis nula es mayor al que se

¹³ Este 5% es equivalente a establecer un nivel de confianza del 95%.

acepta (1 de cada 4, en lugar de 1 de cada 20). De modo que no se puede rechazar esa hipótesis (si se hiciera, el riesgo de equivocarse sería muy alto).

Por el momento, esto puede parecer un tanto abstracto; en posteriores capítulos se verá cómo se utilizan las pruebas en la práctica.

Para terminar, se resumen los pasos para las pruebas de hipótesis:

- Para desarrollar pruebas se formaliza una hipótesis nula; esta hipótesis nula no necesariamente corresponde a la hipótesis teórica. En el ejemplo anterior, la hipótesis científica del sociólogo es contraria a la hipótesis nula. El procedimiento consiste en probar la hipótesis nula para llegar a conclusiones teóricas por contradicción (es decir, se rechaza que el promedio es igual 55, entonces el sociólogo tiene razón al afirmar que la edad promedio no es 55).
- Se escoge un nivel de significancia α (puede ser del 0.05, 0.1 o el deseado). Con el valor p que se obtiene del paquete estadístico, se decide: (a) rechazar la hipótesis nula si el valor p es menor al nivel de significancia determinado; (b) no rechazar la hipótesis nula si el valor p obtenido es igual o mayor al nivel de significancia.
- Se concluye sustantivamente según la decisión tomada respecto a la hipótesis nula.

Se podrá notar que la gran utilidad de interpretar de los valores p es que tiene suficientemente generalidad para poder concluir cualquier prueba de hipótesis (lo más importante es tener claro cuál es la hipótesis nula). Sin embargo, su uso exige una interpretación cuidadosa y las conclusiones de una investigación no deben basarse únicamente en un valor p sin tomar en cuenta las magnitudes de las correlaciones y de los efectos, la importancia práctica del hallazgo y la posibilidad de replicar el resultado en posteriores estudios (Cox, 1982; Wasserstein y Lazar, 2016).

Comentarios finales

Se han visto algunos aspectos fundamentales de la teoría de la inferencia clásica; sin embargo, no constituye la única teoría estadística. Existe, además, un paradigma denominado bayesiano, enfoque que rivaliza con la inferencia clásica y

que recientemente se ha revitalizado con el desarrollo computacional, necesario para muchos de sus cálculos. La perspectiva bayesiana permite incorporar las creencias subjetivas del investigador y “actualizarlas” con los datos del fenómeno. En los trabajos de Jackman (2000 y 2004), puede leerse una breve introducción al enfoque con aplicaciones en ciencia política.

Entre las críticas más recurrentes a la estadística clásica, se encuentra la asignación arbitraria de un nivel de significancia (como 5%) y el hecho de que los tamaños de muestra afecten las pruebas de hipótesis (ver Lindley, 2000).

Este último fenómeno es especialmente pertinente al trabajar con grandes, inmensas bases de datos (no muestras de mil personas, sino registros de millones, como puede ser un padrón electoral). En estos casos, es fácil rechazar las hipótesis nulas solamente por el tamaño grande de muestra —recuérdese que n divide la variabilidad en el error, por lo que si n es enorme, el error es mínimo—, llegando a inferir todo tipo relaciones (incluso espurias) entre los más disímiles fenómenos. Con este tipo de datos la inferencia clásica puede ser inútil, mientras que los métodos y algoritmos de la “minería de datos” (también llamada *big data*) resultarían más adecuados (para una introducción panorámica, consultar Riquelme, Ruiz y Gilbert, 2006).

Ejercicios

1. En la Encuesta Nacional de Hogares de 2013 (INEC, 2013) se estimó que el ingreso per cápita promedio en el país es de 328 688 colones, con un margen de error de $\pm 17\,996$ colones. Construya el intervalo de confianza al 95% e interprételo.
2. También en la Encuesta Nacional de Hogares de 2013 (INEC, 2013) se estimó el porcentaje de hogares clasificados como pobres, según la metodología de línea de pobreza. Se estima que un 20.7% de los hogares son pobres, con un margen de error de ± 0.98 puntos porcentuales. Construya el intervalo de confianza al 95% e interprételo.
3. Los datos del censo 2011 de Costa Rica (INEC, 2011) indican que la población total es de 4 301 712 personas, pero no se precisa ningún margen de error. ¿A qué se debe esto? ¿En qué se diferencia un dato censal de un dato proveniente de una encuesta?

Opcional. Calcule el margen de error y construya el intervalo de confianza al 95% para el caso de intención de voto por un candidato del 66.8%, según una encuesta realizada a 1200 personas.

CAPÍTULO 3

COMPARACIÓN DE DOS MEDIAS

Introducción

¿Es el promedio de ingresos de los hombres igual al de las mujeres? ¿Las notas que se otorgan a un gobierno difieren entre jóvenes y adultos? ¿Es el grado de satisfacción con las políticas públicas igual entre habitantes de regiones urbanas y rurales? Estas son algunas preguntas que pueden surgir en el estudio de la política y los datos de encuestas o de otras fuentes aportan información relevante para contestarlas.

Puede distinguirse que en estos casos se está cuestionando sobre una variable métrica (de intervalo o razón): el ingreso, una nota otorgada al gobierno, una escala de satisfacción. Los promedios de estas variables se comparan entre dos grupos independientes: hombres y mujeres, jóvenes y adultos, pobres y no pobres. Es decir, se codifican como variables categóricas binarias o dicotómicas.

Para este tipo de datos, la comparación de dos muestras independientes es posible utilizar la llamada prueba t de Student. El nombre de este método proviene del trabajo de William S. Gosset (1876-1937), químico y matemático, que trabajaba en la cervecería Guinness. Al estudiar procesos de producción (como las cantidades apropiadas de levadura en la cerveza), obtuvo resultados novedosos para la teoría estadística; pero, puesto que la compañía mantenía una política que restringía las publicaciones científicas de sus empleados (para evitar difusión de información confidencial), Gosset publicó su artículo bajo el seudónimo de “Student” o estudiante en inglés (Salsburg, 2002).

Procedimiento

A modo de ejemplo, se quiere comparar si la nota con que califican las personas a la Defensoría de los Habitantes es igual entre hombres y mujeres con base en los datos de la encuesta de noviembre de 2013 del Centro de Investigación y Estudios Políticos (CIEP, 2012-2014) de la Universidad de Costa Rica. Al preguntar, “¿Qué nota, de 0 a 10, le pondría a la Defensoría de los Habitantes, donde 0 es la peor y 10 la mejor?”, el promedio global fue 6.84.

Pero, ¿será la calificación promedio de la Defensoría igual entre hombres y mujeres? Es decir, en términos estadísticos se formulan las hipótesis:

- Hipótesis nula: nota promedio entre hombres = nota promedio entre mujeres.
- Hipótesis alternativa: nota promedio entre hombres \neq nota promedio entre mujeres.

En el ejemplo, claramente se compara una variable continua o métrica (nota de 0 a 10) en dos grupos identificados por una variable categórica (sexo), por lo que se puede aplicar la prueba *t* para muestras independientes.

El procedimiento común en SPSS se explica a continuación.

Recuadro 3.1

*Resumen del procedimiento para la prueba *t* en SPSS*

Analizar → Comparar medias → Prueba T para muestras independientes

En la ventana izquierda se selecciona la variable métrica cuyas medias se quieren comparar y se traslada a Variables para contrastar.

En esta misma ventana, se selecciona la variable categórica o con aquella que se determinan los grupos 1 y 2. Se traslada a Variables de agrupación y se definen los valores de la variable que indican los grupos respectivamente.

Aceptar.

Para ejecutar la prueba con los datos del ejemplo, se abre la ventana de Prueba T para muestras independientes y se trasladan las variables de interés. La variable

para contrastar es aquella de naturaleza métrica (la nota de la Defensoría) y la variable de agrupación la categórica (sexo de la persona encuestada) (figura 3.1).

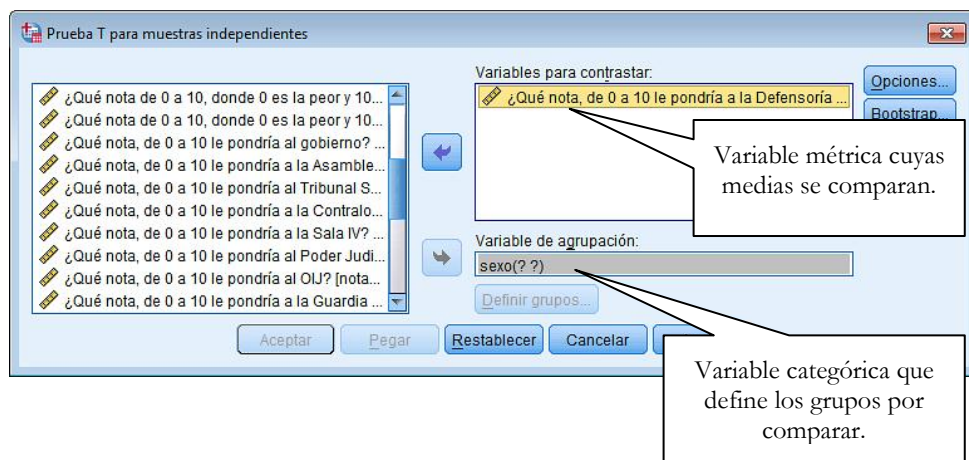


Figura 3.1. Ventana de la prueba T para muestras independientes en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Ahora, se definen los grupos de la variable categórica sexo, hombres con 0 y mujeres con 1, pues así están codificadas las categorías en la base de datos (figura 3.2).

El paquete también ofrece la opción de punto de corte, es decir, que si la variable no tiene grupos predefinidos, como en el caso de variables métricas, pueden categorizarse de manera *ad hoc*. El punto de corte indicado crea un grupo mayor o igual al número que se escriba y otro menor a este. Por ejemplo, si se tiene una variable métrica de edad, pueden generarse dos grupos para la prueba; si se define 45 años como punto de corte, serán (a) 45 o más y (b) menos de 45 (o 44 o menos). Esta transformación de una variable métrica en una categórica muestra el uso creativo de los niveles de medición, referidos en el capítulo 1. Pero si la variable ya es categórica, como en el ejemplo, entonces se sigue con Continuar y Aceptar.¹⁴

¹⁴ Para este y los demás procedimientos del texto se recuerda que, en lugar de Aceptar, puede utilizarse Pegar; con esta última opción, se registra la secuencia de comandos en una ventana de sintaxis. Luego simplemente se ejecutan las líneas de comandos. La gran ventaja que ello implica es la comodidad para repetir o replicar el procedimiento.

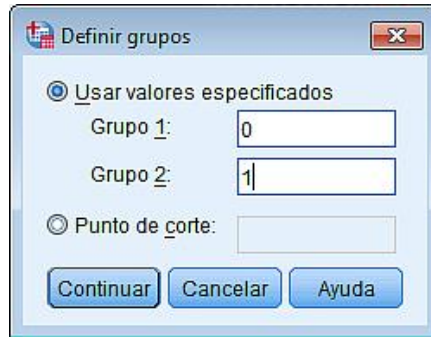


Figura 3.2. Definición de los grupos para la variable categórica en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Adicionalmente, en la sección de Opciones (figura 3.3) es posible definir cómo se construye el intervalo de confianza. El porcentaje preestablecido es 95% (es decir, utiliza el valor 1.96 para construir el intervalo de confianza); pero, perfectamente, se podría definir 99% para obtener un intervalo de confianza al 99% (calculando con 2.58).

Ahora se analiza la salida (figura 3.4). Primero, SPSS ofrece un cuadro descriptivo donde se indica el tamaño de muestra para cada grupo, la media, su desviación estándar y el error estándar de la media.¹⁵ Como puede verse, la nota promedio entre los hombres es 6.52 mientras que en las mujeres 7.09. Aunque intuitivamente la nota promedio de las mujeres parece superior a la de los hombres, recuérdese que la inferencia estadística conlleva un margen de error que debe tomarse en cuenta al sacar conclusiones. Por ello, se aplica la prueba *t*.

¹⁵ Como se vio en el capítulo anterior, este último consiste en dividir la desviación estándar entre la raíz cuadrada de su tamaño de muestra.

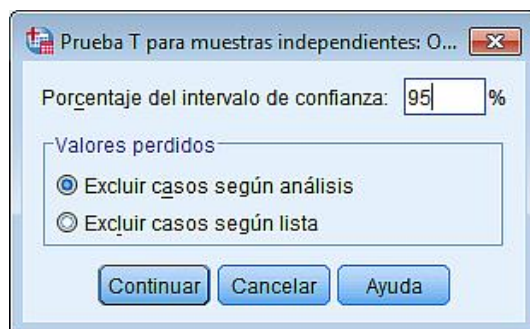


Figura 3.3. Porcentaje del intervalo de confianza en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Estadísticos de grupo					
	sexo Sexo	N	Media	Desviación típ.	Error típ. de la media
notadefensoría ¿Qué nota, de 0 a 10, le pondría a la Defensoría de los Habitantes?	1,00 hombre	257	6.5253	2.31004	.14410
	1,00 mujer	335	7.0866	2.01825	.11027

Prueba de muestras independientes										
		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95% Intervalo de confianza para la diferencia	
									Inferior	Superior
¿Qué nota, de 0 a 10, le pondría a la Defensoría de los Habitantes?	Se han asumido varianzas iguales	3.739	.054	-3.149	590	.002	-.56128	.17826	-.91138	-.21117
	No se han asumido varianzas iguales			-3.093	509.656	.002	-.56128	.18145	-.91775	-.20480

Figura 3.4. Salida de la prueba t para muestras independientes en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

El segundo cuadro indica el resultado de la prueba. La primera fila muestra el resultado si se asume igualdad de variancias; el segundo, si esto no se puede sostener. Este es un supuesto importante para la prueba que se examinará más adelante; por el momento, véase solamente el resultado de la fila superior (cuando se asume igualdad de variancias).

Se puede observar que la probabilidad de equivocarse al rechazar la hipótesis nula, la cual indica que las notas promedio entre hombres y mujeres son iguales, si fuera cierta es 0.002. Es decir, la probabilidad de cometer el error tipo 1 es 0.002 (muy baja). Si se trabaja con un nivel de significancia (α) de 0.05, entonces se puede rechazar la hipótesis nula.

En un modo más simple, como el valor $p = 0.002$ es menor al $\alpha = 0.05$, entonces se rechaza la hipótesis de que son iguales los promedios. Por ende, hay diferencias estadísticamente significativas en la nota de la Defensoría de los Habitantes entre hombres y mujeres.

Otra forma de ver el resultado es el siguiente: se formuló la hipótesis nula, esta dice que el promedio entre los hombres es igual al promedio entre las mujeres:

$$\text{Hipótesis nula: } \text{promedio}_{\text{hombres}} = \text{promedio}_{\text{mujeres}}$$

Pero lo anterior es equivalente a:

$$\text{Hipótesis nula: } \text{promedio}_{\text{hombres}} - \text{promedio}_{\text{mujeres}} = 0$$

En otras palabras, se prueba si la diferencia entre promedios es estadísticamente igual a 0 (es decir, las medias son iguales).

Ahora bien, se puede notar que la salida de SPSS indica una diferencia de medias igual a -0.561 (el resultado de restar las medias: $6.53 - 7.09 = -0.56$). También, incluye un intervalo de confianza calculado al 95% de forma similar a la vista en el capítulo 2, donde el grupo 1 corresponde a los hombres y el 2 a las mujeres:¹⁶

$$(\bar{X}_1 - \bar{X}_2) \pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} =$$

$$(6.53 - 7.09) \pm 1.96 \sqrt{\frac{2.31^2}{257} + \frac{2.02^2}{335}} = -0.56 \pm 0.36.$$

¹⁶ Cuando la muestra es pequeña (aproximadamente menor a 30 en cada grupo), no se calcula con el 1.96 que proviene de una distribución normal, sino con valores de la distribución t de Student.

Es decir, que el intervalo calculado $[-0.92, -0.20]$ contiene el valor real de la diferencia entre los promedios (con un 95% de confianza). Obsérvese que el intervalo va de un número negativo a otro también negativo, por lo que no incluye al cero; y si el intervalo no incluye al cero, entonces no se puede afirmar que la diferencia de promedios sea cero como indicaba la hipótesis nula. Por lo tanto, debido a este razonamiento también se rechaza la hipótesis de igualdad de medias.¹⁷

En conclusión, como el valor p es más bajo que 0.05, se rechaza la hipótesis nula de que las medias son iguales: hombres y mujeres tienen diferentes notas promedio para la Defensoría. Las mujeres otorgaron una nota más alta que los hombres y esto es un resultado confirmado por el análisis. También se puede concluir lo mismo diciendo que la diferencia entre los promedios no es cero, es decir, los promedios de notas no son iguales.

Nota sobre igualdad de variancias

La prueba t supone que la variabilidad dentro de cada grupo sea igual (en el ejemplo, dentro del grupo de hombres y dentro del grupo de mujeres), en otras palabras, que las variancias o las desviaciones estándar sean estadísticamente iguales. Eso se puede examinar con la prueba de Levene que muestra la salida de SPSS (figura 3.4). En el ejemplo visto, su significancia es 0.054, por lo que las variancias no son distintas (no se rechaza la hipótesis nula de que las variancias son iguales con $\alpha = 0.05$). Si fuesen distintas, entonces se debería examinar el resultado inferior (que en este caso es muy similar y lleva a las mismas conclusiones) aduciendo el incumplimiento.

Por su parte, la fórmula expuesta para calcular el intervalo de confianza de la diferencia corresponde a asumir variancias o desviaciones estándar diferentes (hay s_{hombres} y s_{mujeres}). Cuando se puede suponer igualdad de variancias, el intervalo requiere el cálculo de una variancia combinada que no se muestra en este documento.

¹⁷ Para calcular diferencias entre dos porcentajes p_1 y p_2 de muestras independientes, se puede utilizar la misma fórmula simplemente sustituyendo la desviación estándar s por $\sqrt{\hat{p} * (100 - \hat{p})}$, es decir:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{\frac{\hat{p}_1(100 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(100 - \hat{p}_2)}{n_2}} \text{ (Hernández, 2010, p. 118).}$$

Comentarios finales

Nótese que para este método se ha hablado de dos grupos independientes de observaciones, es decir, que no hay observaciones repetidas entre ellas o relacionadas (por ejemplo, no serían independientes si se compara un promedio medido antes con otro medido después para un mismo grupo). La gran limitación que presenta la prueba es que compara solamente dos grupos; para hacerlo con tres o más grupos o tratamientos, se utiliza el análisis de variancia (ANOVA), el cual se expondrá próximamente.

Este método, como muchos otros, se desarrolló originalmente para datos de naturaleza experimental, en que el propio diseño impone controles para poder obtener con pureza los efectos de un tratamiento. En estudios observacionales, más comunes en ciencia política, su uso debe ser cuidadoso. Por lo general, hay variables que no se pueden controlar artificialmente, sino que se incorporan al análisis. Por ello, una prueba como la de comparación de medias, la cual implica un único “tratamiento” (la variable categórica), debe utilizarse con cautela y preferiblemente como un paso previo para análisis posteriores que controlen por otras variables adicionales. Es decir, para el ejemplo visto, no se sabe si la valoración promedio mayor entre las mujeres respecto a la de los hombres ocurre porque son mujeres o por alguna otra variable asociada con ser de este sexo.

Ejercicios

1. Con los datos de la encuesta de noviembre de 2013 del CIEP, pruebe si los promedios de notas otorgadas a la entonces presidenta Laura Chinchilla (variable *notalaura*) son estadísticamente iguales entre hombres y mujeres (variable *sexo*). Utilice un nivel de significancia (α) del 5%.
2. Con los datos de la encuesta de noviembre de 2013 del CIEP, compare la nota promedio otorgada al gobierno (variable *notagobierno*) entre personas con nivel educativo de primaria o menos y aquellas con grado universitario (variable *educacionrec*). Utilice un nivel de significancia (α) del 5%. Sugerencia: explore primero cómo está codificada la variable de educación para poder definir correctamente los grupos.

3. En los datos de la encuesta de noviembre de 2013 del CIEP, la variable edad está medida en años cumplidos. Interesa conocer si, entre personas jóvenes (de 35 años o menos), la opinión sobre la iglesia católica es diferente respecto a las personas adultas (de más de 36 años). Compare el promedio de la nota otorgada a la iglesia católica (variable *notaiglesiaca1*) entre los dos grupos y concluya con un nivel de significancia (α) del 5%. Sugerencia: puede utilizar la opción de punto de corte o recodificar edad en una nueva variable.

Opcional. Utilice los datos del cuadro 3.1 y la fórmula para el intervalo de confianza al 95% con variancias diferentes para determinar si los promedios de edades son estadísticamente iguales entre hombres y mujeres encuestados en noviembre de 2013 por el CIEP. Confirme el resultado manual con SPSS.

Cuadro 3.1. Edades promedio

	Promedio de edad	Desviación estándar	<i>n</i>
Hombres	43.75	17.99	269
Mujeres	45.91	16.01	366

Fuente: CIEP (2012-2014).

CAPÍTULO 4

ANÁLISIS DE VARIANCIA DE UN FACTOR

Introducción

Ya se ha visto que si se quieren comparar dos promedios con el fin de establecer si son estadísticamente diferentes, se puede utilizar la llamada prueba *t* de comparación de medias. ¿Qué pasa si se requiere comparar tres, cuatro o más medias? Para ello se utiliza el análisis de variancia, abreviado comúnmente como ANOVA.¹⁸

El análisis de variancia o ANOVA fue creado por uno de los fundadores de la estadística moderna, Ronald A. Fisher, gracias a su trabajo en la finca experimental de Rothamstead en Inglaterra. ¿Cuál fue el problema que enfrentó Fisher y cuál solución propuso? Como lo relata Salsburg (2001, pp. 46-48), en Rothamstead se ejecutaban pruebas de fertilizantes en los cultivos y —antes de la llegada de Fisher— el procedimiento usual consistía en dividir el terreno en dos partes y aplicar un tipo de fertilizante a cada una. Ahora bien, los terrenos no eran homogéneos: en algunas partes crecía más maleza que en otras, había más agua en determinados lugares, cambiaba el tipo de tierra, habían distintos nutrientes, la inclinación del terreno variaba, etc.; de tal manera resultaba difícil distinguir los efectos “reales” de los fertilizantes que se examinaban.

Por ejemplo, para probar el efecto de dos fertilizantes distintos sobre un cultivo, podría diseñarse un terreno dividido en cuatro partes: una norte donde se drena más el agua y en la sur, menos. Entonces se aplica a una mitad del norte el fertilizante A y a la otra mitad el fertilizante B; igualmente en terreno sur. Sin embargo, incluso diseñando el experimento de esta forma hay factores que no se

¹⁸ La sigla ANOVA proviene del inglés *analysis of variance*; en algunos textos en español se denomina ANDEVA.

están observando y, por lo tanto, tampoco se controlan (como la composición del suelo, etc.), por lo que no es posible decir si las diferencias en los cultivos resultantes se deben a los tipos de fertilizantes. Esto significa que hay factores *confusores* que impiden detectar los efectos de un tratamiento.

La original propuesta de Fisher consistió en que la asignación de los tratamientos (el tipo de fertilizante) se hiciera al azar: un orden aleatorio no sigue ningún patrón fijo y los efectos que no se controlan se estarían cancelando entre sí sin relacionarse con el tratamiento (lo cual se fundamenta en una demostración matemática, naturalmente).¹⁹

En este capítulo, se estudiará el ANOVA llamado de un factor o de una vía, es decir, que implica solo un tratamiento.

Un ejemplo experimental

Supóngase que se tiene un grupo 18 de estudiantes universitarios, escogidos al azar de la lista completa del registro, que se convocan a un laboratorio de ciencia política para realizar un experimento; todos aceptan con interés y firman su anuencia a participar en el estudio, con lo cual se cumplen los requisitos éticos de la universidad.

A cada estudiante se le ofrecen dos materiales: una foto de una persona y un cuestionario con una pregunta. Ahora bien, hay tres tipos de fotos: una con la persona sonriente, otra con la misma persona seria y otra con la misma persona con una expresión facial neutral.²⁰ Como son 18 estudiantes, hay 6 fotografías por tipo de expresión facial, las cuales se asignan de forma aleatoria al grupo de universitarios. Estas fotografías corresponden al tratamiento o factor.

¹⁹ Fisher no fue, sin embargo, el primero en sugerir la asignación aleatoria en los experimentos; Charles Peirce y Joseph Jastrow implementaron un cuidadoso experimento, en el que se recurrió a los naipes para aleatorizar, 50 años antes, aunque no tuvieron el mismo impacto que Fisher (Stigler, 1978).

²⁰ Se evita utilizar retratos de personalidades que los participantes puedan reconocer para no incorporar valoraciones previas en el experimento (las cuales se convertirían en un factor confusor).

Las personas participantes observan la foto que les tocó y luego se dirigen al cuestionario que pregunta lo siguiente: “En una escala donde 0 significa nada seguro y 10 completamente seguro, ¿cuánto se inclinaría en votar usted por este candidato presidencial?”. Los resultados obtenidos se resumen en el cuadro 4.1.

Cuadro 4.1. Resultados del experimento

A: Candidato sonriente	B: Candidato serio	C: Candidato neutro
9	2	5
10	4	6
8	6	7
9	3	4
8	5	8
10	4	6
Promedio _A : 9.0	Promedio _B : 4.0	Promedio _C : 6.0

Fuente: elaboración propia.

Como se observa en la última fila del cuadro 4.1, los promedios dados a cada candidato son 9.0, 4.0 y 6.0. Ahora, ¿cómo saber si las diferentes notas promedio provienen de la variación en expresión facial del candidato? En ese sentido, el interés es conocer si las expresiones faciales influyen sistemáticamente en la valoración de los candidatos (para una reseña de estudios similares sobre la apariencia de los candidatos, ver Lawson *et al.*, 2010). Para ello, los investigadores aplican el análisis de variancia de Fisher, que se verá paso a paso.

Primero, se plantean las hipótesis estadísticas:

- Hipótesis nula: las medias de las notas dadas al candidato son iguales.
- Hipótesis alternativa: las medias no son iguales (al menos un par es diferente).

Como es usual en los procedimientos estadísticos, se probará la hipótesis nula de que los promedios A, B y C son iguales, para luego concluir si no son iguales. Nótese que no se está probando si todos son diferentes, solamente si al menos dos promedios no son iguales. Tampoco se prueba si uno es mayor que otro. Lo que se examinará es si el tratamiento (las distintas expresiones faciales del candidato) tiene efectos no nulos sobre las notas dadas por los participantes.

Seguidamente, los investigadores hacen un gráfico (figura 4.1) donde ubican las notas dadas; además, trazan una línea en los promedios para cada grupo (candidato) y otra para el promedio total de todas las notas (línea segmentada).

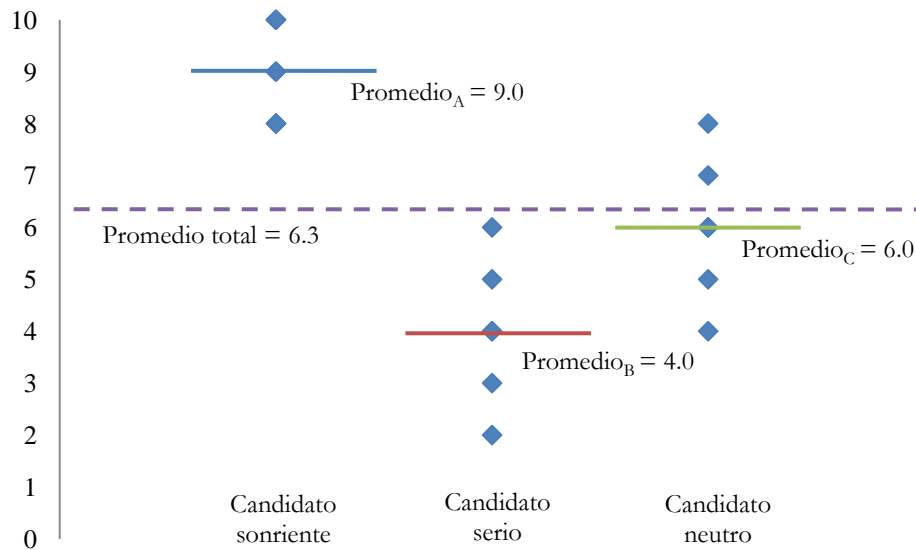


Figura 4.1. Gráfico del experimento de candidatos.

Fuente: elaboración propia.

Véase que cada puntuación se agrupa alrededor de su promedio y algunos puntos están más cerca que otros; esta distancia en cada grupo se denomina variancia intragrupo o dentro del grupo. A su vez, los promedios de cada factor o tratamiento (A, B y C) se alejan del promedio total: esta es la variancia entre grupos. Además, cada punto individual de todos los tratamientos se aleja del promedio total, lo que corresponde a la variabilidad total. Luego se puede demostrar que la variabilidad total es igual a la variabilidad entre los grupos más la variabilidad intragrupo o dentro de cada grupo:

$$\text{Variabilidad total} = \text{variabilidad entre} + \text{variabilidad dentro.}$$

El punto clave es que si el factor es significativo, es decir, las diferencias se deben en realidad al factor (el tipo de expresión facial), entonces la variabilidad entre los grupos es mayor que la variabilidad dentro de cada grupo porque la variabilidad

total está explicada por las diferencias entre los grupos por el factor. Por ello, la variabilidad dentro de los grupos es denominada *error*. Piénsese que si las observaciones no varían dentro de cada grupo, o sea, los únicos valores por grupo corresponden al promedio de cada tratamiento, entonces el error es cero y la variabilidad total es perfectamente explicada por la variabilidad entre grupos.

Cuadro 4.2. Resultado del ANOVA para el experimento de candidatos

	Suma de cuadrados	Grados de libertad	Cuadrado medio	Significancia
Entre grupos	76.0	2	38.0	0.000
Dentro de grupos	24.0	15	1.6	
Total	100.0	17		

Fuente: elaboración propia.

Cuando los investigadores ejecutan un ANOVA en algún paquete estadístico, la salida es similar a la presentada en el cuadro 4.2. Se tienen filas para las tres fuentes de variación mencionadas: entre grupos, dentro de grupos y la total. En la primera fila se indican la suma de cuadrados, es decir, las desviaciones cuadradas respecto a los promedios.²¹ Los grados de libertad provienen del número de observaciones y sirven para calcular el cuadrado medio (que es igual a la suma de cuadrados entre los grados de libertad).

Últimamente interesa llegar al valor *p* o la significancia. Como es usual, este indica la probabilidad de equivocarse rechazando la hipótesis nula (las medias son iguales) siendo esta cierta. En este caso, el valor *p* es 0.000, lo cual es menor a cualquier nivel de significancia y por ello se rechaza la hipótesis nula de que las medias son iguales, es decir, al menos dos de ellas son diferentes. Sustantivamente, los investigadores concluyen que la expresión facial incide en obtener distintas valoraciones promedio para los candidatos.

²¹ Nótese que $76.0 + 24.0 = 100$, o sea, que la suma de cuadrados entre grupos más la suma de cuadrados dentro de grupos es igual a la suma de cuadrados total, mostrándose la igualdad citada de la variabilidad total.

Además, cuando el factor es significativo (el valor p es menor al alfa establecido), entonces el cuadrado medio entre grupos es mayor al cuadrado medio dentro de grupos y se puede decir que los promedios no son iguales, ya que las desviaciones (suma de cuadrados) corregidas por los grados de libertad entre grupos “capturan” la variabilidad total. Pero si el factor no es significativo, no hay ninguna implicación necesaria respecto a los cuadrados medios, puede ser mayor dentro de grupos o menor entre grupos; por ello es necesario concluir con base en la significancia.

Procedimiento

Para aplicar un ANOVA de un factor, es necesario contar con dos variables:

- una categórica o cualitativa (el factor) que define los grupos o tratamiento;
- otra que sea métrica o continua cuyas medias se comparen.

El procedimiento general en SPSS se presenta a continuación.

Recuadro 4.1

Resumen del procedimiento para análisis de variancia en SPSS

Analizar → Analizar medias → ANOVA de un factor

Trasladar la variable continua de la ventana izquierda a la derecha llamada Lista de dependientes. Igualmente trasladar de la primera a la segunda el factor.

En Opciones, activar Descriptivos. Luego Continuar.

Aceptar.

Aunque el ejemplo anterior utilizó el ANOVA para datos experimentales, también se puede aplicar en el contexto de datos observacionales, aunque se debe tener cautela en las conclusiones. Con los datos de la encuesta de noviembre de 2013 del CIEP, se compararán los promedios de la valoración del Tribunal Supremo de Elecciones (TSE) según la región donde vive la persona encuestada.

Por lo tanto, el factor (tratamiento) es la región: metropolitana, resto del Valle Central y resto del país. Las medias que se comparan vienen del puntaje entre 0

(muy malo) y 10 (excelente) que responde a “¿Qué nota, de 0 a 10, le pondría al Tribunal Supremo Elecciones?”.

Entonces, en SPSS se busca en el menú Analizar, Analizar Medias y ANOVA de un factor para abrir la ventana presentada en la figura 4.2.

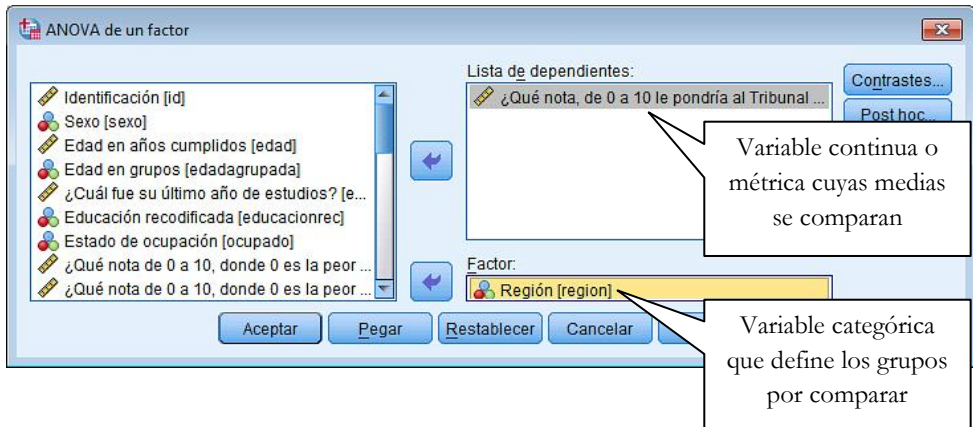


Figura 4.2. Ventana de ANOVA de un factor en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

En factor se desplaza la variable que defina a los grupos o tratamientos, en este caso Región. En la lista de variable dependiente, se introduce la variable de intervalo o métrica con la cual se quieren comparar las medias, o sea, la calificación del TSE.

Luego, en Opciones se seleccionan las opciones de Estadísticas Descriptivas y Aceptar (figura 4.3).

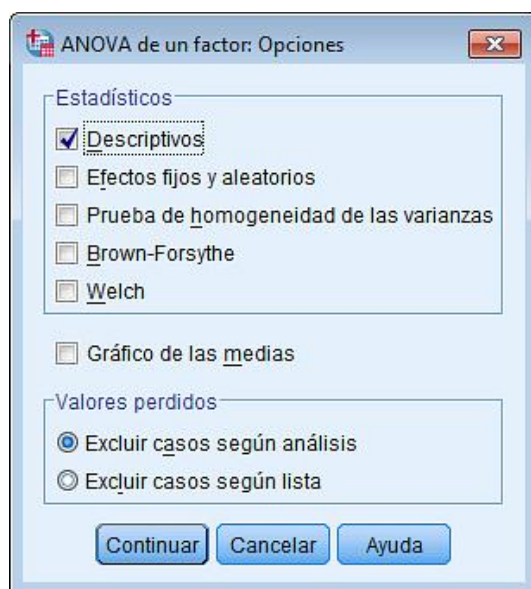


Figura 4.3. Ventana de opciones del ANOVA de un factor en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Las salidas, primero, ofrecen una descripción de las variables (figura 4.2). En la región metropolitana, la puntuación promedio para el TSE fue de 6.32, en el resto del Valle Central 6.71 y en el resto del país 6.99. Pero se sigue sin saber si estas medias son *estadísticamente* iguales o no.

En el segundo cuadro se encuentra el resultado del ANOVA. El valor p es 0.024, por lo que con un 0.05 de significancia se puede rechazar la hipótesis nula. En otras palabras, las medias no son iguales al nivel de significancia del 5%. Puede verse, además, que el cuadrado medio entre grupos es mayor al cuadrado medio dentro de grupos, es decir, que las notas varían más de una región a otra que dentro de cada región. Por lo tanto, se puede concluir que las personas de distintas regiones no otorgan las mismas calificaciones promedio al TSE.

Descriptivos

notatse ¿Qué nota, de 0 a 10, le pondría al Tribunal Supremo de Elecciones?

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1,00 Metropolitana	326	6.3190	2.48368	.13756	6.0484	6.5896	.00	10.00
2,00 Resto del Valle Central	157	6.7070	2.32120	.18525	6.3411	7.0729	.00	10.00
3,00 Resto del país	107	6.9907	2.00234	.19357	6.6069	7.3744	.00	10.00
Total	590	6.5441	2.37143	.09763	6.3523	6.7358	.00	10.00

ANOVA

notatse ¿Qué nota, de 0 a 10, le pondría al Tribunal Supremo de Elecciones?

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	42.019	2	21.010	3.771	.024
Intra-grupos	3270.335	587	5.571		
Total	3312.354	589			

Figura 4.4. Resultados del ANOVA de una vía en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Como se indicó previamente, con el análisis de variancia no se puede decidir cuál es la nota mayor, es decir, si la nota promedio del TSE en el resto del país es mayor a la nota en el resto del Valle Central o a la de la región metropolitana. Solamente se sabe que no son iguales.

Sin embargo, en algunos casos es posible comparar los intervalos de confianza para cada media para determinar cuáles promedios pueden ser mayores y menores según los traslapes. El intervalo para el resto del país es [6.607 , 7.374] y se traslapa con el intervalo para el resto del Valle Central [6.341 , 7.073], por lo que son estadísticamente iguales. Pero el intervalo para el resto del país no se cruza con el de la región metropolitana, pues su límite superior es 6.590 y el límite inferior para el resto del país es 6.607, de modo que estos promedios son estadísticamente diferentes.

Este procedimiento de comparación de intervalos de confianza contiene dos complicaciones. Primero, si son muchos los promedios, el número de pares que

habría que comparar sería exagerado.²² El segundo, desde el punto de vista estadístico, es que no se está fijando un error tipo 1 simultáneamente en todas las comparaciones, sino que se aumenta por cada comparación, debido a que ya los intervalos están contruidos al 95% de confianza, es decir, cada uno asume un 5% de error tipo 1. Lo anterior sería similar a ejecutar pruebas t (vistas en el capítulo 3) para cada par posible de medias que conlleva los mismos inconvenientes de comparar intervalos de confianza (Agresti y Franklin, 2013).

Comentarios finales

- Tanto la anécdota de Fisher como el ejemplo de los candidatos, se aplicaron en el contexto de experimentos controlados donde la diferencia de efectos se puede atribuir al factor o tratamiento, ya que todas las demás fuentes de variación se consideran como error aleatorio. En esos casos la causalidad es clara.
- Con datos observacionales —como los de una encuesta—, si bien el ANOVA es útil para reconocer diferencia entre medias, es arriesgado pensar en causas y efectos, pues la asignación de tratamientos no es aleatoria. Por ejemplo, puede que los niveles de acceso a medios de comunicación y a la educación formal diferenciados por región sean los verdaderos causantes de la variabilidad en la percepción del TSE; para controlar las demás variables y no atribuir al factor efectos causales que no le corresponden (lo cual se llama sesgo de variable omitida y se abarcará en el capítulo 7) es necesario recurrir a otros métodos como el análisis de regresión múltiple.
- Se recuerda de nuevo que el ANOVA permite analizar si las medias son iguales estadísticamente, pero no se sabe cuáles de ellas son diferentes. Además, comparar por intervalos de confianza o comparar medias por pruebas t no constituyen estrategias óptimas por las razones ya señaladas. La alternativa es recurrir a los contrastes múltiples (ver Agresti y Franklin, 2013, pp. 693-695).

²² Si k es el número de promedios, las comparaciones se calculan como $k(k-1)/2$. Por ejemplo, con 7 promedios (p. g. las provincias de Costa Rica), entonces se tendrían que hacer 21 comparaciones.

- Al igual que con la prueba t , el ANOVA conlleva ciertos supuestos, como normalidad en la distribución, en los cuales no se ahondaron, pero su uso riguroso exige su diagnóstico.

Ejercicios

1. Con los datos del cuadro 4.1 o con otros de un experimento efectuado en el aula, utilice SPSS para realizar un análisis de variancia y concluir si existen efectos significativos del tratamiento experimental al 5%.
2. Con los datos de la encuesta de noviembre de 2013 del CIEP, utilice la variable *edadagrupada* como factor para probar si entre distintos grupos etarios las notas promedio que se otorgan a la Asamblea Legislativa son iguales con una significancia del 5%.
3. La variable *indicemedios* de la encuesta de noviembre de 2013 del CIEP corresponde al conteo de medios de información que utiliza cada persona (entre 0 y 6). Aplique el análisis de variancia para constatar si entre personas con diferente nivel educativo (*educacionrec*) el uso promedio de fuentes de información es igual (al 5%).
4. En el capítulo 3, se vio la prueba t para comparar dos medias, mientras que el análisis de variancia se presentó para el caso de tres o más medias. ¿Qué ocurre al aplicar un ANOVA para comparar dos medias? Replique el ejercicio 1 del capítulo 3, compare los resultados e intente explicar qué ocurre.

CAPÍTULO 5

MEDIDAS DE ASOCIACIÓN

Introducción

Entre la descripción y la relación causal, se puede establecer un grado intermedio al examinar dos o más variables: la asociación.²³ Con ella se puede entender que una característica varía de forma conjunta con otra, pero ello no basta para establecer una dirección de tipo causa-efecto, pues no hay una secuencia temporal donde la causa precede al efecto. Estos son algunos ejemplos de asociación en ciencia política:

- Existe una asociación entre sistemas electorales y sistemas de partidos (Duverger, 1957).
- Las democracias tienden a no luchar entre sí (Maoz y Abdolali, 1989).
- Los ciudadanos con bajo estatus social (educación e ingreso) son más frecuentes entre aquellas personas políticamente inactivas (Verba y Nie, 1972).
- La confianza social se relaciona con las redes de compromiso cívico (Putnam, 1993).
- El tamaño de la población de un país se asocia con la existencia del federalismo (Lijphart, 1999).

Existen múltiples medidas de asociación con diversos cálculos y coeficientes, pero que resultan específicas para la relación entre los niveles de medición de las variables, según sean nominales, ordinales y de intervalo o de razón (ver capítulo 1). En este capítulo se estudiarán dos medidas de asociación muy comunes, la

²³ Aunque en ocasiones se le llama también correlación, se reservará este nombre para la medida, en particular, de correlación de Pearson.

primera para relacionar dos variables categóricas (nominales) y la segunda para dos variables continuas (de intervalo o de razón).

Prueba *chi* cuadrado para tablas de contingencia

Cuando se tienen variables categóricas (pueden ser medidas nominalmente u ordinalmente, pero sin poner atención al orden, es decir, interpretándolas como nominales), como pueden ser sexo (hombre/mujer), grupo etario (jóvenes/adultos/adultos mayores), participación electoral (votó/no votó), es posible arreglar los datos de dos variables en un único cuadro denominado tabla de contingencia.

Estas tablas deben ser exhaustivas y mutuamente excluyentes: cada observación (como personas encuestadas) debe pertenecer a una categoría de cada variable (exhaustividad) y solamente a una categoría (exclusividad mutua).

Por ejemplo, con los datos de la encuesta de noviembre de 2013 realizada por el CIEP, se construyó una tabla de contingencia al cruzar las variables categóricas sexo y simpatía partidaria (ver cuadro 5.1). En este caso, cada persona encuestada se ubica en una categoría de sexo y simpatía partidaria y a la vez no puede estar en más de una categoría. Puesto que cada variable posee dos categorías, se genera una tabla de dos filas y dos columnas (de forma abreviada, una tabla 2 x 2).

Cuadro 5.1. Sexo y simpatía partidaria (valores absolutos)

Sexo	Simpatiza con algún partido		Total
	Sí	No	
Hombres	80	189	269
Mujeres	104	262	366
Total	184	451	635

Fuente: CIEP (2012-2014).

La misma tabla puede verse en términos relativos o de porcentajes en tres sentidos. Primero, por el porcentaje de simpatía según sexo, es decir, por columnas (cuadro 5.2). Los datos de esta tabla se pueden interpretar descriptivamente de la siguiente manera: entre los simpatizantes de algún partido

hay 43.5% de hombres y 56.5% de mujeres, mientras que entre los no simpatizantes existe 41.9% de hombres y 58.1% de mujeres.

Cuadro 5.2. Sexo y simpatía partidaria (porcentajes por columnas)

Sexo	Simpatiza con algún partido		Total
	Sí	No	
Hombres	43.5%	41.9%	42.4%
Mujeres	56.5%	58.1%	57.6%
Total	100.0%	100.0%	100.0%

Fuente: CIEP (2012-2014).

Los porcentajes también se pueden ordenar por filas como en el cuadro 5.3. Mediante la lectura de la tabla de contingencia de sexo según simpatía partidaria, se encuentra que, en los hombres, 29.7% simpatiza con algún partido, mientras 70.3% no; entre las mujeres, 28.4% simpatiza con algún partido y 71.6% no lo hace.

Cuadro 5.3. Sexo y simpatía partidaria (porcentajes por filas)

Sexo	Simpatiza con algún partido		Total
	Sí	No	
Hombres	29.7%	70.3%	100.0%
Mujeres	28.4%	71.6%	100.0%
Total	29.0%	71.0%	100.0%

Fuente: CIEP (2012-2014).

Finalmente, es posible interpretar porcentajes en relación con el total de la muestra analizada (cuadro 5.4). Bajo esta lógica, se puede leer que un 12.6% del total de encuestados corresponden a hombres que simpatizan con algún partido, un 29.8% a hombres no simpatizantes, 16.4% a mujeres que simpatizan y 41.3% a mujeres que no simpatizan con ninguno.

Cuadro 5.4. Sexo y simpatía partidaria (porcentajes según total)

Sexo	Simpatiza con algún partido		Total
	Sí	No	
Hombres	12.6%	29.8%	42.4%
Mujeres	16.4%	41.3%	57.6%
Total	29.0%	71.0%	100.0%

Fuente: CIEP (2012-2014).

Como es común en el análisis estadístico, luego de observar los datos, se buscará alguna prueba que permita sostener inferencias sobre la asociación entre las variables, es decir, siguiendo el ejemplo anterior, ¿hay alguna relación entre el sexo y la simpatía por partidos políticos?

Para ello se presentará la prueba *chi* cuadrado de independencia, creada por el estadístico inglés Karl Pearson (1857-1946) y que corresponde a la prueba estadística más antigua aún en uso (Agresti y Franklin, 2012, p. 544).

Para introducir la prueba, en primer lugar, es necesario esclarecer qué se entiende por independencia. Existe independencia estadística entre dos variables categóricas cuando sus porcentajes, según alguna variable, son iguales para cualquier categoría de la otra variable. Por ejemplo, una tabla como la presente en el cuadro 5.5 indicaría independencia de la variable *Y* respecto a la variable *X*, ya que no importa la categoría de *X* (puede ser la 1 o la 2), el porcentaje según *Y* es igual.²⁴

Cuadro 5.5. Ejemplo de independencia entre variables

Variable <i>X</i>	Variable <i>Y</i>		Total
	Categoría 1	Categoría 2	
Categoría 1	120 (60%)	80 (40%)	200 (100%)
Categoría 2	90 (60%)	60 (40%)	150 (100%)
Total	210 (60%)	140 (40%)	350 (100%)

Fuente: elaboración propia.

²⁴ La conclusión de independencia sería la misma si se examinan los porcentajes por columnas.

Por lo tanto, con la prueba *chi* cuadrado se establecen las siguientes hipótesis:

- Hipótesis nula: las variables son estadísticamente independientes.
- Hipótesis alternativa: las variables no son estadísticamente independientes (*i. e.* son dependientes).

La prueba considera si una tabla hipotética con valores independientes es significativamente diferente de la tabla observada de datos reales. Si el valor p obtenido es menor al alfa establecido (*p. g.* 0.05), entonces se rechaza la hipótesis nula de independencia, es decir, existe una asociación entre ambas variables.

La lógica del cálculo es la siguiente. Retomando la teoría de probabilidades (ver Hernández, 2010), se 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. Por ejemplo, si se quiere saber cuál es la probabilidad de sacar un 2 en un dado de seis caras y luego un 5 (eventos que son independientes pues un resultado no se relaciona con otro), entonces se calcula:

$$\begin{aligned} &\text{probabilidad}_{\text{cara 2}} * \text{probabilidad}_{\text{cara 5}} = \\ &\frac{\# \text{ de caras}}{\text{total de caras}} * \frac{\# \text{ de caras}}{\text{total de caras}} = \\ &\frac{1}{6} * \frac{1}{6} = 0.03 \end{aligned}$$

Por analogía, con los datos del cuadro 5.1, si se piensa que las probabilidades de ser hombre y la probabilidad de simpatizar por un partido son independientes, entonces la probabilidad de la ocurrencia de ambas debe ser igual a su producto:

$$\begin{aligned} &\text{probabilidad}_{\text{hombres}} * \text{probabilidad}_{\text{simpatizantes}} = \\ &\frac{\# \text{ de hombres}}{\text{total de personas}} * \frac{\# \text{ de simpatizantes}}{\text{total de personas}} = \\ &\frac{184}{635} * \frac{269}{635} = 0.12. \end{aligned}$$

Esta sería, por lo tanto, la probabilidad si los eventos son independientes, la cual, al multiplicarse por el total de personas (635), permite calcular el llamado “valor

esperado”: 77.9, o 78 al redondear. Es decir, si fuesen independientes, debería haber 78 hombres con simpatías partidarias. Pero la frecuencia observada (real) es 80. Luego de obtener los valores esperados para cada celda (se muestran en el cuadro 5.6), bajo el supuesto hipotético de que son categorías independientes, se comparan las frecuencias reales y esperadas para el total k de celdas por medio de la siguiente fórmula que calcula el estadístico *chi* cuadrado (χ^2):

$$\chi^2 = \sum_{i=1}^k \frac{(\text{valor observado}_i - \text{valor esperado}_i)^2}{\text{valor esperado}_i}.$$

Si la diferencia entre observados y esperados es estadísticamente “grande” (mayor a lo esperado por el azar), entonces el *chi* cuadrado es relativamente grande y las variables no son independientes. Por el contrario, si los valores esperados son semejantes a los observados, el *chi* cuadrado calculado es relativamente pequeño y se creería que las variables son independientes.

Cuadro 5.6. Valores observados y esperados

Sexo	Simpatiza con algún partido			
	Sí		No	
	Valor observado	Valor esperado	Valor observado	Valor esperado
Hombres	80	77.9	189	191.1
Mujeres	104	106.1	262	259.9

Fuente: elaboración propia.

Con los datos del ejemplo, se calcula el siguiente estadístico *chi* cuadrado que resulta ser relativamente pequeño y, por ende, parece mostrar independencia:

$$\chi^2 = \frac{(80 - 77.9)^2}{77.9} + \frac{(104 - 106.1)^2}{106.1} + \frac{(189 - 191.1)^2}{191.1} + \frac{(262 - 259.9)^2}{259.9} = 0.132.$$

Para confirmar el resultado, se procede a calcular el valor p por medio de SPSS. Además, nótese que la prueba señala únicamente si son independientes, pero no la intensidad de la dependencia, para ello se puede recurrir al coeficiente V de Cramer que se detallará también en seguida.

Procedimiento para prueba *chi* cuadrado

Se recuerda que el propósito es analizar si las variables sexo y simpatía partidaria son independientes o no y cuán fuerte es la asociación, basándose en los datos de la encuesta de noviembre de 2013 del CIEP. Para comprobar lo anterior se sigue el siguiente procedimiento en SPSS.

Recuadro 5.1

*Resumen del procedimiento para la prueba *chi* cuadrado en SPSS*

Analizar → Estadísticos descriptivos → Tablas de contingencia

Trasladar las variables a Filas y Columnas.

En Estadísticos, seleccionar Chi cuadrado y Phi y V de Cramer. Continuar.

Aceptar.

En SPSS se busca en el menú Analizar, Estadísticos descriptivos y Tablas de contingencia, con lo cual se abre la ventana presentada en figura 5.1. Primero, se trasladan las variables hacia las casillas de filas y columnas. El orden es arbitrario y depende de cómo se quiera visualizar la tabla, pero los resultados son los mismos.

Luego, en Estadísticas (figura 5.2), se solicita el cálculo de *chi* cuadrado y el *Phi* y la *V* de Cramer para variables nominales (estos dos últimos se ubican en una misma opción). Después Continuar.

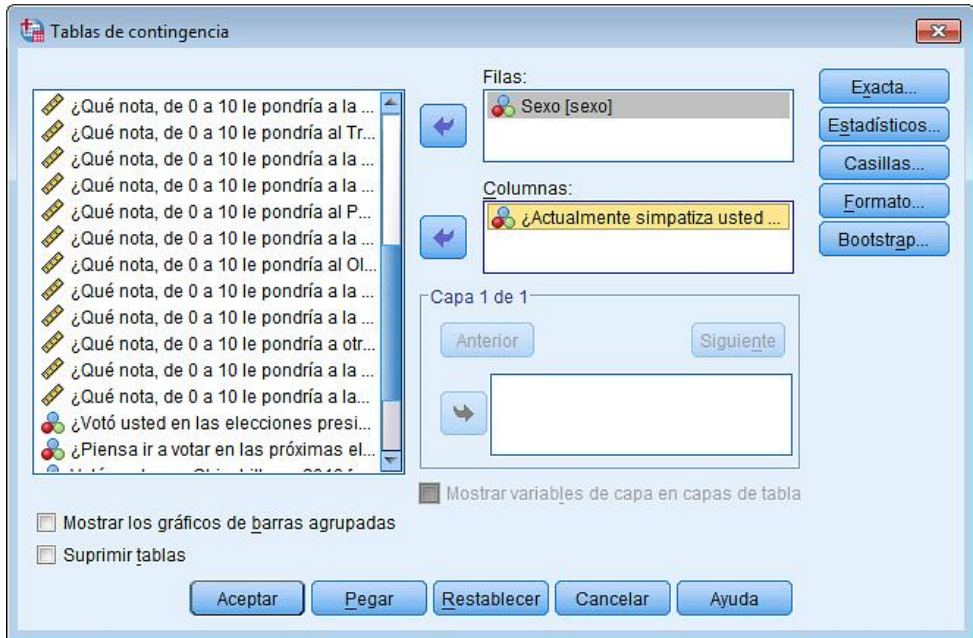


Figura 5.1. Ventana de tablas de contingencia en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

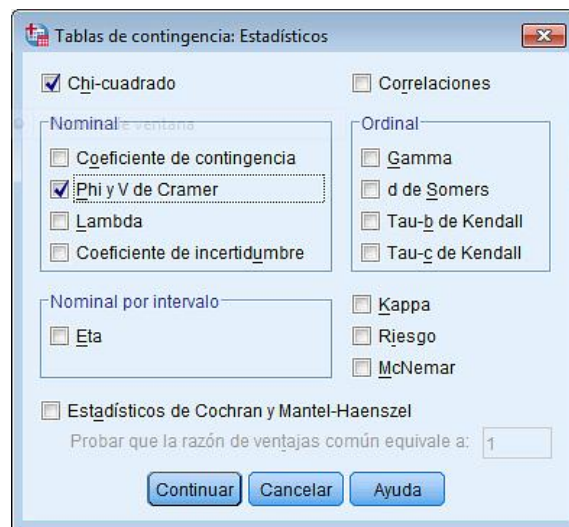


Figura 5.2. Ventana para definir estadísticas en tablas de contingencia en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

En Celdas (figura 5.3) pueden especificarse los porcentajes según columnas, filas o el total de la muestra. En este caso, se pedirá por filas para examinar cómo se distribuyen los hombres y mujeres dependiendo de si simpatizan o no.²⁵

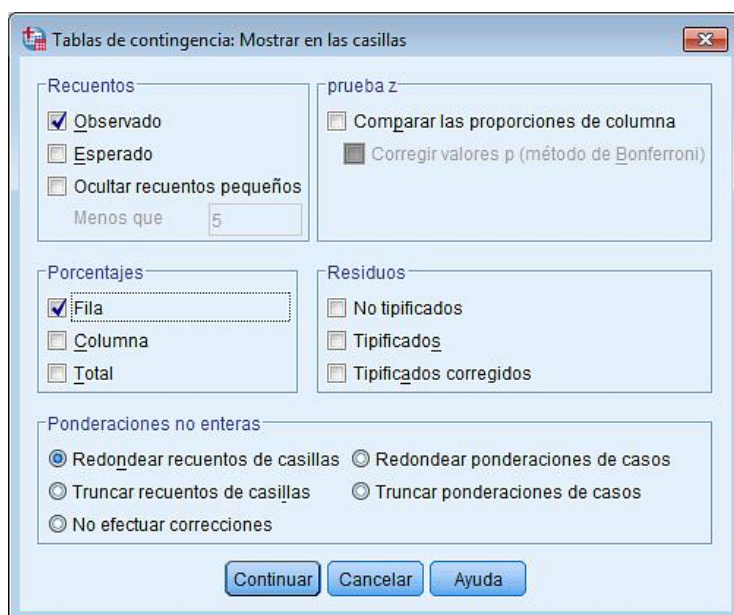


Figura 5.3. Ventana para definir celdas en tablas de contingencia en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

²⁵ También se puede solicitar el cálculo de los valores esperados tal y como se vio en el cuadro 5.6.

Al ejecutar el procedimiento anterior, se obtiene entre los resultados una tabla de contingencia que combina las categorías de sexo con simpatía (figura 5.4) igual a la presentada anteriormente en el Cuadro 5.3.

Tabla de contingencia sexo Sexo * simpatiza ¿Actualmente simpatiza usted con algún partido político?

			simpatiza ¿Actualmente simpatiza usted con algún partido político?		Total
			,00 no simpatiza	1,00 simpatiza	
sexo Sexo	,00 hombre	Recuento	189	80	269
		% dentro de sexo Sexo	70.3%	29.7%	100.0%
	1,00 mujer	Recuento	262	104	366
		% dentro de sexo Sexo	71.6%	28.4%	100.0%
Total		Recuento	451	184	635
		% dentro de sexo Sexo	71.0%	29.0%	100.0%

Figura 5.4. Resultado de tablas de contingencia en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	.132 ^a	1	.716	.724	.391
Corrección por continuidad ^b	.076	1	.783		
Razón de verosimilitudes	.132	1	.716		
Estadístico exacto de Fisher					
Asociación lineal por lineal	.132	1	.716		
N de casos válidos	635				

- a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 77,95.
- b. Calculado solo para una tabla de 2x2.

Figura 5.5. Resultado de la prueba *chi* cuadrado en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

También, se calcula el resultado de la prueba *chi* cuadrado de independencia en la primera fila de la siguiente tabla (figura 5.5). En ella, se indica el valor del estadístico *chi* cuadrado y su valor *p* (la significancia) que es 0.716: la probabilidad de equivocarse rechazando la hipótesis de independencia es del 72% y evidentemente superior a un nivel de significancia usual como el 5%. Por lo

tanto, se decide no rechazar la hipótesis nula de independencia, por lo que las variables sexo y simpatía partidaria no están relacionadas.

Si las variables estuvieran relacionadas, aún no se sabría la fuerza de la asociación. El coeficiente de Cramer aporta información en este sentido. Esta medida se calcula con base en el valor estadístico *chi* cuadrado (χ^2) que, en el ejemplo, es 0.132 (ver figura 5.5), de la siguiente manera (Acock y Stavig, 1979):

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

donde *n* es el tamaño de la muestra y se debe multiplicar por (filas – 1) o por (columnas – 1), dependiendo de cuál número sea el más pequeño o el mínimo. La medida varía entre 0 que significa que no hay relación y 1 cuando la asociación es perfecta. Para el ejemplo anterior, en el que no hay que buscar un mínimo ya que el número de filas es igual al de columnas, se calcula de la siguiente forma:

$$V \text{ de Cramer} = \sqrt{\frac{0.132}{635 * (2 - 1)}} = 0.014.$$

En SPSS (figura 5.6), el resultado es exactamente igual al cálculo manual (0.014), que es muy bajo y por lo tanto refleja escasa asociación, siendo este resultado consistente con lo obtenido en la prueba de independencia de *chi* cuadrado.

Medidas simétricas			
		Valor	Sig. aproximada
Nominal por nominal	Phi	.014	.716
	V de Cramer	.014	.716
N de casos válidos		635	

Figura 5.6. Resultado del coeficiente V de Cramer en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Correlación bivariada: coeficiente de correlación de Pearson

Para establecer la relación entre dos variables métricas o continuas, una de las medidas de asociación más utilizadas es el coeficiente de correlación lineal (abreviado por su símbolo r). Esta medida, que se ha convertido en una pieza de información fundamental en las investigaciones cuantitativas, es otro aporte de Pearson, el mismo creador de la prueba *chi* cuadrado vista previamente.

La correlación indica cuán fuerte es una relación lineal entre dos variables X y Y , esta se calcula con la covarianza entre las variables (que determina su dependencia lineal) dividida entre las desviaciones estándar de cada variable (s_X y s_Y), las que estandarizan o eliminan el efecto de escalas de medición de las variables. Es decir,

$$r_{XY} = \frac{\text{covarianza}_{XY}}{s_X * s_Y}.$$

El valor de la correlación puede oscilar entre -1 y 1 , por lo que reporta, además, una dirección: números positivos significan que cuanto mayor sea la X , mayor es la Y (y viceversa, a mayor Y mayor X), mientras que los números negativos indican una relación inversa, a mayor X menor Y (y viceversa).

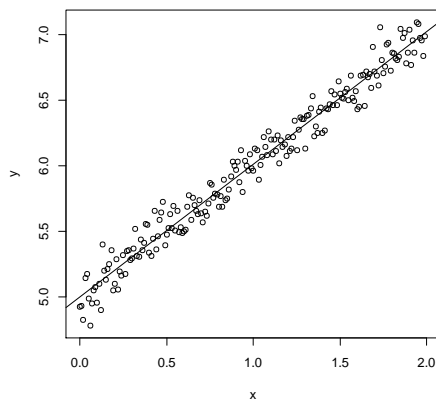
Por ejemplo, si se calcula la correlación entre edad en años cumplidos e ingreso anual promedio, una correlación mayor a 0 se interpretaría como que cuantos más años tenga una persona, mayor será su ingreso. Una correlación menor a 0 es lo opuesto: a mayor edad, menor ingreso. Los r iguales a 1 y -1 indican siempre correlación lineal perfecta. Si la correlación fuese 0 , entonces las variables no están correlacionadas (o la correlación es nula).

Ahora bien, surge la pregunta sobre cómo leer niveles intermedios de correlación como 0.7 , -0.5 , 0.2 , etc. Aunque muchas veces los libros de texto ofrecen reglas prácticas para su interpretación, la fuerza de la correlación depende más del área de estudio. Por ejemplo, en un campo donde se sabe que la relación entre dos variables es teóricamente fuerte y en una investigación se encuentra un r de 0.6 , esta correlación podría considerarse decepcionante y baja. Pero si tan solo se

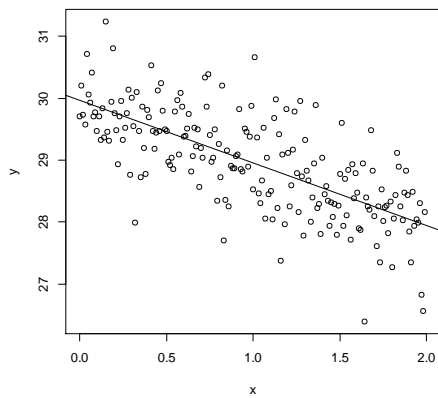
esperaba una tenue relación entre dos variables, se puede ser menos exigente y considerar un r de 0.5 como moderadamente alto.²⁶

En la figura 5.7, se presentan gráficos de dispersión entre dos variables X y Y que ejemplifican distintos niveles de correlación: en el gráfico superior izquierdo se muestran datos con una alta correlación positiva ($r = 0.98$); en el gráfico superior derecho se construyó con datos con una alta correlación negativa o inversa ($r = -0.70$); y en el gráfico inferior izquierdo, se ilustra una baja correlación positiva ($r = 0.28$). Por otra parte, se destaca que la correlación de Pearson se establece para asociación lineal entre variables, por lo que no aplica en relaciones curvilíneas o de otro tipo no lineal. Una aplicación sobre datos no lineales llevaría a conclusiones erróneas. Por ejemplo, en el gráfico inferior derecho de la figura anterior, se observa que el r de Pearson es igual a 0.96, lo que implica una correlación casi perfecta, aunque la relación es curvilínea, por lo que el cálculo de correlación lineal de Pearson es improcedente.

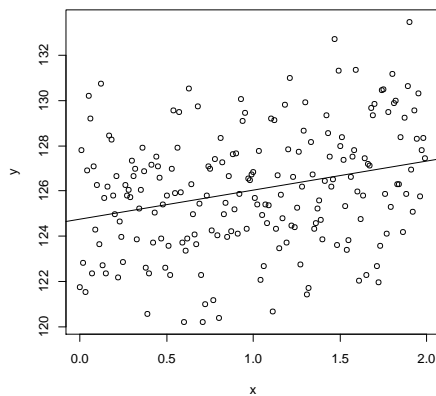
²⁶ La interpretación de valores intermedios aplica también para el coeficiente V de Cramer, la diferencia es que este último no indica dirección, pues siempre es positivo.



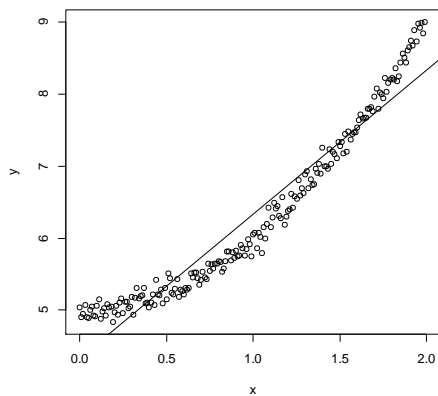
$$r = 0.98$$



$$r = -0.70$$



$$r = 0.28$$



$$r = 0.96$$

Figura 5.7. Gráficos de dispersión con distintos coeficientes de Pearson.

Fuente: elaboración propia.

Procedimiento para correlación de Pearson

Para ejemplificar el procedimiento, se recurre a la encuesta de noviembre de 2013 del CIEP en la que se preguntó por calificaciones de 0 a 10 para diversas instituciones. Supóngase que una investigadora piensa analizar la percepción sobre instituciones sin control partidario directo, como la Sala Constitucional (o Sala Cuarta) y el Tribunal Supremo de Elecciones (TSE). Se esperaría que una mayor calificación en la primera institución se relacione con otra mayor en la segunda (o al revés), por lo que la correlación de Pearson ofrecería una medición pertinente de la asociación.

Recuadro 5.2

Resumen del procedimiento para correlación de Pearson en SPSS

Analizar → Correlaciones → Bivariadas

Trasladar las variables de interés.

En Coeficientes de correlación seleccionar Pearson. En Prueba de significación dejar Bilateral.

Aceptar.

En el menú de Correlaciones Bivariadas en SPSS, se seleccionan ambas variables de la base de datos correspondiente. Se especifica la utilización del coeficiente de Pearson y la significancia Bilateral (figura 5.8).

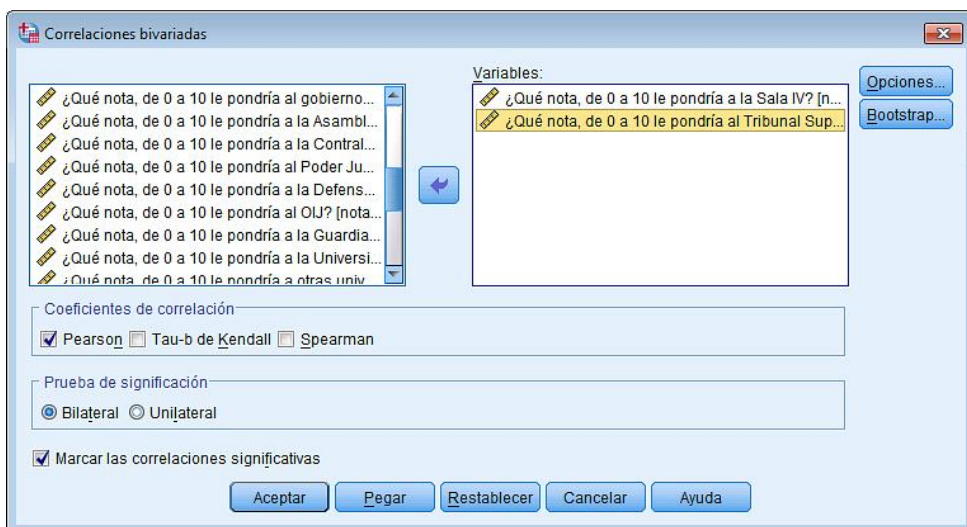


Figura 5.8. Ventana de correlaciones bivariadas en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Luego, se obtiene como resultado una matriz de correlaciones entre la calificación del TSE y la Sala Constitucional. El coeficiente de Pearson entre las dos instituciones es de 0.554, lo cual se puede interpretar como moderadamente alto, ya que el estudio es exploratorio y no se tenían muchas expectativas de antemano para la relación entre variables. Al ser r mayor a 0 o positivo, se puede constatar que la relación es positiva: un mayor puntaje al TSE se relaciona con un mayor puntaje a la Sala Cuarta.

Puede notarse que la matriz de correlación se caracteriza por ser simétrica: su diagonal actúa como un espejo entre dos esquinas. Así, la correlación entre TSE y Sala Cuarta se encuentra tanto en la esquina inferior izquierda como en la superior derecha. En la diagonal están las correlaciones de la cada variable con ella misma: lógicamente esta correlación es perfecta (igual a 1).

Correlaciones		
	notasalacuarta ¿Qué nota, de 0 a 10, le pondría a la Sala IV?	notatse ¿Qué nota, de 0 a 10, le pondría al Tribunal Supremo de Elecciones?
notasalacuarta ¿Qué nota, de 0 a 10, le pondría a la Sala IV?	Correlación de Pearson Sig. (bilateral) N	.554** .000 568
notatse ¿Qué nota, de 0 a 10, le pondría al Tribunal Supremo de Elecciones?	Correlación de Pearson Sig. (bilateral) N	.554** .000 553

** . La correlación es significativa al nivel 0,01 (bilateral).

Figura 5.9. Resultado de correlaciones bivariadas en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Además, se obtiene un nivel de significancia que indica la probabilidad de rechazar la hipótesis nula de que la correlación sea cero en la población. En estos casos, se puede rechazar la hipótesis nula con un nivel de significancia de 0.05 o 5%. Vale aclarar que, aunque es común encontrar el valor p para la correlación de Pearson, lo usual es centrarse en la magnitud de la correlación, pues si esta última es alta o si la muestra es grande, difícilmente no será significativa.

Comentarios finales

Existen más medidas de asociación de las que se abarcaron aquí, que fueron específicas para relaciones entre dos variables asumidas como nominales, por un lado, y entre dos variables métricas, por el otro. Para variables categóricas con escala ordinal, por ejemplo, es necesario aplicar otras coeficientes (consultar Agresti, 2007; Gutiérrez-Espeleta, 2010; Sánchez, 2005).

Recuérdese, además, que la prueba *chi* cuadrado y el coeficiente de correlación se utilizan en relaciones bivariadas, por lo que conllevan la debilidad de no controlar por otras posibles variables influyentes.

Ejercicios

1. La encuesta preelectoral de noviembre de 2013 del CIEP preguntó tanto si la persona piensa ir a votar en 2014 como si votó en las elecciones nacionales de 2010. Asumiendo una relación teórica del hábito del voto, se quiere analizar si el comportamiento del voto en 2010 está vinculado con el de 2014. Utilice la prueba *chi* cuadrado para determinar la independencia al 5% y el coeficiente V de Cramer para conocer la fuerza de la asociación. Además, interprete los resultados de manera sustantiva.
2. Emplee los datos de noviembre de 2013 del CIEP para analizar la correlación entre la edad de las personas encuestadas y las notas otorgadas a los candidatos presidenciales Johnny Araya, José María Villalta y Luis Guillermo Solís. Sugerencia: puede utilizar una única matriz de correlaciones para la interpretación.

CAPÍTULO 6

REGRESIÓN LINEAL SIMPLE POR MÍNIMOS CUADRADOS ORDINARIOS

Introducción

Sir Francis Galton (1822-1911), investigador en biometría y uno de los pioneros de la estadística moderna, descubrió un fenómeno interesante al comparar las estaturas de los padres y de los hijos: los hijos de padres muy altos tienden a ser más bajos que sus padres, mientras que los hijos de padres pequeños tienden a ser más altos que sus progenitores. A este hecho lo llamó “regresión a la media”. Si la regresión a la media no existiera, las poblaciones humanas tenderían a los extremos y la humanidad estaría compuesta por un grupo de personas altísimas y otro grupo de personas bajísimas (Salsburg, 2001, pp. 12-13).

La regresión caracteriza múltiples fenómenos de la naturaleza y del comportamiento, como el siguiente caso: un instructor de vuelo explica que no felicita a los pilotos cuando ejecutan correctamente una maniobra, ya que la mayoría de las veces, cuando el piloto repite la misma maniobra, la hace mal. En cambio, cuando les grita por equivocarse, suelen realizar bien la siguiente maniobra. En realidad, lo que actúa es la regresión y no la reacción del instructor: cuando el cadete realizó el intento exitoso en realidad tuvo suerte (independientemente de que lo felicitaran o no), mientras que su ejecución mala era seguida de una buena ejecución porque su habilidad lo hacía regresar a su promedio de éxito. Otros fenómenos que oscilan alrededor de la media son las notas obtenidas por estudiantes a lo largo de un curso, las ventas de un negocio, el desempeño de un golfista... en ocasiones, un buen resultado está seguido por uno malo y esto genera extrañeza; en realidad, el nivel real no está en estos extremos sino en su promedio (ejemplo tomados de Kahneman, 2012).

El análisis de regresión es un método poderoso y flexible que permite examinar relaciones entre todos los tipos de variables (categóricas y métricas). Por medio de la regresión es posible lograr varios objetivos investigativos: describir, explicar y predecir.

Pero la existencia de diversos tipos de variable dependiente (cualitativa o cuantitativa) y de su nivel de medición implica que se deba escoger entre diferentes modelos de regresión. En este texto, se estudiarán puntualmente dos de ellos:

- Regresión lineal (simple y múltiple) por mínimos cuadrados ordinarios (MCO o conocido por sus siglas en inglés como OLS, por *ordinary least squares*): aplica cuando la variable dependiente es métrica. También es llamado “modelo gaussiano” porque los errores asumen una distribución probabilística normal.
- Regresión logística: cuando la variable es categórica. Aunque existen diversos modelos, se verá solamente el caso de una categórica binaria (en el capítulo 8).

Conceptos

En los modelos de regresión se establece una relación causal o asociación entre una variable independiente –también llamada predictor, variable explicativa o covariable– denotada como X y una variable dependiente o respuesta Y . En general, el interés al aplicar los modelos de regresión radica en:

- Conocer el efecto promedio de la variable X sobre la Y (magnitud del efecto).
- Determinar si el efecto es estadísticamente significativo (es decir, no nulo).
- Establecer el poder predictivo o explicativo del modelo teórico.

Ahora bien, en la investigación cuantitativa no siempre se puede hablar de una causalidad en términos de condiciones necesarias y suficientes, sino de efectos de causas (Mahoney y Goertz, 2006). Es decir, no se intenta responder preguntas de qué causa un fenómeno o evento determinado, sino cuáles son los efectos sobre un fenómeno de una causa o un conjunto de ellas, las cuales provienen de una teorización previa que identificó los posibles factores causales o variables independientes.

En cuanto a los modelos de regresión con datos transversales, tampoco se puede comprobar la causalidad en sentido estricto, pues se toman mediciones recopiladas en un mismo periodo y la causalidad exige una variable independiente que preceda temporalmente a la variable dependiente (más adelante se ahondará este problema). Por ello, los datos longitudinales son preferibles para establecer causalidad, aunque no sean tan robustos como los experimentos controlados (Frees, 2004, p. 11). De modo que la relación entre X y Y se interpreta, finalmente, no como “ X causa Y ”, sino como “ X es un factor asociado a Y ” o “ X tiene efectos sobre Y ”.

Etapas

Se pueden establecer cinco etapas al realizar un análisis de regresión (de cualquier tipo):

- Especificación del modelo: se formula un modelo teórico que relaciona la(s) variable(s) X con la respuesta Y , según ciertos parámetros desconocidos. La relación asumida proviene de la teoría establecida o de una exploración previa de los datos.
- Estimación: se estima el valor de los parámetros según los datos que se tienen.
- Evaluación: se examina la adecuación del modelo (también conocida como la bondad de ajuste).
- Diagnósticos: se chequea que los supuestos del modelo estadístico se cumplan.
- Correcciones: en caso de incumplirse algún supuesto, se busca corregir.

El libro se centra en las tres primeras etapas: especificación, estimación y evaluación. Las dos faltantes deberían verse en un curso más amplio o estudiarse para el caso de una aplicación rigurosa (por ejemplo, una tesis).

Modelo

Regresión simple se refiere a que se tiene una única variable independiente o predictor. Cuando hay más de una variable independiente, se llama regresión múltiple.

En términos generales, el modelo es el siguiente:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

donde

- Y_i corresponde a los valores que puede adoptar la variable dependiente según los datos;
- X_i son los valores de la variable independiente según los datos;
- β_0 es el parámetro denominado constante o intercepto;
- β_1 es el parámetro denominado pendiente;
- ε_i son los errores (también llamados residuos) que se definen como la diferencia entre el valor de predicho (\hat{Y}_i) por el modelo y el valor observado (Y_i), por lo que $\varepsilon_i = Y_i - \hat{Y}_i$;
- i es cada observación (una persona encuestada, un cantón, un país, etc.).

El objetivo de la regresión es estimar los valores de los parámetros, también llamados coeficientes de regresión, β_0 y β_1 que hagan mínimos los errores (la diferencia entre valores observados y predichos). Esta estimación se realiza por el procedimiento de *mínimos cuadrados ordinarios*. Las fórmulas para calcular los coeficientes son relativamente simples (aunque con muchos datos es preferible recurrir a paquetes estadísticos para la estimación). 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 su parte, el intercepto se puede determinar con base en el estimador de la pendiente y las medias de las variables X y Y :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Sin embargo, retomando lo visto en el capítulo 2, ¿cómo saber que estas fórmulas de cálculo estiman correctamente los parámetros desconocidos? La teoría estadística (específicamente, el teorema Gauss-Markov) indica que si ciertos supuestos se cumplen, entonces las fórmulas anteriores permiten obtener los *mejores estimadores lineales insesgados*²⁷ (ver Gujarati y Porter, 2010).

En otras palabras, si para calcular un promedio desconocido se sabe que el mejor estimador es la media universalmente conocida (suma de los valores de las

²⁷ En la literatura en inglés, se conocen con el acrónimo BLUE por *best linear unbiased estimators*.

observaciones entre el número de observaciones), las fórmulas obtenidas por mínimos cuadrados ordinarios –bajo los supuestos de normalidad, linealidad y otros– conllevan a estimar, de manera insesgada y eficiente, los parámetros desconocidos β_0 y β_1 .

Ejemplo

Una investigadora supone que cuanto mayor sea el número de partidos políticos, más opciones tienen los electores y por ello se incentiva la concurrencia a las urnas. Es decir, formula la hipótesis siguiente: cuanto mayor sea el número de partidos políticos en el país, mayor es la participación electoral promedio.

Cuadro 6.1. Base de datos el modelo de regresión simple

País (año)	<i>NEPE</i>	<i>PARTIEDAD</i> (%)
Argentina (2007)	4.74	72.24
Bolivia (2005)	2.62	63.44
Chile (2005)	6.57	59.64
Colombia (2006)	8.59	44.15
Costa Rica (2006)	4.63	63.96
Ecuador (2006)	5.79	79.91
El Salvador (2009)	2.92	72.39
Guatemala (2007)	7.75	57.19
Honduras (2005)	2.65	60.55
México (2006)	3.59	63.26
Nicaragua (2006)	3.45	74.16
Panamá (2004)	4.47	80.31
Uruguay (2004)	2.61	93.14

Fuente: OIR (2014b) y Pignataro (2012).

Para medir el número de partidos en el país, se utiliza el indicador denominado número efectivo de partidos electorales (*NEPE*).²⁸ La participación se mide

²⁸ El número efectivo de partidos políticos es un índice propuesto por Laakso y Taagepera (1979) como una medida para tomar el tamaño relativo de los partidos según sus proporciones en votos o en escaños.

como el total de votos entre la población en edad de votar en porcentaje (*PARTIEDAD*). Los datos recopilados se presentan en el cuadro 6.1.

En primer lugar, se grafican los datos en un diagrama de dispersión (figura 6.1). Con una simple inspección gráfica, es posible notar una relación inversa o negativa: cuanto mayor es el número de partidos, menor la participación. Parecería que los datos tienden a refutar la hipótesis investigativa. Sin embargo, se utilizará el análisis de regresión para saber si la variable independiente es significativa estadísticamente, cuál es la magnitud y dirección de sus efectos y cuál es el poder predictivo del modelo.

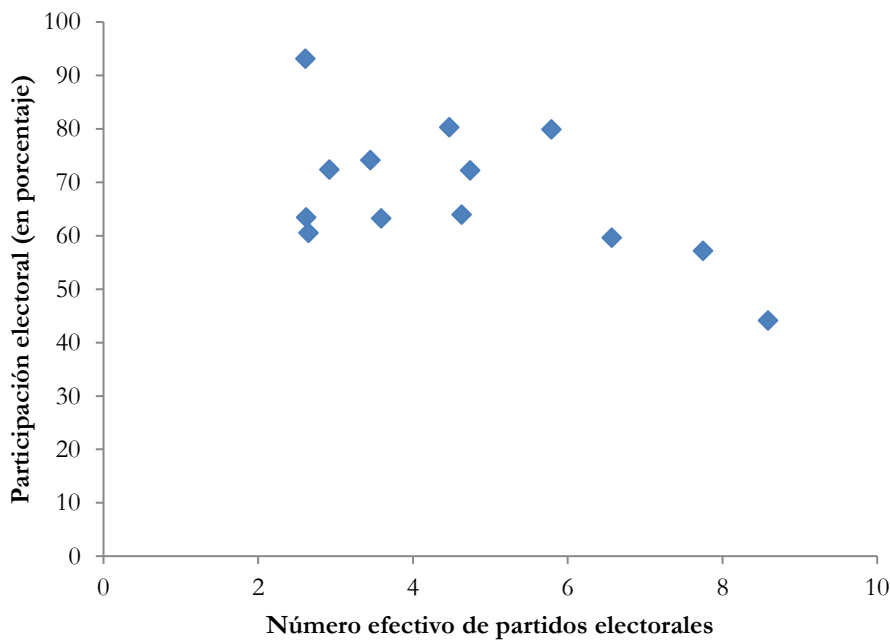


Figura 6.1. Gráfico de dispersión de número efectivo de partidos electorales por participación electoral.

Fuente: elaboración propia con base en OIR (2014b) y Pignataro (2012).

Especificación del modelo. Se propone el modelo de regresión:

$$PARTIEDAD_i = \beta_0 + \beta_1 NEPE_i + \varepsilon_i,$$

donde PARTIEDAD es la variable dependiente (Y) que se quiere explicar a partir de la variable independiente (X) NEPE. Los i corresponden a las observaciones o países; β_0 y β_1 son los parámetros desconocidos que se quieren estimar.

Primero, se verá una forma cómoda de calcularlos manualmente y luego en SPSS.

Estimación. Con las fórmulas antes vistas, es fácil calcular los estimadores de mínimos cuadrados ordinarios teniendo pocos datos y con una hoja de cálculo. En el cuadro 6.2, se muestran los datos originales de las variables X y Y (columnas 1 y 2). A estas observaciones se les resta su promedio en cada caso (columnas 3 y 4). Luego, las desviaciones del promedio para X y Y se multiplican entre sí (columna 5). Al sumar el resultado anterior, se obtiene el numerador de la fórmula de $\hat{\beta}_1$. El denominador consiste en elevar al cuadrado las desviaciones de X (columna 6) con su promedio (que ya se habían calculado en la columna 2) y sumar estos cuadrados.

Cuadro 6.2. Cálculo manual de los estimadores de mínimos cuadrados

X_i	Y_i	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
4.74	72.24	0.10	4.21	0.40	0.01
2.62	63.44	-2.02	-4.59	9.29	4.10
6.57	59.64	1.93	-8.39	-16.15	3.71
8.59	44.15	3.95	-23.88	-94.20	15.57
4.63	63.96	-0.01	-4.07	0.06	0.00
5.79	79.91	1.15	11.88	13.61	1.31
2.92	72.39	-1.72	4.36	-7.53	2.97
7.75	57.19	3.11	-10.84	-33.65	9.64
2.65	60.55	-1.99	-7.48	14.91	3.98
3.59	63.26	-1.05	-4.77	5.03	1.11
3.45	74.16	-1.19	6.13	-7.33	1.43
4.47	80.31	-0.17	12.28	-2.14	0.03
2.61	93.14	-2.03	25.11	-51.10	4.14
$\bar{X} = 4.64$	$\bar{Y} = 68.03$			Suma = -168.80	Suma = 48.00

Fuente: elaboración propia.

Se tiene entonces,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{-168.80}{48.0} = -3.52.$$

El intercepto, a su vez, se obtiene con el resultado anterior y los promedios de X y Y :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 68.03 - (-3.52) * (4.64) = 84.36.$$

Por lo tanto, con base en los coeficientes de regresión estimados, la ecuación de regresión en términos del valor promedio de la variable dependiente se puede escribir así:

$$\bar{Y} = 84.36 - 3.52\bar{X}.$$

Esto significa que el modelo se ajusta para un promedio de la participación, dados distintos valores en el número efectivo de partidos. Podrá observarse que desaparece el término del error, ya que en un modelo ajustado este es igual a cero.

Ahora bien, hay otros resultados que también interesan, como la significancia estadística de los coeficientes. En SPSS es posible obtener esta última, así como los estimadores de regresión. Claramente el procedimiento computacional es más eficiente que el manual, sin embargo, los resultados son iguales. Para la ejecución en SPSS, se deben seguir los siguientes pasos.

Recuadro 6.1

Resumen del procedimiento de regresión lineal en SPSS

Analizar → Regresión → Lineales

Trasladar la variable dependiente a la casilla con ese nombre (debe ser métrica); igualmente pasar la variable independiente. En Estadísticos, seleccionar Estimaciones, Intervalos de confianza y Ajuste del modelo.

Aceptar.

Al abrir la ventana de regresión lineal (figura 6.2), se seleccionan las variables de análisis, trasladando las variables independiente y dependiente donde corresponden.

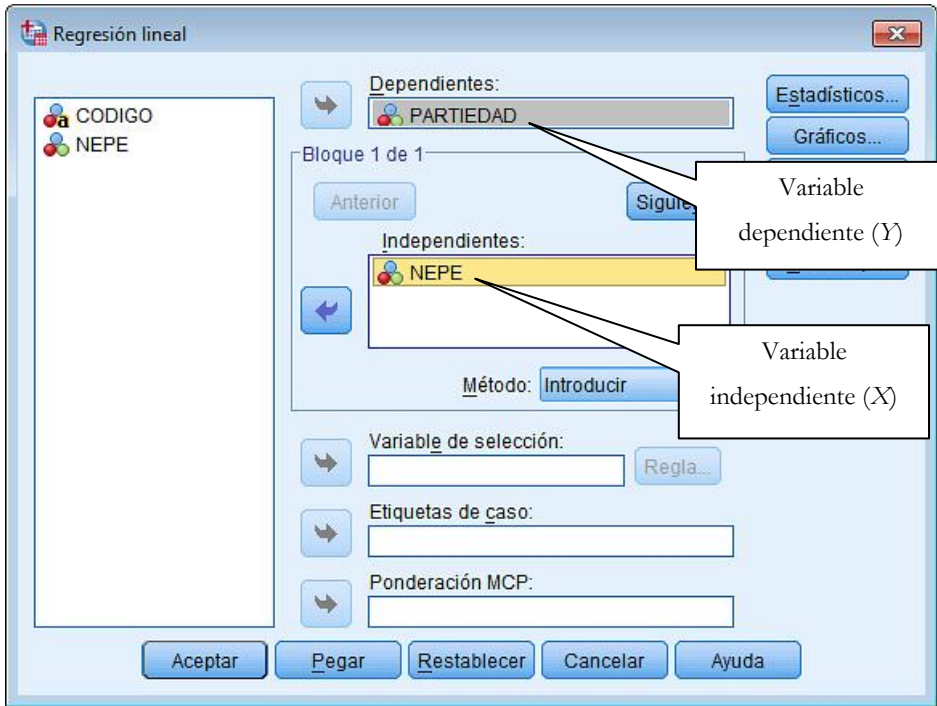


Figura 6.2. Ventana de regresión lineal en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

En Estadísticos, se seleccionan las opciones Estimaciones, Intervalos de confianza (se puede especificar el nivel) y Ajuste del modelo (figura 6.3). Luego, Continuar y Aceptar.

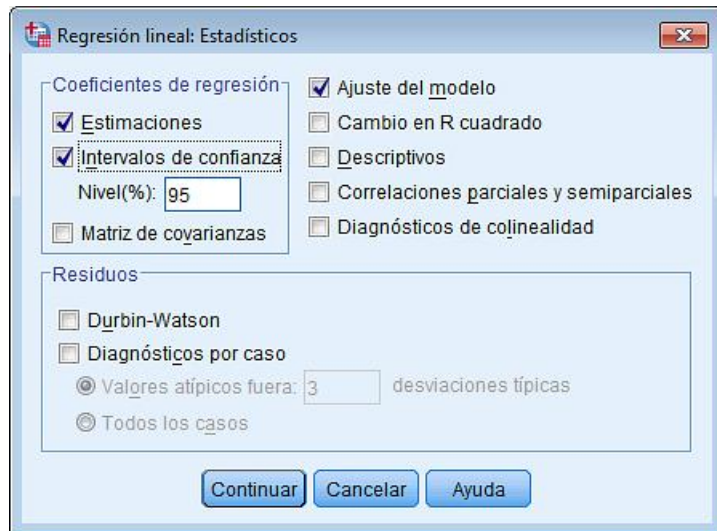


Figura 6.3. Ventana de opciones de estadísticos para regresión lineal en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

En la pantalla de resultados de SPSS, se obtienen las siguientes salidas mostradas en la figura 6.4.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.563 ^a	.317	.255	10.77727

a. Variables predictoras: (Constante), NEPE

ANOVA^b

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	593.599	1	593.599	5.111	.045 ^a
	Residual	1277.644	11	116.149		
	Total	1871.244	12			

a. Variables predictoras: (Constante), NEPE

b. Variable dependiente: PARTIEDAD

Coefficientes^a

Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
	B	Error típ.	Beta			Límite inferior	Límite superior
1 (Constante)	84.360	7.819		10.789	.000	67.150	101.569
NEPE	-3.517	1.556	-.563	-2.261	.045	-6.940	-.093

a. Variable dependiente: PARTIEDAD

Figura 6.4. Salida de regresión lineal simple en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Evaluación. Con las salidas anteriores, la investigadora obtiene los valores estimados insesgados y eficientes de los parámetros que minimizan el error. Estos son los betas señalados como coeficientes no estandarizados por el paquete SPSS. El $\hat{\beta}_0$ estimado es 84.360 y el $\hat{\beta}_1$ estimado es -3.517 (iguales a los calculados manualmente), por lo que puede establecer el siguiente modelo estimado, utilizando los nombres de las variables:

$$\overline{PARTIEDAD} = 84.360 - 3.517\overline{NEPE}.$$

También es posible expresarlo en términos de los valores ajustados o predichos para la variable dependiente con la siguiente simbología:

$$\widehat{PARTIEDAD}_i = 84.360 - 3.517\overline{NEPE}_i.$$

Ahora bien, ¿cómo se interpretan los resultados?

Primero, la constante o el intercepto ($\hat{\beta}_0$) es el valor promedio de Y cuando X es igual a cero porque

$$\widehat{PARTIEDAD} = 84.360 - 3.517(0) = 84.369.$$

Sin embargo, este coeficiente no siempre es interpretable (puede que no tenga sentido teórico o lógico). En el caso del ejemplo, se diría que la participación electoral promedio es del 84.36% cuando el número efectivo de partidos electorales es cero. Claramente, la competencia partidaria es necesaria en los sistemas democráticos y por ello la interpretación del intercepto no tiene sentido. Más importante resulta interpretar la pendiente.

La pendiente $\hat{\beta}_1$ significa el cambio en la variable dependiente promedio (\bar{Y}) cuando X aumenta en 1 unidad. Algebraicamente, el cambio de \bar{Y} es la diferencia o resta entre \bar{Y} con $X = 1$ y \bar{Y} con $X = 0$. Entonces,

$$\bar{Y}_{X=1} - \bar{Y}_{X=0} = [\beta_0 + \beta_1(1)] - [\beta_0 + \beta_1(0)]$$

$$\bar{Y}_{X=1} - \bar{Y}_{X=0} = \beta_0 + \beta_1 - \beta_0$$

$$\bar{Y}_{X=1} - \bar{Y}_{X=0} = \beta_1.$$

Es decir, el cambio en Y promedio cuando X aumenta una unidad es igual al valor del coeficiente β_1 .

En el ejemplo, el cambio en Y promedio es de 3.517 por cada unidad de X . Esto se traduce, en términos sustantivos, a que por cada partido efectivo electoral adicional, la participación promedio disminuye (porque el signo obtenido es negativo) 3.517 (que es coeficiente β_1) puntos porcentuales en promedio.

Se destaca que, al ser una relación lineal, el cambio es igual si se aumenta de 2 a 3 partidos, de 6 a 7 o de 8 a 9. También se puede calcular el cambio promedio al aumentar con dos partidos políticos adicionales: la participación disminuye $2 * -3.517 = -7.034$ puntos porcentuales en promedio.

Además de indicar magnitud (en términos de cambio promedio en Y), la pendiente o β_1 indica también dirección (al igual que el coeficiente de correlación

de Pearson). En este caso es negativa, por ello se habla de disminución; si fuera positiva, se interpretaría como aumento promedio en Y .

Por otro lado, la significancia estadística de los coeficientes es de gran importancia. La hipótesis nula que se establece es que el coeficiente de regresión es igual a cero. Si el valor p es menor al alfa establecido, el coeficiente es significativamente distinto de cero (se puede rechazar la hipótesis nula). En el ejemplo, $\hat{\beta}_1$ es significativo al 5%, pues su valor p es 0.045. Nótese, además, que el intervalo de confianza del coeficiente estimado no incluye al cero; por ello, se puede hablar de efectos estadísticamente significativos para la variable NEPE.

Finalmente, en la segunda tabla se indica el coeficiente de correlación (r) de Pearson (exactamente el mismo que se obtendría por el procedimiento visto en el capítulo anterior) y el coeficiente de determinación o R cuadrado (R^2). El coeficiente de determinación, evidentemente, es el cuadrado de la correlación ($0.563^2 = 0.317$); se interpreta como el porcentaje de variancia de la variable dependiente explicada por el modelo. En este caso, el modelo explica un 31.7% de la variación en la participación electoral.

Con todos estos resultados, la investigadora concluye que el número de partidos efectivos electorales produce efectos estadísticamente significativos sobre el porcentaje de participación electoral (significancia de la variable independiente); que con cada partido adicional la participación disminuye 3.5 puntos porcentuales *en promedio* (magnitud y dirección de la variable independiente) y que su modelo explica un 31.7% de la variación en la participación electoral (poder predictivo). Ahora bien, al igual que con otros coeficientes, la interpretación del R^2 depende del campo de conocimiento. Un 31.7% en el estudio comparado de participación electoral podría considerarse bajo, pero en otras áreas como en cultura política podría ser bastante alto.

Desde el punto de vista sustantivo, puede rechazar su hipótesis de que un mayor número de partidos políticos incentiva la participación, ya que el análisis de regresión muestra lo contrario: a mayor número de partidos, menor es la participación promedio.

Además, con base en su ecuación $\widehat{PARTIEDAD}_i = 84.360 - 3.517NEPE_i$, se puede graficar la recta de mejor ajuste para sus datos (figura 6.5).

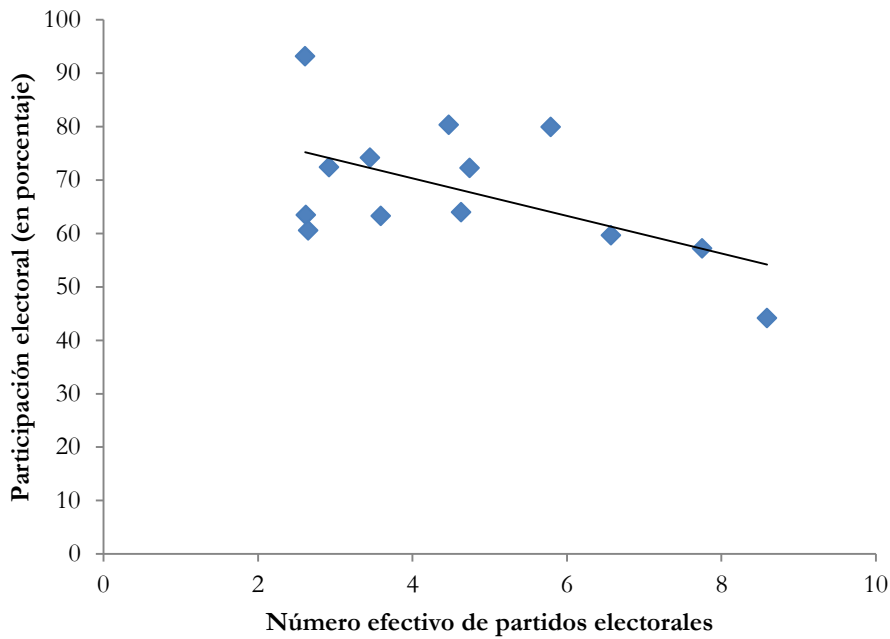


Figura 6.5. Recta de mejor ajuste según el modelo estimado.

Fuente: elaboración propia con base en OIR (2014b) y Pignataro (2012).

Comentarios finales

- Con la ecuación de regresión ajustada es posible predecir valores de participación electoral para datos del número de partidos efectivos que no existían originalmente. Por ejemplo, ¿cuál sería la participación esperada si hay nueve partidos políticos efectivos? Se calcularía: $\widehat{PARTIEDAD}_{X=9} = 84.360 - 3.517 * 9.0 = 52.71\%$.
- El caso simple –solo una variable independiente– se introduce primordialmente con fines didácticos. En la práctica, los fenómenos políticos (y muchos otros) tienen más de una variable explicativa, es decir, son multicausales, o bien las posibles causas se relacionan entre sí de manera que es difícil distinguir la “verdadera”. Para el ejemplo

anterior, ¿cómo saber si es el efecto del número de los partidos lo que se está observando y no un efecto del sistema electoral que se relaciona con el número de partidos y hace parecer que son estos últimos los que inciden en la participación? Para modelar más de una variable independiente y poder controlar estas explicaciones alternativas que podrían contener las verdaderas variables influyentes, se verá el modelo de regresión múltiple en el siguiente capítulo.

- Hay una estrecha relación entre el análisis de regresión y algunos contenidos que se vieron anteriormente: correlación y análisis de variancia. Esto no es casual, pues matemáticamente están vinculados. Véase, por ejemplo, que en las salidas de SPSS (figura 6.4) hay una tabla de ANOVA donde se indica una suma de cuadrados de regresión (en lugar de tratamiento). Si se divide la suma de cuadrados de regresión entre la suma de cuadrados total, se obtiene el coeficiente de determinación que, a la vez, es lo mismo que el coeficiente de correlación de Pearson al cuadrado:

$$R^2 = \frac{593.599}{1871.244} = 0.317 = 0.563^2.$$

- El modelo de regresión por mínimos cuadrados conlleva ciertos supuestos que deberían diagnosticarse para que los estimadores obtenidos en realidad sean los “mejores” (recordando el teorema Gauss-Markov). Principalmente, se refieren a que la relación entre las variables sea lineal, que los errores sigan una distribución normal y constante en su variabilidad y que no existan valores extremos.

Ejercicios

1. Se aduce teóricamente que la tasa de éxito de aprobación de leyes presentadas por el Ejecutivo es producto de la mayoría parlamentaria con la que cuenta el presidente. Compruebe esta hipótesis con un modelo de regresión utilizando los datos de García (2009) (cuadro 6.3), donde la variable dependiente corresponde a la tasa de éxito legislativo (*TASAEXITO*) y la variable independiente es el porcentaje total de legisladores del partido del gobierno (*LEGIS*). Estime el

modelo en SPSS, interprete el coeficiente de regresión, su significancia y el coeficiente de determinación. Concluya sustantivamente respecto a la hipótesis.

Cuadro 6.3. Mayoría parlamentaria y éxito legislativo

Gobierno	<i>LEGIS (%)</i>	<i>TASAEXITO (%)</i>
Néstor Kirchner (2003-2007)	53.6	68.8
Lula da Silva (2002-2006)	45.3	79.7
Ricardo Lagos (2000-2006)	47.5	72.9
Álvaro Uribe (2002-2006)	63.0	73.4
Abel Pacheco (2002-2006)	33.3	30.9
Lucio Gutiérrez (2003-2005)	32.0	30.4
Manuel Zelaya (2006-2009)	48.0	78.8
Vicente Fox (2000-2006)	36.4	74.1
Mireya Moscoso (1999-2004)	39.4	77.5
Luis González Macchi (1998-2002)	54.8	65.0
Alejandro Toledo (2001-2005)	46.7	74.3
Jorge Battle (2000-2005)	55.2	70.3
Jaime Lusinchi (1984-1989)	59.8	89.5

Fuente: García (2009).

2. Con las fórmulas presentadas en el capítulo, obtenga manualmente o en una hoja de cálculo los coeficientes de regresión con los datos anteriores y confirme su solución con el resultado en SPSS.

CAPÍTULO 7

REGRESIÓN LINEAL MÚLTIPLE POR MÍNIMOS CUADRADOS ORDINARIOS

Introducción

Previamente, se estudió la regresión simple como una forma de modelar una relación lineal en la que una variable independiente X produce efectos o está asociada con una variable dependiente Y . El objetivo era encontrar una ecuación lineal que se ajustara de la mejor forma a los datos que se poseen para saber si el efecto de la X es significativo, su dirección, magnitud y poder explicativo o predictivo.

Sin embargo, en la ciencia política, como en muchas otras disciplinas, difícilmente se puede suponer que un solo factor sea el responsable de un fenómeno particular, es decir, que sea su única causa. A su vez, los experimentos, que permiten controlar por el diseño todas las fuentes de variación, no están siempre disponibles o son impracticables. Por este motivo, es lógico suponer que hay más de un factor causal o múltiples variables independientes, las cuales se querrán modelar.

Al asumir un modelo multicausal con una variable dependiente métrica, es posible estimar un modelo de regresión a través de los mínimos cuadrados ordinarios. Aunque 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), su especificación, estimación y evaluación incluyen ciertos aspectos particulares a los que se debe atender particularmente y se tratarán al final del capítulo.

Modelo

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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i,$$

donde Y_i corresponde a los valores que puede adoptar la variable dependiente; X_{ki} son los valores de las variables independientes para cada observación; β_k constituyen los parámetros por estimar; ε_i son los errores (también llamados residuos) que se definen como la diferencia entre el valor de predicho (\hat{Y}_i) por el modelo y el valor observado (Y_i); i es cada observación (una persona encuestada, un cantón, un país, etc.); k es el número de variables independientes.

Nótense los siguientes aspectos:

- Los puntos suspensivos indican que se pueden incorporar muchas variables independientes (el límite de su número se tratará luego).
- Cada variable independiente X_k tiene un coeficiente a su lado, es decir, su relación o pendiente con la variable Y . Además, se incluye un intercepto o constante β_0 .
- El modelo es aditivo: los efectos de las variables independientes se suman.
- Los efectos son directos: cada variable independiente afecta directamente la dependiente (no a través de otra independiente).

Muchas de estas particularidades se pueden variar y, por ejemplo, especificar modelos que incluyan efectos multiplicativos de variables independientes llamadas “interacciones”. Este libro se centrará en el modelo aditivo presentado (para otras especificaciones posibles del modelo, ver Gujarati y Porter, 2010, capítulo 13).

Ejemplo

A un equipo de investigadores le interesa detectar si algunas características sociodemográficas están asociadas con la evaluación o aprobación que hacen las personas del gobierno, o sea, si personas con diferente edad, sexo, nivel educativo, estado de ocupación y simpatía partidaria a su vez califican de distintas maneras al gobierno. Para fortuna del equipo, cuentan con los datos de la encuesta de

noviembre de 2013 realizada por el Centro de Investigación y Estudios Políticos (CIEP), de tal modo no tendrán que recolectar ellos mismos la información.

En el cuadro 7.1 se resumen las variables independientes (factores sociodemográficos) de interés, junto con la variable de simpatía partidaria como variable de control y la variable dependiente (evaluación gubernamental).

Cuadro 7.1. Tabla de codificación

Nombre de la variable	Código en la base de datos	Significado y valores
<i>Dependiente</i>		
Evaluación gubernamental	GOBIERNO	Nota del 0 (muy mala) a 100 (muy buena)*
<i>Independientes</i>		
Edad	EDAD	Edad en años cumplidos
Sexo	SEXO	1=mujer 0=hombre
Estado de ocupación	OCUPADO	1=ocupado 0=desempleado/inactivo
Nivel educativo	EDUCACION	Escala de 0 (ninguna educación formal) a 6 (educación universitaria completa)
Simpatía partidaria	SIMPATIZA	1=sí 0=no

*Nota: la escala con que se preguntó era de 0 a 10, pero se transformó de 0 a 100 para facilitar la interpretación de los resultados del análisis de regresión.

Fuente: elaboración propia con base en CIEP (2012-2014).

En la tabla de codificación, se incluye el código con que se denominan las variables y las escalas con las que se midieron. Para interpretar resultados en un análisis de regresión, es muy importante tener claro cómo se midió cada variable en una base de datos. Obsérvese que las variables evaluación gubernamental, edad y nivel educativo tienen una medición métrica, mientras que sexo, estado de ocupación y simpatía partidaria son categóricas de tipo binario o dicotómico (es decir, dos categorías). Para la regresión, las variables categóricas se deben

codificar con 0 y 1, donde la asignación es arbitraria (p. g. mujer puede ser 0 o 1), pero explicitando siempre el significado atribuido.²⁹

Especificación del modelo. Con base en estas variables y sus mediciones particulares en la encuesta, se propone estimar el siguiente modelo por regresión:

$$GOBIERNO_i = \beta_0 + \beta_1 EDAD_i + \beta_2 SEXO_i + \beta_3 OCUPADO_i \\ + \beta_4 EDUCACION_i + \beta_5 SIMPATIZA_i + \varepsilon_i.$$

Como la variable dependiente (evaluación gubernamental) se mide de forma métrica, entonces se puede estimar por mínimos cuadrados ordinarios. Obsérvese que se tiene un intercepto (β_0) y cinco pendientes (una para cada variable independiente). Es decir, se tiene un modelo multidimensional que ya no se puede dibujar en un gráfico de dispersión de dos ejes y donde el cálculo de los estimadores es más engorroso en comparación con el modelo de regresión simple con una variable independiente. En lugar de fórmulas sencillas, encontrar los estimadores implica el uso de álgebra matricial; por lo tanto, para estimar el modelo se utilizará directamente el siguiente procedimiento en SPSS.

²⁹ Pueden incluirse variables categóricas de más de dos categorías recurriendo a las llamadas variables indicadoras o *dummy*. Aunque estas no se consideran en este texto, el procedimiento es simple: para cada categoría se crea una variable de 0 y 1, pero se deben incluir solo $k - 1$ variables indicadoras en el modelo. Por ejemplo, si la variable ocupado estuviera compuesta por tres categorías “ocupado”, “desempleado” y “pensionado”, se pueden construir tres indicadoras codificadas de la siguiente manera:

ocupado (1=ocupado y 0=otros casos);
desempleado (1=desempleado y 0=otros casos);
pensionado (1=pensionado y 0=otros casos).

En el modelo se incluyen solamente dos de las anteriores porque hay tres categorías; la que se deja por fuera es llamada la categoría de referencia, y la interpretación de las variables indicadoras incluidas se realiza respecto a esta última.

Recuadro 7.1

Resumen del procedimiento de regresión lineal múltiple en SPSS

Analizar → Regresión → Lineales

Trasladar la variable dependiente a la casilla con ese nombre (debe ser métrica); igualmente, pasar las variables independientes.

En Estadísticos, seleccionar Estimaciones, Intervalos de confianza y Ajuste del modelo.

Aceptar.

Estimación. Primero se trasladan las variables a sus respectivas casillas (dependientes e independientes). Antes de utilizar este comando de regresión, es importante haber explorado previamente las variables con frecuencias o descriptivos, declarar los valores perdidos y efectuar las recodificaciones necesarias (por ejemplo, en 0 y 1 para las categóricas).

Luego se seleccionan, en la ventana Estadísticos (figura 7.2), las opciones Estimaciones, Intervalos de confianza y Ajuste del modelo. Luego, Continuar y Aceptar.

La salida del paquete se muestra en la figura 7.3.

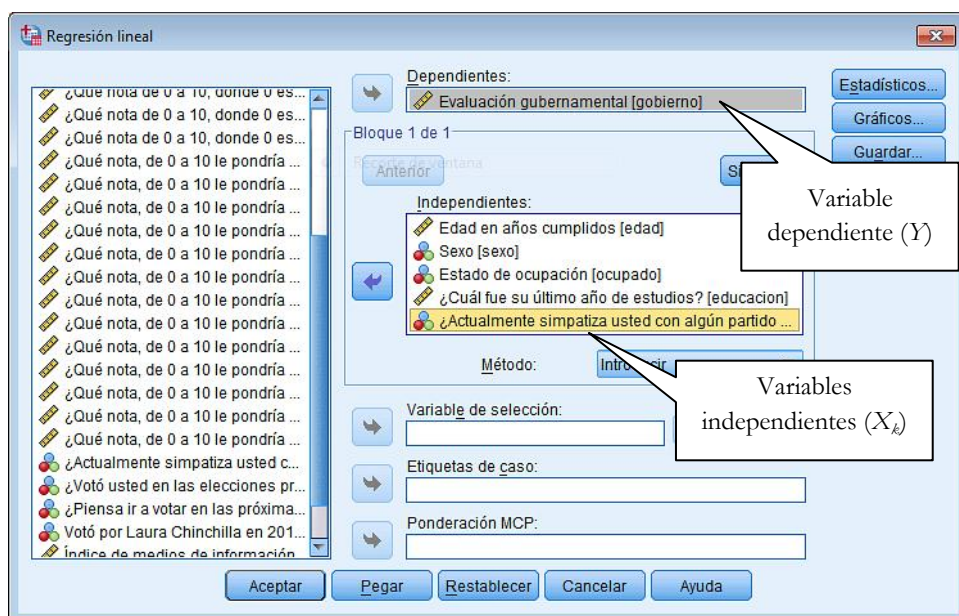


Figura 7.1. Ventana de regresión lineal en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

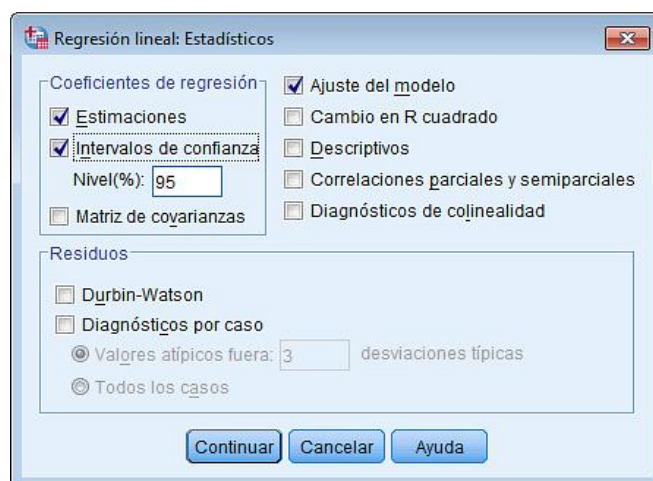


Figura 7.2. Ventana estadística de regresión lineal en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.213 ^a	.045	.037	24.19843

a. Variables predictoras: (Constante), simpatiza ¿Actualmente simpatiza usted con algún partido político?, sexo Sexo, edad Edad en años cumplidos, educacion ¿Cuál fue su último año de estudios?, ocupado Estado de ocupación

ANOVA^b

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	17201.497	5	3440.299	5.875	.000 ^a
	Residual	363635.185	621	585.564		
	Total	380836.683	626			

a. Variables predictoras: (Constante), simpatiza ¿Actualmente simpatiza usted con algún partido político?, sexo Sexo, edad Edad en años cumplidos, educacion ¿Cuál fue su último año de estudios?, ocupado Estado de ocupación

b. Variable dependiente: gobierno Evaluación gubernamental

Coefficientes^a

Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
	B	Error típ.				Límite inferior	Límite superior
1(Constante)	44.047	4.302		10.240	.000	35.599	52.494
edad	.164	.061	.112	2.710	.007	.045	.283
sexo	1.434	2.043	.029	.702	.483	-2.578	5.446
ocupado	-.746	2.160	-.015	-.345	.730	-4.988	3.496
educacion	-1.446	.631	-.099	-2.292	.022	-2.685	-.207
simpatiza	5.710	2.138	.105	2.671	.008	1.511	9.908

a. Variable dependiente: gobierno Evaluación gubernamental

Figura 7.3. Salida de regresión lineal múltiple en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Evaluación. Una vez estimado el modelo mediante SPSS, el equipo de investigación procede a analizar los resultados y evaluar el ajuste del modelo utilizando un nivel de significancia (α) del 5%. En el cuadro 7.2, se resumen los datos de interés provenientes de la salida anterior (figura 7.3).

Cuadro 7.2. Resultados del modelo de regresión sobre evaluación gubernamental

Variable	Coefficiente	Error estándar	Significancia
Constante	44.047	4.302	0.000
Edad	0.164	0.061	0.007
Sexo	1.434	2.043	0.483
Estado de ocupación	-0.746	2.160	0.730
Nivel educativo	-1.446	0.631	0.022
Simpatía partidaria	5.710	2.138	0.008
R ²	0.045		
R ² ajustado	0.037		

Fuente: elaboración propia con base en los resultados obtenidos en la figura 7.3.

A partir del modelo de regresión múltiple que se estimó, se puede concluir lo siguiente:

- Se encuentran tres variables significativas al 5% (porque sus valores p son menores a 0.05) que son edad, simpatía partidaria y nivel educativo. Los coeficientes de sexo y estado de ocupación no son significativamente distintas de cero (sus valores p son mayores a 0.05).
- Por cada año de edad cumplida, la nota de aprobación presidencial aumenta 0.164 en promedio, con todas las demás variables constantes.³⁰ En palabras simples, cuanto mayor es la persona encuestada, mejor valora al gobierno. No obstante, el efecto de cada año de edad es pequeño, pero también se puede interpretar (por la linealidad) como cambios por 10 años de edad: al incrementarse 10 años en la edad de la persona, la nota promedio de aprobación presidencial aumenta 1.64 en promedio, con todas las variables constantes.
- Respecto al nivel educativo, por cada nivel de educación alcanzado, la nota promedio otorgada al gobierno disminuye 1.446 puntos, con todas las otras variables constantes. Por lo tanto, cuanto mayor es el grado educativo, peor se valora la gestión gubernamental.

³⁰ Esta aclaración, “con todas las demás variables constantes”, indica que se está describiendo el cambio promedio en Y dependiendo solo de la variable en cuestión y dejando fijas las demás variables independientes. En economía, se suele utilizar para ello la expresión *ceteris paribus* (Wooldridge, 2010, p. 3).

- Como simpatía partidaria es una variable categórica, su interpretación es distinta a la de variables métricas como edad y nivel educativo: entre las personas con simpatía partidaria, la nota de aprobación presidencial es 5.710 puntos mayor en promedio respecto a los no simpatizantes, con todas las variables constantes. Es decir, quienes no simpatizan con ningún partido político, califican peor al gobierno.
- Sexo y estado de ocupación también presentan coeficientes, pero puede considerarse inadecuado darles alguna interpretación, ya que el valor p indica que no son estadísticamente diferentes de cero. Así que debe evitarse hablar de efectos positivos o negativos cuando no sean significativas las variables; más bien, el efecto es *nulo*.
- El intercepto o constante se lee de la siguiente forma: cuando todas las variables independientes tienen valor cero —es decir, entre hombres, con cero años de edad (lo cual no tiene sentido sustantivo), desempleados, sin estudios y sin simpatía partidaria— la nota promedio hacia el gobierno es de 44.047 puntos.
- Si se desea identificar cuál de las variables tiene más peso, es necesario leer los coeficientes estandarizados o tipificados (ver la columna respectiva en la figura 7.3). Un mayor coeficiente estandarizado, en valor absoluto, indica un mayor peso o importancia. En el ejemplo, edad, simpatía y educación son las variables con coeficientes tipificados mayores y, por ende, las de más peso en el modelo. Como se vio antes, también son estadísticamente significativas.
- Para evaluar la calidad del modelo (también llamada la bondad de ajuste), se tienen dos medidas: el “ R^2 ” y el “ R^2 ajustado” (o “ R cuadrado corregida” en SPSS). Ambos se interpretan como proporción, que se puede transformar en porcentaje al multiplicar por 100, de variancia explicada de la variable dependiente. Pero el segundo, el R^2 ajustado, en el caso de regresión múltiple resulta una medida más adecuada, pues corrige por el número de variables independientes incluidas. De esta forma, se puede decir que el modelo explica un 4.5% de la variabilidad de la nota de aprobación presidencial (según el R^2) o un 3.7% (según el R^2 ajustado).
- Se puede concluir que los factores sociodemográficos que influyen en la valoración del gobierno son edad y nivel educativo. Ser hombre o mujer

y ocupado o desempleado no tiene incidencia en la nota promedio; pero, en general, el poder predictivo del modelo con estas variables es muy bajo. Hay factores relevantes que no se incluyeron y que explicarían la variabilidad de las notas hacia el gobierno.

- Finalmente, se puede expresar el modelo estimado de la siguiente manera:

$$\widehat{GOBIERNO}_i = 44.047 + 0.164EDAD_i + 1.434SEXO_i - 0.746OCUPADO_i - 1.446EDUCACION_i + 5.710SIMPATIZA_i.$$

- Con esta ecuación ajustada es posible, además, predecir valores para nuevas observaciones, al igual que en el caso de regresión simple.

Problemas comunes en regresión múltiple

¿Cuántas variables independientes incluir? En el ejemplo mostrado se incluyeron cinco variables independientes para 635 observaciones (personas participantes en la encuesta). Aunque el anterior constituía una investigación hipotética, en situaciones reales el número de variables necesarias debe provenir de la teoría o las teorías que sustenten un modelo. Existen métodos que automatizan la selección de las “mejores” variables (es decir, según su significancia), aunque se desaconseja este tipo de operación en los casos probatorios de teoría en ciencia política.

En general, el número de variables no debe superar al número de observaciones, ya que se indeterminaría el modelo y no podría ser estimado. Por ejemplo, si se tiene una ecuación matemática con una única incógnita, es posible despejar esta última y conocer el valor:

$$10 = 4a + 2.$$

En ese caso, $a = 2$. Pero si se plantea la siguiente ecuación:

$$10 = 4a + 2b.$$

las soluciones (los valores de a y b) son infinitas. Para determinarlos se requiere de un sistema de dos ecuaciones. En regresión, cada ecuación se plantea para una observación i y las incógnitas son los coeficientes de regresión. Por lo tanto, el

número de parámetros por estimar (incógnitas) no puede ser mayor al número de observaciones (ecuaciones).

El dilema metodológico radica en que, por un lado, se deben incluir todas las variables necesarias para explicar un fenómeno según el marco teórico, pero, por otro lado, se debe conservar la parsimonia o simplicidad del modelo (lo que algunos llaman, su elegancia), evitando incorporar variables irrelevantes (es decir, que no aportan nada a la explicación).

En ocasiones (aunque difícilmente en el campo de datos de encuestas), el número de casos es bastante limitado; esto ocurre, por ejemplo, en estudios comparados entre países u otras unidades políticas (81 cantones en Costa Rica, 18 países en América Latina, etc.). En estos diseños es fácil sobrecargar el modelo con demasiadas variables, lo cual se denomina “sobreespecificarlo”. Las reglas prácticas sugieren no más de una variable independiente por cada 8 o 10 observaciones.

Sesgo de variable omitida. Si bien los modelos intentan no sobredeterminar el diseño y ser parsimoniosos, existe un peligro mayor al especificar un modelo: el sesgo de la variable omitida. Esto ocurre cuando se atribuye un efecto de una variable independiente X_1 sobre una 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 establecería una relación espuria o falsa entre X_1 y Y .

Por ejemplo, supóngase que los investigadores consideran que la única variable que influye en la evaluación gubernamental es si la persona trabaja o no; por ello, establecen un modelo simple donde la variable independiente es estado de ocupación y encuentran efectos significativos al 5% (figura 7.4).

Sin embargo, el modelo de regresión múltiple ya visto indicaba que la variable estado de ocupación no es significativa *al controlar las demás variables que pueden influir y que están relacionadas con ella*. Asumir un modelo simple donde la ocupación tiene efectos significativos es una relación espuria y conlleva un sesgo de variable omitida. Por ejemplo, el nivel educativo está relacionado con el estado de ocupación y las personas empleadas tienen mayor nivel educativo alcanzado promedio; así que el supuesto efecto negativo del estado de ocupación en

realidad refleja el nivel educativo que es significativo y con signo negativo en el modelo múltiple.

Coeficientes ^a							
Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
	B	Error típ.	Beta			Límite inferior	Límite superior
1 (Constante)	50.494	1.326		38.079	.000	47.890	53.098
ocupado	-4.240	1.974	-.086	-2.148	.032	-8.116	-.364

a. Variable dependiente: gobierno Evaluación gubernamental

Figura 7.4. Modelo de regresión con solo la variable estado de ocupación.

Fuente: elaboración propia con base en el paquete SPSS.

Obsérvese que solo con incluir educación, el efecto del estado de ocupación ya desaparece (figura 7.5).

Coeficientes ^a							
Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
	B	Error típ.	Beta			Límite inferior	Límite superior
1 (Constante)	55.862	2.237		24.971	.000	51.468	60.255
ocupado	-2.189	2.080	-.044	-1.052	.293	-6.272	1.895
educacion	-1.821	.613	-.125	-2.969	.003	-3.025	-.617

a. Variable dependiente: gobierno Evaluación gubernamental

Figura 7.5. Modelo de regresión con las variables estado de ocupación y nivel educativo.

Fuente: elaboración propia con base en el paquete SPSS.

Multicolinealidad. Es común que las variables independientes presenten cierta correlación entre sí, por ejemplo: la educación y el ingreso individual comúnmente están asociados. Cuando las variables independientes están *fuertemente* correlacionadas se genera un problema denominado como multicolinealidad. Esto puede afectar las estimaciones, agrandando los errores estándar de los coeficientes que podrían parecer iguales a cero cuando no lo son (Gujarati y Porter, 2010).

Una forma simple de examinar si existe multicolinealidad es construyendo una matriz de correlaciones (figura 7.6).³¹

Correlaciones						
		edad	sexo	ocupado	educacion	simpatiza
edad	Correlación de Pearson	1	.063	-.195**	-.310**	.047
	Sig. (bilateral)		.112	.000	.000	.233
	N	635	635	635	635	635
sexo	Correlación de Pearson	.063	1	-.286**	-.085	-.014
	Sig. (bilateral)	.112		.000	.031	.717
	N	635	635	635	635	635
ocupado	Correlación de Pearson	-.195**	-.286**	1	.334*	-.072
	Sig. (bilateral)	.000	.000		.000	.070
	N	635	635	635	635	635
educacion	Correlación de Pearson	-.310**	-.085	.334*	1	.021
	Sig. (bilateral)	.000	.031	.000		.604
	N	635	635	635	635	635
simpatiza	Correlación de Pearson	.047	-.014	-.072	.021	1
	Sig. (bilateral)	.233	.717	.070	.604	
	N	635	635	635	635	635

** . La correlación es significativa al nivel 0,01 (bilateral).

* . La correlación es significante al nivel 0,05 (bilateral).

Figura 7.6. Matriz de correlaciones para examinar multicolinealidad.

Fuente: elaboración propia con base en el paquete SPSS.

Para el modelo de regresión múltiple del ejemplo no se encuentran correlaciones suficientemente fuertes para dar indicios de multicolinealidad (las correlaciones entre variables categóricas y categóricas con métricas deben leerse con cautela y, preferiblemente, ver su asociación por métodos más pertinentes como tablas de contingencia y pruebas χ^2). Otro síntoma de multicolinealidad es la presencia de un coeficiente de determinación (R^2) muy alto pero con coeficientes de regresión no significativos (Gujarati y Porter, 2010, p. 337).

Si bien la multicolinealidad es un problema grave y común, existen estrategias para resolverla. Lo primero es constatar que dos variables no estén midiendo lo

³¹ En SPSS se puede construir la propia matriz con la opción de Correlaciones Bivariadas (capítulo 5), o bien en la ventana de Estadísticos en Regresión Lineales, seleccionando la opción Descriptivos.

mismo. Por ejemplo, si se incluyen como variables independientes el número de partidos políticos y un índice de fragmentación partidaria, que en el fondo también está calculando cuántos partidos hay, se está partiendo de una doble e innecesaria medición que podría ocasionar problemas de multicolinealidad.

Otra solución, pero más sofisticada, es el análisis factorial (capítulo 10), pues tiene como objetivo reducir el número de variables en factores no correlacionados entre sí.

Endogeneidad. La regresión supone un modelo teórico donde las variables X_k producen efectos sobre la Y (es decir, las variables independientes preceden temporalmente a las dependientes). Pero, ¿qué pasa si la Y también puede generar efectos sobre alguna X ? Esto se le llama causalidad recíproca o endogeneidad, es decir, que existe una doble secuencia causal.³² Algunos ejemplos son los siguientes:

- Si se analizan los efectos de los sistemas electorales sobre los partidos políticos (siguiendo la agenda de Duverger, 1957), se puede afrontar endogeneidad, pues se argumenta que los partidos, como actores políticos y legislativos, buscan a su vez alterar las reglas electorales (Benoit, 2004).
- Un típico caso de endogeneidad se encuentra en el estudio del desarrollo económico y la democracia. Desde las teorías de la modernización, se puede argumentar que países con crecimiento económico son más proclives a ser democráticos (Lipset, 1959). Pero también se sostiene que las democracias garantizan un mejor funcionamiento de la economía (Przeworski *et al.*, 2000).
- Para el ejemplo del presente capítulo, la variable independiente simpatía partidaria podría ser endógena respecto a la evaluación gubernamental, en tanto una valoración positiva de esta última tiene la posibilidad de generar simpatía con el partido en el gobierno (o una evaluación negativa, desafección con los partidos).

³² Estrictamente, “endogeneidad” se refiere a la correlación entre las variables explicativas y los errores. Una de las fuentes de dicha correlación es la causalidad recíproca, pero hay otras (ver Wooldridge, 2010, pp. 54-55).

La endogeneidad es uno de los problemas más frecuentes en la política comparada (Franzese, 2007; Przeworski, 2007). Aunque las soluciones estadísticas pueden alcanzar niveles muy sofisticados, una estrategia sencilla consiste en utilizar variables independientes rezagadas temporalmente. Por ejemplo, en el caso anterior de crecimiento económico y democracia, para disminuir la endogeneidad se puede medir el PIB siempre en un año anterior respecto al del índice de democracia y evitar que el tipo de régimen incida en el desempeño económico.

Sin embargo, con datos transversales, donde no es posible rezagar variables porque hay una única medición en el tiempo, lo anterior no aplica. Un remedio consiste en buscar una variable alternativa —llamada instrumento— que esté parcialmente correlacionada con aquella que es endógena, pero que sea exógena respecto a la dependiente. Este es el enfoque de variables instrumentales en econometría (ver Wooldridge, 2010, capítulo 5).

Comentarios finales

- Al igual que con muchos otros métodos estadísticos, la regresión está orientada hacia la prueba de teoría previamente desarrollada por otras metodologías; por ello, es esencial fundamentar teóricamente las variables que se incluyan para evitar encontrar efectos significativos pero espurios. Además, no solamente se agregan variables de la teoría que se prueba, sino también de teorías o explicaciones rivales.
- Aunque el ejemplo no formalizó hipótesis, la teoría puede tener predicciones no solo de que ciertas variables son significativas, sino también la dirección de sus efectos. Por ejemplo, es posible que alguien haya argumentado que las personas de mayor edad son más críticas del gobierno por su “desencanto” con el sistema político (en términos técnicos, esperaba un coeficiente negativo para la variable edad). En este caso, se refutaría la hipótesis, ya que con la edad aumenta la aprobación.
- Se vio que el coeficiente de determinación (R^2) indica el porcentaje de variancia explicada, por lo que vale preguntarse, ¿qué pasa con la variancia no explicada? Ante ello existen dos posiciones: (a) la de aquellos que afirman la existencia de componentes aleatorios en el mundo por los

que nunca se podrán hacer predicciones perfectas (que expliquen el 100% de la variancia); (b) la de quienes ven el mundo de manera determinista y el 100% se podría explicar si se incluyeran todas las correctas variables explicativas (King, Keohane y Verba, 1994, p. 59).

- Cuando se ajusta un modelo y se encuentran algunas variables que no son significativas, es posible estimar de nuevo el modelo sin estas variables. Con ello se obtendrían distintos coeficientes y un R^2 más pequeño. Por lo tanto, se alcanza mayor parsimonia pero con menor porcentaje de explicación.
- Al tratar con regresión múltiple, inevitablemente se cuenta con muchas variables y, por lo general, cada una tiene su propia unidad y forma de medición. De tal forma, resulta importante construir la tabla de codificación (como la del cuadro 7.1). También es clarificador proporcionar un esquema gráfico o incluso la ecuación de regresión para mostrar el modelo teórico que se busca probar.
- Los problemas de sesgo de variable omitida, multicolinealidad y endogeneidad no son exclusivos del modelo lineal por mínimos cuadrados ordinarios, aplican también en otros, como la regresión logística del capítulo 8.
- Una aplicación rigurosa del modelo debería examinar con detalle los supuestos sobre la distribución de los residuos, así como los de especificación que ya se explicaron.

Ejercicios

1. En la base de datos de la encuesta de noviembre de 2013 del CIEP, la variable poderes consiste en un índice de 0 a 100 construido con base en las calificaciones otorgadas al gobierno, a la Asamblea Legislativa y al Poder Judicial. Estime un modelo de regresión múltiple con las siguientes variables independientes: edad, sexo, estado de ocupación, nivel educativo, si simpatiza con algún partido político y si votó por Laura Chinchilla (Partido Liberación Nacional) en 2010.
2. Según el modelo estimado en el ejercicio 1, interprete los coeficientes de regresión y su significancia, así como el coeficiente de determinación.

3. Discuta si pueden darse problemas de sesgo de variable omitida, multicolinealidad y endogeneidad en el modelo estimado en el punto 1.

Cuadro 7.3. Codificación de variables

Variable	Código	Significado y valores
<i>Dependiente</i>		
Índice de poderes	PODERES	Nota del 0 (muy mala) a 100 (muy buena)
<i>Independientes</i>		
Edad	EDAD	Edad en años cumplidos
Sexo	SEXO	1=mujer 0=hombre
Estado de ocupación	OCUPADO	1=ocupado 0=desempleado/inactivo
Nivel educativo	EDUCACION	Escala de 0 (ninguna educación formal) a 6 (educación universitaria completa)
Simpatía partidaria	SIMPATIZA	1=sí 0=no
Votó por Laura Chinchilla	VOTOLAURA2010	1=sí 0=no

Fuente: elaboración propia con base en CIEP (2012-2014).

CAPÍTULO 8

REGRESIÓN LOGÍSTICA

Introducción

La regresión lineal estimada por mínimos cuadrados ordinarios (de los capítulos 6 y 7) se prefiere en casos donde la variable dependiente Y se ha medido de forma métrica o continua. Cuando la variable dependiente se mide de forma categórica, el modelo resulta inapropiado.

Uno de los modelos utilizados para variables dependientes cualitativas es el logístico, muy común en ciencia política, sociología y epidemiología. Se presentará una breve introducción al modelo logístico para el caso binario, dando énfasis a la interpretación de los resultados.

El interés con el modelo logístico es explicar el comportamiento de una variable dependiente Y categórica con base en una o muchas variables independientes X_k . En el caso binario o dicotómico, la variable dependiente asume dos valores: 1 y 0.

Estos son algunos ejemplos de variables dicotómicas que se quieren explicar y en los cuales se aplican modelos logísticos en investigaciones:

- Participación electoral: votó o se abstuvo (Ramírez, 2010).
- Condición de pobreza: sí o no (Mora y Pérez, 2009).
- Simpatiza con un partido político: sí o no (Norris, 2004).
- Referéndum: apoyo o rechazo al Tratado de Libre Comercio (Vargas-Cullell y Rosero Bixby, 2006; Treminio, 2010).
- Apoyo bipartidario o apoyo dividido entre partidos a leyes en el Congreso (Souva y Rhode, 2007).
- Momento de decisión del voto: temprano o tardío (Catellani y Alberici, 2012).

La codificación de las categorías con 1 y 0 en cierto modo es arbitraria, lo importante es identificar en la pregunta de investigación qué es lo que se busca explicar y asignar a esta categoría el 1. Por ejemplo, si se busca determinar los factores que impulsan a una persona a votar, es preferible, para el modelo logístico, codificar a los votantes con 1 y a los abstencionistas con 0. Por el contrario, si el interés radicara en los abstencionistas, entonces se le asignaría a los no votantes el 1. Por otro lado, las variables independientes pueden ser tanto categóricas como métricas.

Modelo

Si la variable dependiente Y está codificada con 0 y 1, es fácil observar cómo se complica el ajuste de una línea recta, como en los modelos normales por mínimos cuadrados ordinarios (ver figura 8.1).

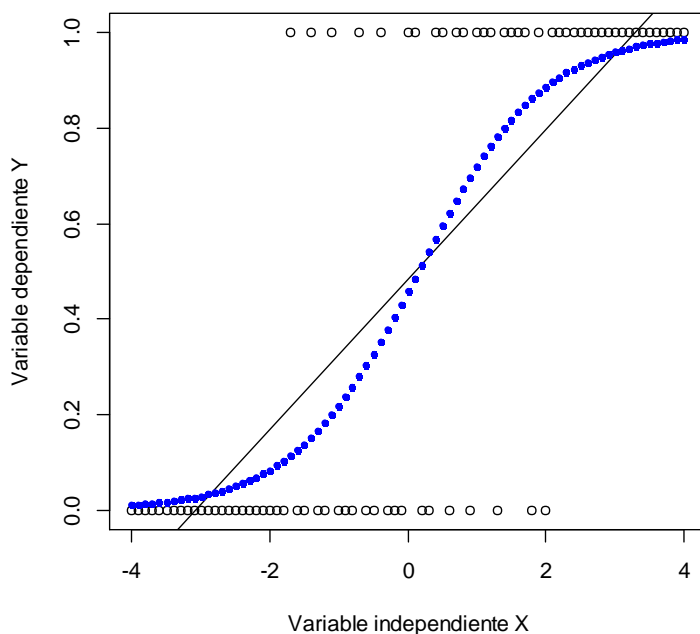


Figura 8.1. Ilustración del ajuste lineal por mínimos cuadrados y de la curva logística.

Fuente: elaboración propia.

Sin embargo, una curva en forma de “S” (la línea punteada azul de la figura 8.1) permite adecuarse mejor a la naturaleza categórica de la variable dependiente y acercarse más a los valores de la variable dependiente.

El modelo logístico se propone como una solución sumamente útil para ajustar la curva en forma de “S” por sus propiedades matemáticas y la interpretación sustantiva que permite (Hosmer, Lemeshaw y Sturdivant, 2013, p. 7).

Al igual que con el modelo normal, se tiene una ecuación lineal de la siguiente forma:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k,$$

donde X_k son los valores de las variables independientes para cada observación; β_k son los parámetros por estimar; k es el número de variables independientes. Hasta el momento no hay nada nuevo. Lo que se hace ahora es utilizar una función matemática que produzca una curva como la que se visualiza en la figura 8.1. Dicha función es la siguiente:

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

donde e corresponde al número irracional que es aproximadamente 2.718.

Esta función genera valores entre 0 y 1, por lo que se pueden predecir probabilidades de pertenencia a las dos categorías de la variable dependiente. En términos de ocurrencia del fenómeno codificado (por ejemplo, que una persona sea un votante, donde votar es 1 y no votar 0), se plantea la siguiente expresión:

$$P(Y = 1) = \frac{e^z}{1 + e^z},$$

donde $P(Y = 1)$ denota la probabilidad de votar. Sustituyendo z por la ecuación de regresión, entonces se tiene:

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k}}.$$

Con esta ecuación, el modelo logístico predice probabilidades de pertenencia a las categorías 1 y 0. Al igual que en regresión múltiple normal, se deben estimar los

parámetros β_k , pero no se hace por mínimos cuadrados ordinarios, sino por el método de máxima verosimilitud.

Un aspecto importante por aclarar es que la regresión logística es un modelo lineal en sus parámetros desde el esquema de los modelos lineales generalizados (Dobson y Barnett, 2008; ver el apéndice B de este libro) y gracias a la transformación logito. Se puede demostrar que al aplicar la transformación logito

$$\ln \left[\frac{P(Y = 1)}{1 - P(Y = 1)} \right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k,$$

se llega a una ecuación similar (excepto por el término error) a la del modelo gaussiano múltiple ya visto anteriormente (capítulo 7).

Ejemplo

Se quiere identificar si algunas variables sociodemográficas están relacionadas con el apoyo al reconocimiento legal de parejas formadas por personas del mismo sexo. Para ello se cuenta con datos de la encuesta de noviembre de 2013 del CIEP, según la cual un 29.3% está de acuerdo con el reconocimiento, un 64.4% en desacuerdo y un 6.3% no sabe o no respondió la pregunta.

Para aplicar el modelo de regresión logística binaria, se recodifica la variable original en una nueva, donde 1 corresponde a apoyar las parejas (31.3%) y 0 rechazo al reconocimiento (68.7%). Los “no sabe o no responde” se declararon como valores perdidos.

El modelo que se quiere estimar incluye, como variables independientes, la edad, el sexo, el estado de ocupación, el nivel educativo y la nota con la que califica a la iglesia católica (cuadro 8.1).

Cuadro 8.1. Tabla de codificación

Nombre de la variable	Código en la base de datos	Significado y valores
<i>Dependiente</i>		
Reconocimiento legal de parejas de personas del mismo sexo	APOYOPAREJAS	1=apoya el reconocimiento legal de las parejas de personas del mismo sexo 0=no lo apoya
<i>Independientes</i>		
Edad	EDAD	Edad en años cumplidos
Sexo	SEXO	1=mujer 0=hombre
Estado de ocupación	OCUPADO	1=ocupado 0=desempleado/inactivo
Nivel educativo	EDUCACION	Escala de 0 (ninguna educación formal) a 6 (educación universitaria completa)
Nota iglesia católica	NOTAIGLESIACAT	Escala de 0 (peor) a 10 (mejor)

Fuente: elaboración propia con base en CIEP (2012-2014).

Especificación. Siguiendo los conceptos introducidos en la sección anterior, el modelo por estimar es el siguiente:

$$P(Y = 1) = \frac{e^z}{1 + e^z},$$

donde

$$z = \beta_0 + \beta_1 EDAD + \beta_2 SEXO + \beta_3 OCUPADO + \beta_4 EDUCACION + \beta_5 NOTAIGLESIACAT.$$

Sustituyendo z , lo anterior es igual a la siguiente expresión:

$$P(Y = 1) = \frac{e^{\beta_0 + \beta_1 EDAD + \beta_2 SEXO + \beta_3 OCUPADO + \beta_4 EDUCACION + \beta_5 NOTAIGLESIACAT}}{1 + e^{\beta_0 + \beta_1 EDAD + \beta_2 SEXO + \beta_3 OCUPADO + \beta_4 EDUCACION + \beta_5 NOTAIGLESIACAT}}.$$

Estimación. Existen dos procedimientos para estimar el modelo en SPSS dependiendo de los paquetes disponibles. Si la versión de SPSS está completa o al menos incluye el módulo de regresión, entonces se siguen estos pasos.

Recuadro 8.1

Resumen del procedimiento convencional de regresión logística en SPSS

Analizar → Regresión → Logística binaria

Trasladar las variables.

En Guardar, seleccionar Probabilidades y Grupo de pertenencia. Continuar.

Aceptar.

En la ventana de regresión logística (figura 8.2), se trasladan las variables independientes (covariables) y dependiente, previamente revisadas, con valores perdidos declarados y las categóricas codificadas en 0 y 1. En la opción de Guardar se selecciona Probabilidades y Grupo de pertenencia. Luego, Continuar y Aceptar.

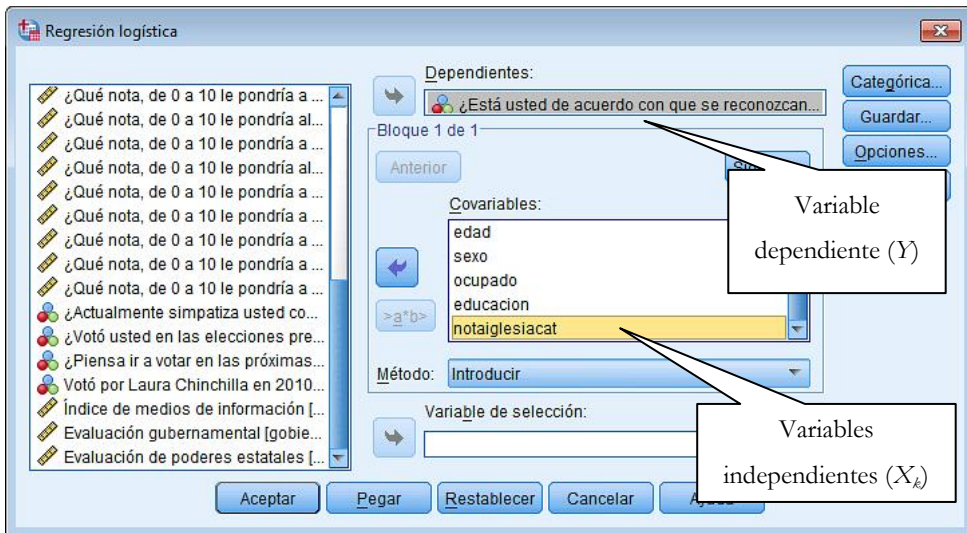


Figura 8.2. Ventana de regresión logística en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

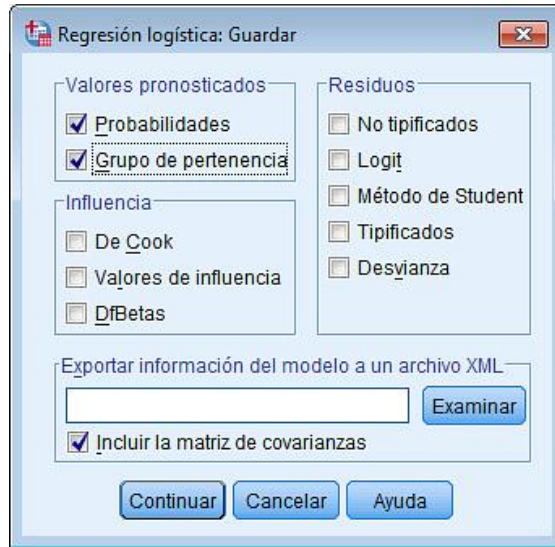


Figura 8.3. Ventana de opciones para guardar de regresión logística en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

En la figura 8.4 se presentan las salidas más importantes del procedimiento. En estas se incluyen, en la segunda columna de la segunda tabla, los coeficientes de regresión estimados del modelo y su significancia en la penúltima columna. La significancia se interpreta igual que en el modelo normal: un valor p menor a la probabilidad del error tipo 1 aceptable (p. g. 0.05) indica que el coeficiente es significativamente distinto de cero y la variable produce efectos no nulos.

Los coeficientes de regresión no se pueden interpretar directamente, pero si se exponencian con la base e , entonces sí adquieren sentido. Exponenciar significa llevar el número e a la potencia del coeficiente. Por ejemplo, para el caso de edad, su beta es -0.015 ; exponenciado es $e^{-0.015}$ que es 0.985. Los coeficientes exponenciados los presenta automáticamente SPSS en la última columna.

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	640.668 ^a	.123	.173

a. La estimación ha finalizado en el número de iteración 4 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Tabla de clasificación^a

Observado		Pronosticado		
		¿Está usted de acuerdo con que se reconozcan legalmente a las parejas formadas por personas del mismo sexo?		Porcentaje correcto
		,00 no	1,00 sí	
Paso 1	¿Está usted de acuerdo con que se reconozcan legalmente a las parejas formadas por personas del mismo sexo?			
	,00 no	359	38	90.4
	1,00 sí	121	59	32.8
	Porcentaje global			72.4

a. El valor de corte es ,500

Variables en la ecuación

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 ^a						
edad	-.015	.006	5.953	1	.015	.985
sexo	-.064	.201	.102	1	.749	.938
ocupado	-.012	.210	.003	1	.955	.988
educacion	.365	.065	31.814	1	.000	1.441
notaiglesiaca	-.103	.035	8.510	1	.004	.902
Constante	-.713	.479	2.220	1	.136	.490

a. Variable(s) introducida(s) en el paso 1: edad, sexo, ocupado, educacion, notaiglesiaca.

Figura 8.4. Salidas de la regresión logística en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Finalmente, la primera tabla indica cómo se clasificaron los casos. Por filas, se indica el total de personas que apoyan o no las parejas (reales en los datos). Por columnas, el número de apoyo o rechazo predicho por el modelo; cada caso se clasifica según la probabilidad predicha y el punto de corte preprogramado es 0.5 (pero este se puede variar). Es decir, si para una persona el modelo predice una

probabilidad menor a 0.5, entonces lo clasifica como no apoyo (0); si fuese mayor o igual a 0.5, lo categoriza como apoyo (1).³³

apoyoparejas	gobierno	poderes	PRE_1	PGR_1
1.00	60.00	56.67	.51334	1.00
.	80.00	26.67	.19986	.00
.00	80.00	80.00	.12369	.00
1.00	70.00	63.33	.21329	.00
.00	40.00	50.00	.62767	1.00
1.00	50.00	53.33	.58499	1.00
.00	.00	20.00	.33794	.00
1.00	60.00	63.33	.55700	1.00
1.00	40.00	40.00	.61846	1.00

Figura 8.5. Variables de probabilidades y grupo de pertenencia en la vista de datos en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Si se revisa la vista de datos en SPSS, se encuentra que se generaron dos variables (figura 8.5): las probabilidades (PRE_1) y el grupo de pertenencia (PGR_1). Se muestra que las probabilidades que son mayores a 0.5 clasifican a las personas en el grupo 1; las menores a ese punto de corte se destinan al grupo 0.

Retornando a la salida de la figura 8.4, el paquete indica que el total de clasificación correcta realizado por el modelo (en otras palabras, el porcentaje de apoyos predichos que sí eran apoyos en realidad y el porcentaje de no apoyos predichos como tales y que no apoyan las parejas en realidad) es de 72.4%.

Si la versión de SPSS disponible no incluye el módulo de regresión, se debe tomar otro camino, algo más largo pero con resultados prácticamente iguales.

³³ Como indican Hosmer, Lemeshow y Sturdivant (2013, p. 171), el inconveniente con la tabla de clasificación es que obliga a pasar de una escala continua (las probabilidades) a otra categórica (0 y 1), originando pérdida de información. Así, si una observación obtiene una probabilidad de 0.48 y otra una de 0.52, con un punto de corte de 0.5 serían clasificados en dos categorías completamente diferentes, aunque en términos de las probabilidades son muy similares.

Recuadro 8.2

Resumen del procedimiento no convencional de regresión logística en SPSS

Analizar → Regresión → Ordinal

Se traslada la variable dependiente a su casilla correspondiente; las variables independientes se mueven hacia Covariables.

En Resultados, seleccionar Estadísticos de bondad de ajuste, Estadísticos de resumen, Estimaciones de los parámetros, Probabilidades de respuesta estimadas y Categoría pronosticada. Continuar.

Aceptar.

Para obtener la tabla de clasificación (que sería inmediata con el comando Logística binaria), se puede construir una tabla de contingencia manualmente.

Analizar → Estadísticos descriptivos → Tablas de contingencia

Trasladar a filas y columnas la variable dependiente y la variable llamada Categoría de respuesta pronosticada, la cual fue creada al realizar la estimación logística. Con ella se obtienen conteos de predicciones correctas e incorrectas para cada valor de la dependiente. Para conocer el porcentaje de clasificación correcta, se deben sumar los casos clasificados correctamente, dividir entre el total de la muestra y multiplicar por cien.

Similar al procedimiento convencional ya visto, con el método de regresión ordinal también se trasladan las variables respectivas (las independientes se incluyen como covariables) (figura 8.6).

En las opciones de Resultado (figura 8.7), se solicitan los Estadísticos de bondad de ajuste, Estadísticos de resumen, Estimaciones de los parámetros, Probabilidades de respuesta estimadas y Categoría pronosticada. Luego, Continuar y Aceptar.

En la figura 8.8, se resumen las salidas de la regresión ordinal, en particular las estimaciones de los parámetros. Para obtener, la clasificación se construye una tabla de contingencia (figura 8.9) definiendo en Filas la variable dependiente (votar) y en Columnas la variable Categorías de respuesta pronosticada que se

creó al ejecutar la regresión ordinal. Luego, Aceptar. En la figura 8.10, se encuentra la tabla de contingencia resultante.

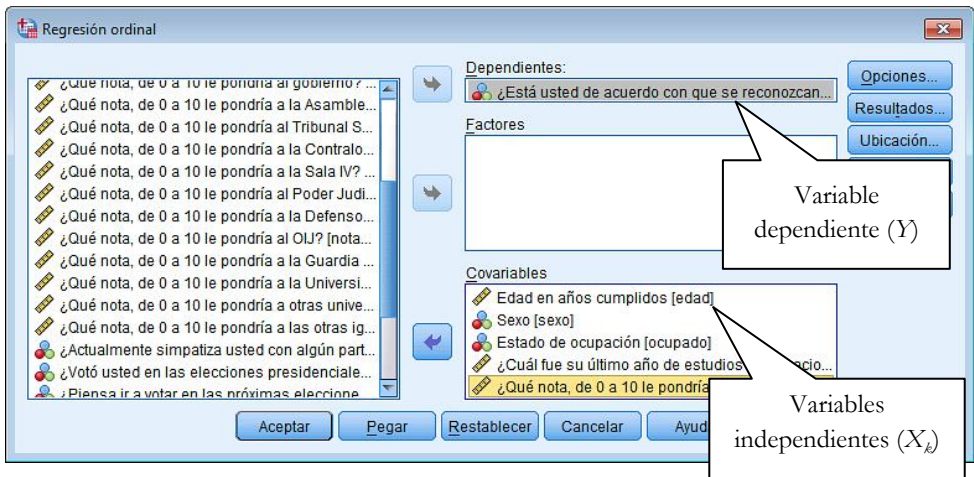


Figura 8.6. Ventana de regresión ordinal en SPSS (procedimiento no convencional de regresión logística).

Fuente: elaboración propia con base en el paquete SPSS.

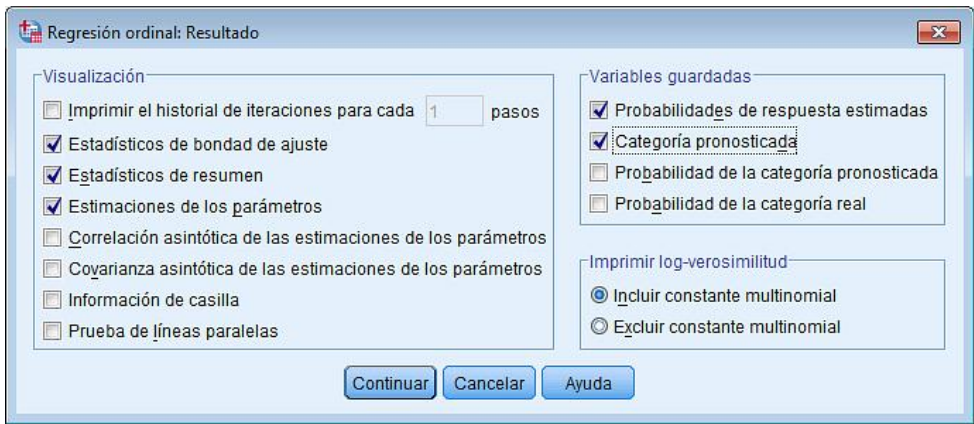


Figura 8.7. Ventana de resultados de regresión ordinal en SPSS (procedimiento no convencional de regresión logística).

Fuente: elaboración propia con base en el paquete SPSS.

Estimaciones de los parámetros

		Estimación	Error típ.	Wald	gl	Sig.	Intervalo de confianza 95%	
							Límite inferior	Límite superior
Umbral	[apoyoparejas = ,00]	.713	.479	2.220	1	.136	-.225	1.651
Ubicación	edad	-.015	.006	5.953	1	.015	-.027	-.003
	sexo	-.064	.201	.102	1	.749	-.457	.329
	ocupado	-.012	.210	.003	1	.955	-.423	.399
	educacion	.365	.065	31.814	1	.000	.238	.492
	notaiglesiacat	-.103	.035	8.510	1	.004	-.173	-.034

Función de vínculo: Logit.

Figura 8.8. Salidas de regresión ordinal en SPSS (procedimiento no convencional de regresión logística).

Fuente: elaboración propia con base en el paquete SPSS.

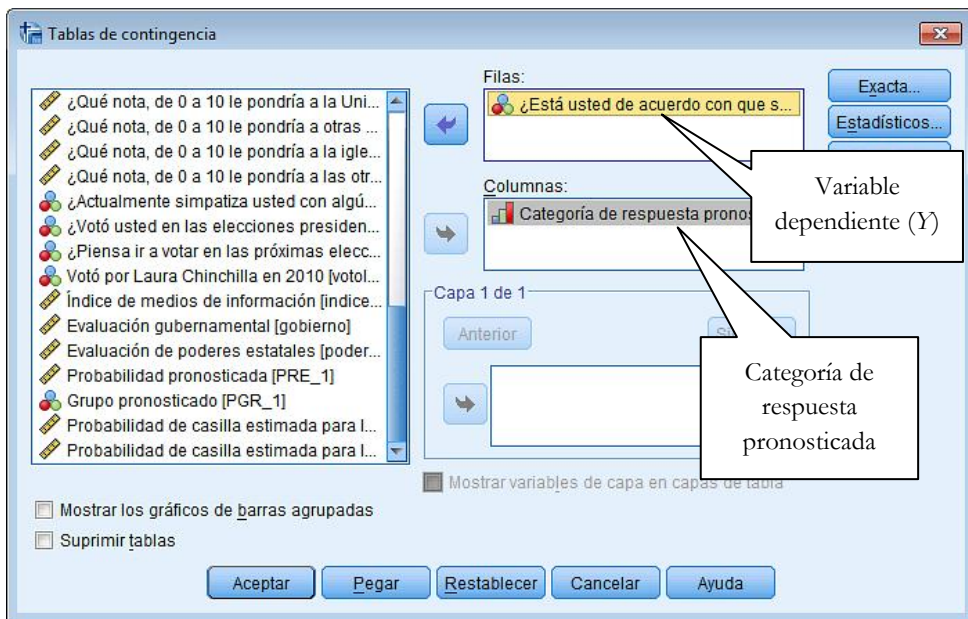


Figura 8.9. Ventana de tablas de contingencia en SPSS (procedimiento no convencional de regresión logística).

Fuente: elaboración propia con base en el paquete SPSS.

Tabla de contingencia apoyoparejas ¿Está usted de acuerdo con que se reconozcan legalmente a las parejas formadas por personas del mismo sexo? * PRE_1 Categoría de respuesta pronosticada

Recuento

	PRE_1 Categoría de respuesta pronosticada		Total
	,00 no	1,00 sí	
apoyoparejas ¿Está usted de ,00 no	359	38	397
acuerdo con que se reconozcan 1,00 sí	121	59	180
legalmente a las parejas formadas			
por personas del mismo sexo?			
Total	480	97	577

Figura 8.10. Tabla de contingencia para obtener el porcentaje de clasificación correcta (procedimiento no convencional de regresión logística).

Fuente: elaboración propia con base en el paquete SPSS.

Utilizando este procedimiento no convencional, los resultados de la significancia y los coeficientes son los mismos (solo préstese atención a que el procedimiento cambia el signo del intercepto, llamado umbral, respecto al procedimiento convencional).

La tabla de clasificación se construyó con la tabla de contingencia entre la respuesta observada (real) y la predicción de las categorías: 59 personas que apoyan las parejas fueron predichas correctamente y 359 personas que rechazan el reconocimiento también se predijeron como rechazo. Por lo tanto, la proporción de clasificación correcta es $\frac{(359+59)}{577} = 0.724$, que en porcentaje sería 72.4% (igual que con el procedimiento convencional de regresión logística).

Evaluación. Con cualquiera de los dos procedimientos, se llega al mismo resultado que se presenta en el cuadro 8.2 para su interpretación.

Se encuentra que las variables edad, educación y nota a la iglesia católica son significativas al 5% para predecir el apoyo al reconocimiento legal de las parejas de personas del mismo sexo. La magnitud de los efectos de las variables se suele leer como cambios en la razón de ventaja (*odds ratio*) respecto a la variable

dependiente.³⁴ Pero aquí la interpretación se centrará en el signo del coeficiente (sin analizar las razones de ventaja): si es positivo, al aumentar una variable el chance de la ocurrencia es mayor; si es negativo, el chance de la ocurrencia disminuye. En el caso de variables independientes cualitativas, si el coeficiente es positivo, el chance es mayor entre el grupo con 1 de la variable independiente; si es negativo, el chance de ocurrencia es mayor entre el grupo con 0 de la variable independiente.

Cuadro 8.2. Resultados del modelo de regresión logística sobre apoyo a parejas del mismo sexo

Variable	Coeficiente	Significancia
Constante	-0.713	0.136
Edad	-0.015	0.015
Sexo	-0.064	0.749
Ocupado	-0.012	0.995
Nivel educativo	0.365	0.000
Nota iglesia católica	-0.103	0.004
Casos clasificados correctamente:	72.4%	

Fuente: elaboración propia con base en la información de la figura 8.4.

De esta manera, los resultados se pueden leer de la siguiente manera:

- Al aumentar la edad, disminuye el chance de apoyar el reconocimiento de las parejas (el signo es negativo), con todas las demás variables constantes.
- Al aumentar el nivel educativo, el chance de apoyar aumenta (el signo es positivo), constantes todas las demás variables.
- Conforme aumenta la nota con que se califica a la iglesia católica, el chance de apoyar el reconocimiento de las parejas disminuye (signo negativo), con todas las variables constantes.
- Las variables sexo y ocupación no tienen efectos significativos.
- En cuanto a la bondad de ajuste del modelo, se clasificaron correctamente un 72.4% de los casos, lo cual sin duda indica un buen

³⁴ El *odds* se traduce aquí como “chance” pero puede resultar impreciso de todas formas. Lo importante es no confundirlo con probabilidad, pues el *odds* es un cociente de probabilidades: la probabilidad de la ocurrencia entre la probabilidad de la no ocurrencia.

rendimiento. En general, para lograr buenas clasificaciones es preferible contar con variables dependientes con bastante variabilidad (como la presente en el ejemplo); es decir, con el evento de estudio cercano al 50% y no a porcentajes muy bajos o muy altos, pues las tablas favorecen el pronóstico hacia los grupos más grandes (Hosmer, Lemeshow y Sturdivant, 2013, p. 17).

En resumen, el modelo ajustado es:

$$z = -0.713 - 0.015EDAD - 0.064SEXO - 0.012OCUPADO \\ + 0.365EDUCACION - 0.103NOTAIGLESIACAT,$$

con

$$P(Y = 1) = \frac{e^z}{1 + e^z},$$

Al igual que con el modelo de regresión normal, con la regresión logística se pueden hacer predicciones, pero en este caso de probabilidades. Por ejemplo: ¿cuál es la probabilidad predicha de que una mujer de 40 años, que no trabaja, con tres niveles de educación y que califica a la iglesia católica con un 9 apoye el reconocimiento de las parejas?

$$z = -0.713 - 0.015(40) - 0.064(1) - 0.012(0) + 0.365(3) - 0.103(9) = -1.209.$$

$$P(Y = 1) = \frac{e^{-1.209}}{1 + e^{-1.209}} = 0.23.$$

Según el modelo, la probabilidad de que una mujer con esas características apoye el reconocimiento legal de las parejas es 0.23.

Comentarios finales

- El modelo logístico estudiado aplica solo para el caso de una variable dependiente binaria, pero existen también modelos para variables con más categorías (el logístico ordinal y el multinomial).
- La regresión logística no es la única utilizada para predecir variables categóricas. Otro modelo muy popular es el “probit”, bastante común en

economía y en ciencia política (ver Przeworski *et al.*, 2000; Wolfinger y Rosenstone, 1980). No obstante, en la práctica difícilmente se encuentran diferencias entre el modelo logístico y el probit (Glasgow y Alvarez, 2008, p. 516).

- Respecto al modelo normal por mínimos cuadrados ordinarios, el logístico requiere de un mayor número de observaciones o casos por variable independiente. Algunos recomiendan un mínimo de 50 por cada variable independiente (Wright, 1995).
- Aunque los paquetes (SPSS incluido, ver figura 8.4) proveen una medida llamada pseudo R^2 como medida de bondad de ajuste, aquí se prefirió darle atención a la tabla de clasificación. Si bien el pseudo R^2 puede servir para comparar modelos, nunca debe interpretarse como una proporción de variancia explicada (como ocurre con la regresión normal) por sus diferencias conceptuales (ver Hosmer, Lemeshow y Sturdivant, 2013).

Ejercicios

1. Se quiere estudiar cuáles factores se asocian significativamente a que una persona piense ir a votar en el periodo preelectoral. Con datos de la encuesta de noviembre de 2013 del CIEP, estime en SPSS un modelo de regresión logística que incluya las variables según se muestra en el Cuadro 8.3.
2. Interprete la significancia y dirección de los coeficientes según lo obtenido en el punto 1. Además, evalúe la bondad de ajuste según el porcentaje de clasificación correcta.
3. Con el modelo estimado, calcule la probabilidad de que una mujer de 29 años con educación universitaria completa, sin simpatías partidarias y que votó en 2010 piense ir a votar en 2014. Sugerencia: recuerde que si estima el modelo por el procedimiento no convencional de regresión ordinal, debe cambiar el signo del intercepto o el umbral.

Cuadro 8.3. Codificación de variables

Nombre de la variable	Código en la base de datos	Significado y valores
<i>Dependiente</i>		
Piensa votar en 2014	VOTO2014	1=piensa ir a votar
<i>Independientes</i>		
Sexo	SEXO	0=no piensa ir a votar
		1=mujer
		0=hombre
Edad	EDAD	Edad en años cumplidos
Nivel educativo	EDUCACION	Escala de 0 (ninguna educación formal) a 6 (educación universitaria completa)
Simpatía partidaria	SIMPATIZA	1=simpatiza con algún partido
		0=no simpatiza
Votó en 2010	VOTO2010	1=sí votó
		0=no votó

Fuente: elaboración propia con base en CIEP (2012-2014).

CAPÍTULO 9

ANÁLISIS DE CONGLOMERADOS

Introducción

Las técnicas denominadas análisis de conglomerados o agrupamientos (también conocidos por el término inglés *clusters*) constituyen poderosas herramientas para la descripción y exploración de datos multivariados. Específicamente, permiten clasificar observaciones o casos en grupos homogéneos, es decir, conformados por entidades con características similares. Esta clasificación facilita no solo resumir información (con lo cual se gana simplicidad descriptiva), sino, además, construir categorizaciones y tipologías.

En ciencia política, se ha recurrido al análisis de conglomerados en clasificaciones de individuos según su comportamiento electoral (Verba y Nie, 1972), países según sus regímenes de bienestar (Martínez Franzoni, 2008), partidos políticos (Altman *et al.*, 2009), elecciones en países latinoamericanos (Pignataro, 2012), entre otras posibles aplicaciones.

Existen numerosas técnicas de agrupamientos. A continuación, se examinará el método aglomerante jerárquico; este se caracteriza por iniciar en un punto donde hay tantos grupos como casos, los cuales se van reuniendo en grupos cada vez más grandes hasta formar un grupo único con todos los casos (Hernández, 2013). El método aglomerante jerárquico se basa en el cálculo de distancias entre los valores respectivos de cada objeto, caso u observación según una o más variables

predefinidas. Estas distancias se pueden determinar por muchas fórmulas distintas, pero los próximos ejemplos utilizarán la distancia euclidiana.³⁵

Si se tienen dos objetos (u observaciones como personas encuestadas, países, etc.), medidos con distintas variables p , la distancia euclidiana se calcula como:

$$d(I_1, I_2) = \sqrt{\sum_{k=1}^p (X_{1k} - X_{2k})^2}.$$

Es decir, la distancia euclidiana calcula la disimilitud o diferencia entre los objetos, de modo que si tienen valores iguales para todas las variables, evidentemente la distancia es de cero. Para ejemplificar, con dos observaciones y cuatro variables se tienen los siguientes datos:

Objeto	X_1	X_2	X_3	X_4
1	7	1	6	8
2	5	3	2	9

Entonces, la distancia euclidiana se calcula así:

$$d(I_1, I_2) = \sqrt{(7 - 5)^2 + (1 - 3)^2 + (6 - 2)^2 + (8 - 9)^2}$$

$$d(I_1, I_2) = \sqrt{(2)^2 + (-2)^2 + (4)^2 + (-1)^2}$$

$$d(I_1, I_2) = \sqrt{25} = 5.$$

Con estas distancias, se establece la secuencia de agrupación de objetos (vinculando a los más semejantes en grupos homogéneos) para efectuar el análisis de conglomerados. Claramente, conforme aumenta el número de casos y de variables, los cálculos manuales se tornan muy intensivos, por lo tanto será conveniente recurrir a los paquetes estadísticos.

³⁵ Otras son la distancia Manhattan, la métrica de Minkowski y la distancia generalizada de Mahalanobis (Aldenderfer y Blashfield, 1984).

Ejemplo 1: clasificación de elecciones

En primer lugar, se tienen definidos una serie de observaciones o casos. En este ejemplo, los casos corresponden a elecciones presidenciales de primera ronda; las características, atributos o variables son dos: (a) participación electoral (porcentaje según electores registrados); y (b) mayoría parlamentaria del candidato ganador (porcentaje de votos obtenidos). Los valores se presentan en el cuadro 9.1. El objetivo, por lo tanto, es clasificar las elecciones en grupos según ambas variables por medio del análisis de conglomerados.

Cuadro 9.1. Datos de elecciones

Elección según país (año)	Participación electoral (%)	Mayoría parlamentaria (%)
Bolivia (2009)	94.55	64.22
Colombia (2010)	49.24	25.18
Nicaragua (2006)	61.23	37.18
Paraguay (2008)	60.34	35.27
Perú (2006)	88.71	21.15
Venezuela (2006)	74.69	85.50

Fuente: Pignataro (2012).

Recuadro 9.1

Resumen del procedimiento para el análisis de conglomerados en SPSS

Analizar → Clasificar → Conglomerados jerárquicos

Seleccionar las variables con las cuales se realiza la clasificación y la variable para darle etiqueta a los casos (en Etiquetar los casos mediante). En Conglomerar, marcar Casos.

En Estadísticos, seleccionar Matriz de distancias.

En Gráficos, seleccionar Dendrograma (en Témpanos, marcar Ninguna).

En Método, señalar en Intervalo la Distancia euclídea. Luego, se puede escoger, en Método de conglomeración, las técnicas de vinculación: Vecino más próximo, Vecino más lejano y Vinculación inter-grupos. (Se deben escoger uno a la vez y luego comparar.) Continuar. Aceptar.

En SPSS, se debe buscar la opción Clasificar y Conglomerados jerárquicos (figura 9.1). En la ventana que se abre, se trasladan hacia el recuadro superior las variables que se quieren utilizar para la clasificación. Abajo se puede trasladar la variable de identificación de los casos (*i. e.* aquella que incluye los nombres de los casos). Indicar conglomeración por Casos.

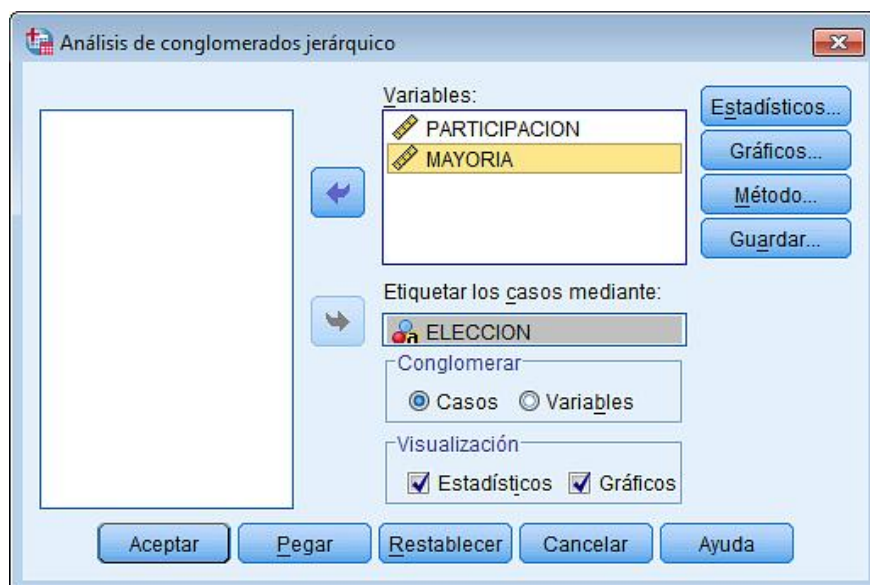


Figura 9.1. Ventana de análisis de conglomerados jerárquicos en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Para observar las distancias calculadas, se indica Matriz de distancias en la subventana de Estadísticos (figura 9.2). En la opción de Gráficos, señalar Dendrograma (figura 9.3) (no hace falta seleccionar un gráfico de témpanos, pues aquí no se va a analizar). Luego, Continuar.

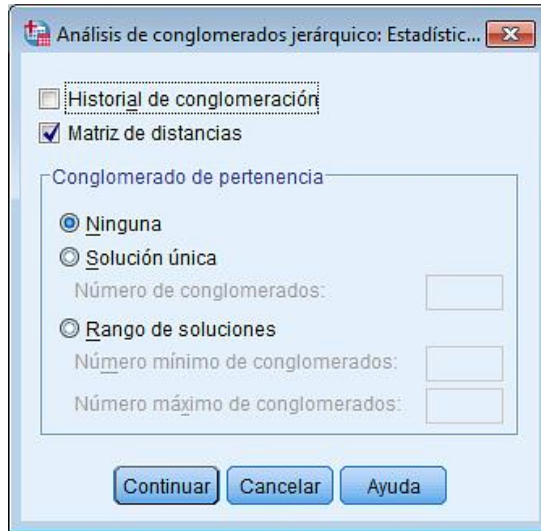


Figura 9.2. Ventana de Estadísticos de Análisis de Conglomerados Jerárquicos en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

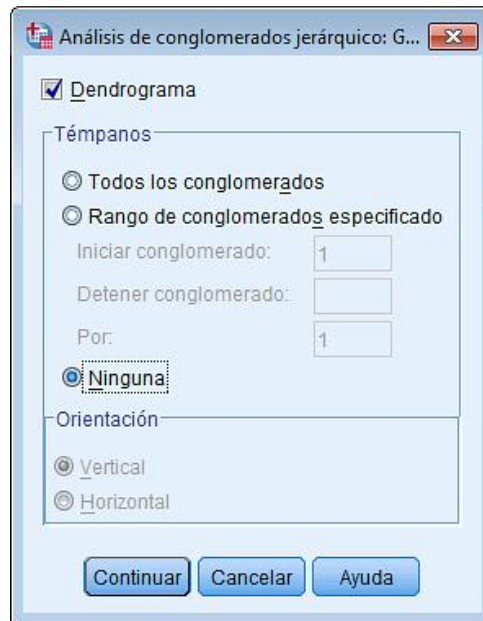


Figura 9.3. Ventana de Gráficos de Análisis de Conglomerados Jerárquicos en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

En Método (figura 9.4), se especifica la distancia euclídea para Intervalo. En Método de conglomeración, existen diversas opciones; por el momento, se va a solicitar la Vinculación intergrupos. Además, en Transformación de valores se pueden estandarizar variables cuando están medidas con distintas escalas. En este ejemplo no aplica, pues tanto la participación como la mayoría parlamentaria se miden con porcentajes. Continuar y Aceptar.

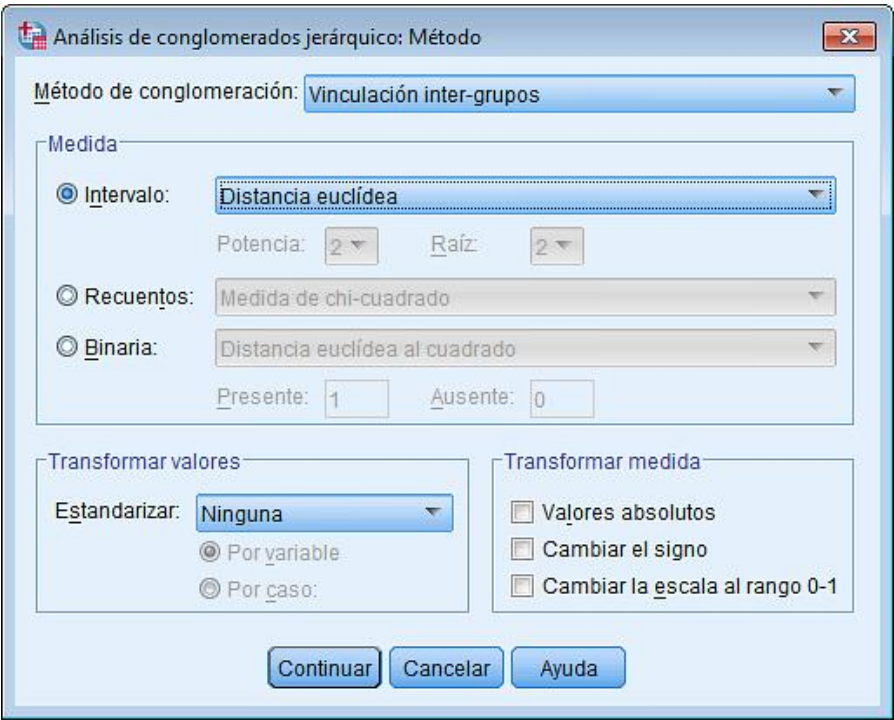


Figura 9.4. Ventana de Método de Análisis de Conglomerados Jerárquicos en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

La primera salida que se obtiene es la matriz de distancias euclidianas (figura 9.5), calculadas exactamente con la fórmula vista en la introducción del capítulo, lo cual se puede comprobar fácilmente. Por ejemplo, la distancia euclidiana entre Bolivia 2009 y Colombia 2010 es:

$$d(BOL2009, COL2010) = \sqrt{(94.55 - 49.24)^2 + (64.22 - 25.18)^2}$$

$$d(BOL2009, COL2010) = \sqrt{2052.9961 + 1524.1216}$$

$$d(BOL2009, COL2010) = \sqrt{3577.1177}$$

$$d(BOL2009, COL2010) = 59.809.$$

Matriz de distancias						
Caso	distancia euclídea					
	1:BOL2009	2:COL2010	3:NIC2006	4:PAR2008	5:PER2006	6:VEN2006
1:BOL2009	.000	59.809	42.911	44.815	43.464	29.108
2:COL2010	59.809	.000	16.963	15.001	39.675	65.469
3:NIC2006	42.911	16.963	.000	2.107	31.814	50.160
4:PAR2008	44.815	15.001	2.107	.000	31.690	52.240
5:PER2006	43.464	39.675	31.814	31.690	.000	65.860
6:VEN2006	29.108	65.469	50.160	52.240	65.860	.000

Esta es una matriz de disimilaridades

Figura 9.5. Matriz de distancias para el ejemplo de elecciones (vinculación intergrupos).

Fuente: elaboración propia con base en el paquete SPSS.

Puede verse que la mayor distancia calculada corresponde a las elecciones de Perú 2006 y Venezuela 2006, pues son las más disímiles entre sí (aunque la participación electoral fue similar; en el primer caso, la mayoría fue del 21.15% y en el segundo de 85.50%) mientras que las elecciones más similares son las de Paraguay 2008 y Nicaragua 2006, pues tienen valores muy parecidos en las dos variables.

Nótese que la matriz de distancias es simétrica porque la disimilitud de un objeto 1 con un objeto 2 es la misma que si se calcula entre el 2 y el 1. Como se dijo anteriormente, no hay disimilitud entre un mismo objeto y por ello la diagonal está compuesta por ceros.

Como resultado del análisis de conglomerados se interpretará el gráfico llamado árbol de clasificación o dendrograma (figura 9.6).

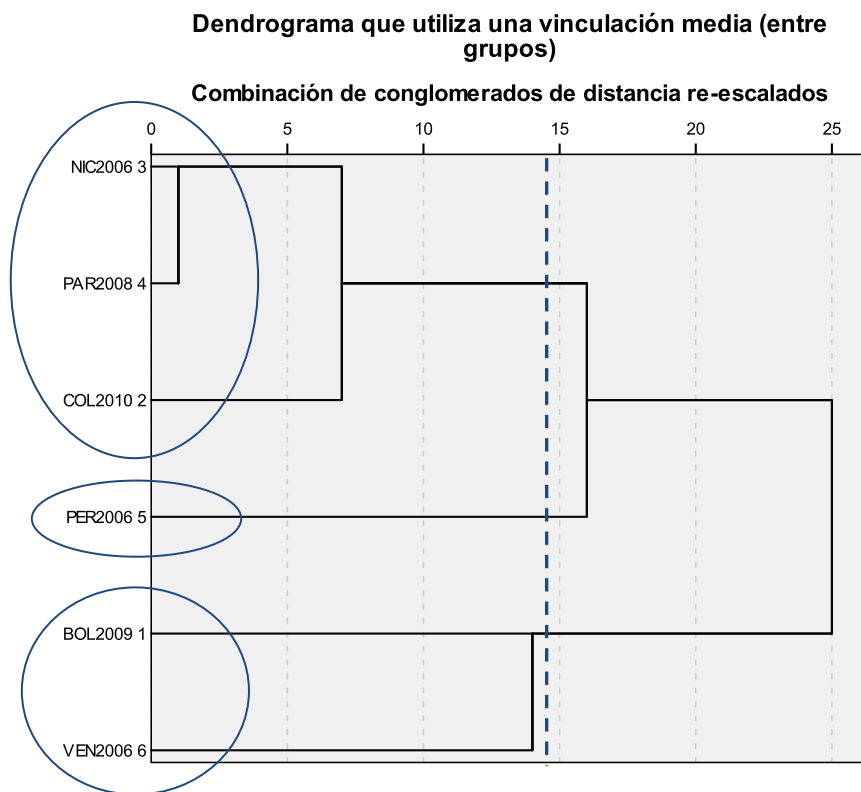


Figura 9.6. Dendrograma de clasificación para el ejemplo de elecciones (vinculación intergrupos).

Fuente: elaboración propia con base en el paquete SPSS.

Como se indicó antes, el análisis de aglomeración jerárquica parte de tantos grupos como casos existan (extremo izquierdo del gráfico); las líneas indican cómo se van formando los grupos y la distancia horizontal muestra la secuencia de conglomeración. La interpretación del dendrograma se realiza con base en la información sustantiva, es decir los datos del cuadro 9.1.

Primero, Paraguay y Nicaragua forman un conglomerado, pues ambas constituyen elecciones con participación y mayoría parlamentaria bajas; luego, se les une Colombia cuya participación y mayoría parlamentaria son aún más bajas. Por otro lado, Bolivia y Venezuela forman un segundo conglomerado, pues comparten alta participación y gran mayoría parlamentaria. Luego Perú (con alta participación pero baja mayoría), se une al grupo de Paraguay, Nicaragua y Colombia y, finalmente, todos los casos se agrupan en un mismo conglomerado.

El punto crucial en la interpretación del dendrograma consiste en decidir el número aceptable de agrupamientos. Esto se define trazando un corte horizontal en la secuencia de conglomeración, dependiendo del número de grupos deseados, el conocimiento sustantivo de los casos y las expectativas teóricas (por ejemplo, si se tienen hipótesis sobre el número y tipo de conglomerados). Así, un observador puede realizar un corte como el mostrado por la línea segmentada en la figura 9.6, con lo cual se definen tres grupos:

- Grupo 1: Paraguay 2008, Nicaragua 2006 y Colombia 2010 (participación y mayoría parlamentaria bajas).
- Grupo 2: Perú 2006 (caso excepcional de participación muy alta y mayoría parlamentaria muy baja).
- Grupo 3: Bolivia 2009 y Venezuela 2006 (casos con participación y mayorías altas).

Gracias a la clasificación del dendrograma es posible, además, construir una tipología para las elecciones estudiadas como una forma adicional de presentar los casos (Cuadro 9.2). Nótese que no existen casos empíricos que correspondan a elecciones con participación alta y mayoría parlamentaria baja, lo cual es un indicio sobre la relación directa existente entre ambas variables.

Cuadro 9.2. Tipología construida para el ejemplo 1

Participación electoral	Mayoría parlamentaria	
	Baja	Alta
Baja	Paraguay 2008, Nicaragua 2006, Colombia 2010	Perú 2006
Alta	(sin casos)	Bolivia 2009, Venezuela 2006

Fuente: elaboración propia con base en la información de la figura 9.6.

No se puede ignorar que la técnica de aglomeración jerárquica permite variar los tipos de enlace que establecen las reglas de formación de los grupos. Se examinarán tres de ellos: vinculación única, completa o promedio. En el cuadro 9.3, se precisa su nomenclatura en español e inglés, así como algunas ventajas y desventajas de cada tipo de vinculación.

Cuadro 9.3. Tipos de enlace en la aglomeración jerárquica

Tipo de enlace	Otros nombres	Nombre en inglés	Ventaja/desventaja
Vinculación única	Vecino más próximo o más cercano	<i>Nearest neighbor</i>	Tiende a formar “cadenas” en la aglomeración (casos individuales que se van anexando a un grupo grande ya formado)
Vinculación completa	Vecino más lejano	<i>Furthest neighbor</i>	Forma conglomerados compuestos por casos muy similares (lo cual no siempre coincide con la estructura teórica)
Vinculación promedio	Vinculación intergrupos o entre grupos	<i>Between-groups linkage</i>	Solución intermedia entre las vinculaciones única y completa

Fuente: elaboración propia con base en Aldenderfer y Blashfield (1984).

Es recomendable comparar las distintas soluciones generadas por cada tipo de enlace. Si se obtienen agrupaciones similares, la clasificación obtenida puede estar respondiendo adecuadamente a la estructura natural de agrupación. Por el contrario, las soluciones muy diferentes indican que difícilmente existe una estructura definida de conglomeración (Hernández, 2013, p. 263). De cualquier manera, es necesario un conocimiento apropiado de los datos y el campo de estudio. En el próximo ejemplo, se compararán las soluciones según distintos enlaces.

Ejemplo 2: clasificación de parlamentos

Se tienen datos para 16 países europeos sobre el número de escaños en el parlamento (cámara baja o única) y el porcentaje de representación femenina en este, cifras actualizadas al 1º de enero de 2013 por Inter-Parliamentary Union (IPU, 2013).

Cuadro 9.4. Datos de parlamentos

País	Número de escaños parlamentarios	Porcentaje de representación femenina
Alemania	620	32.90
Austria	183	27.90
Bélgica	150	38.00
Dinamarca	179	39.10
España	350	36.00
Finlandia	200	42.50
Francia	577	26.90
Grecia	300	21.00
Holanda	150	38.70
Italia	630	21.40
Luxemburgo	60	21.70
Noruega	169	39.60
Portugal	230	28.70
Reino Unido	650	22.50
Suecia	349	44.70
Suiza	200	29.00

Fuente: IPU (2013).

Utilizando el procedimiento en SPSS ya señalado, se procede a realizar el análisis de aglomeración jerárquico y se compararán los distintos dendrogramas según el tipo de vinculación.

Un punto destacable es que el método es sensible a la escala de medición de las variables, por ello se recomienda, en ocasiones, estandarizar las variables (Aldenderfer y Blashfield, 1984, p. 20) y el mismo procedimiento de conglomerados en SPSS provee, en Método, la opción de estandarizar los valores (ver figura 9.7). En el ejemplo, el número de escaños y el porcentaje de representación involucran distintas escalas de medición, de manera que resulta adecuado estandarizarlas con puntuaciones z .

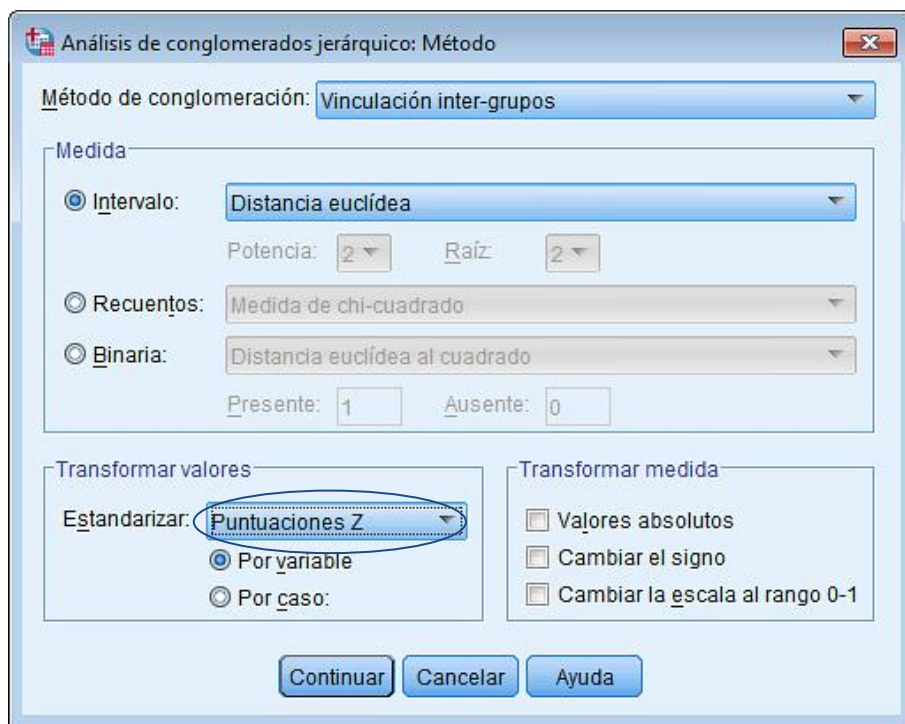


Figura 9.7. Ventana de Método de Análisis de Conglomerados Jerárquicos en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

En la figura 9.8, se encuentra el dendrograma según vinculación única. En este ejemplo, se puede observar una solución encadenada en la que se les suman casos individuales a varios grupos. Una interpretación consiste en examinar tres grupos:

- Grupo 1: Dinamarca, Noruega, Bélgica, Holanda, Finlandia, España y Suecia (parlamentos medianos y pequeños con representación femenina alta).
- Grupo 2: Portugal, Suiza, Austria, Grecia y Luxemburgo (parlamentos medianos y pequeños con representación femenina intermedia y baja).
- Grupo 3: Italia, Reino Unido, Francia y Alemania (parlamentos grandes y representación femenina intermedia y baja).

Las soluciones con la vinculación completa (figura 9.9) y con vinculación intergrupos (figura 9.10) son muy similares entre sí, y la configuración bajo estos dos enlaces es igual a la encontrada con la vinculación única. Gracias a esta

coherencia, la tipología descrita parece ser la representación más adecuada para clasificar a los parlamentos según tamaño y presencia de mujeres.

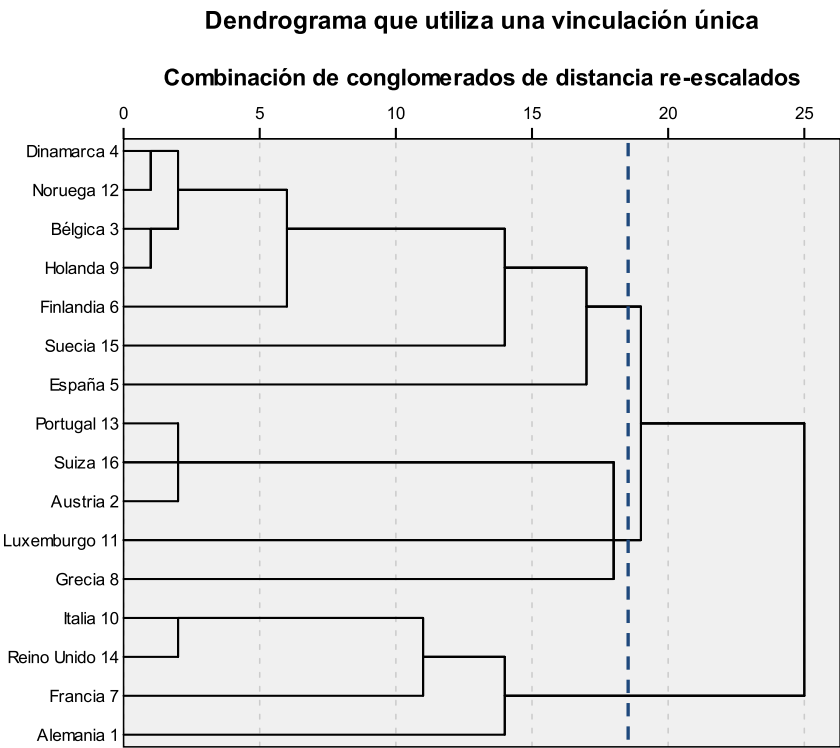


Figura 9.8. Dendrograma de clasificación para el ejemplo de parlamentos (vinculación única).

Fuente: elaboración propia con base en el paquete SPSS.

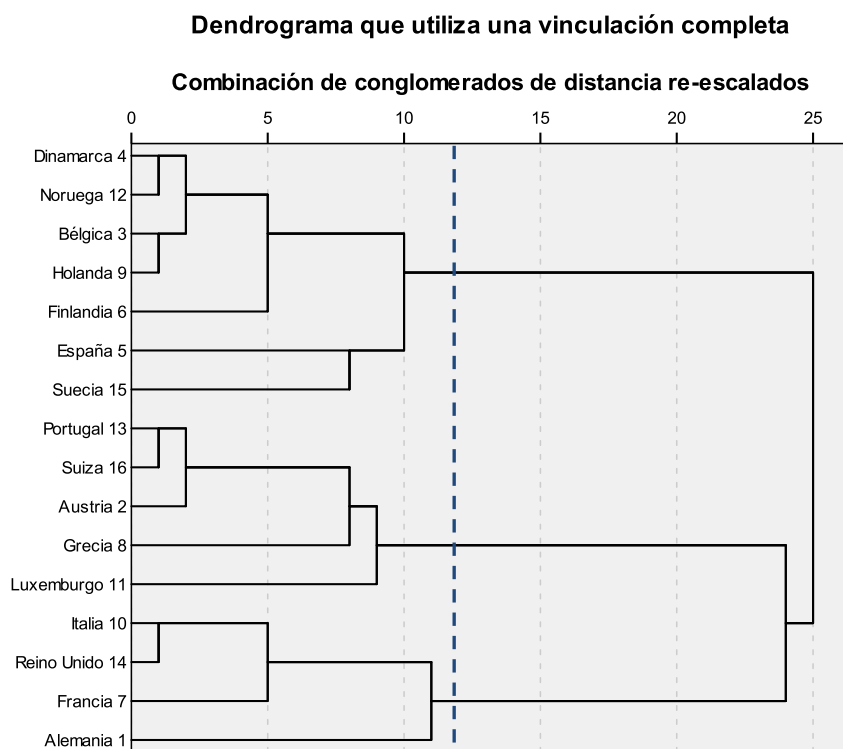


Figura 9.9. Dendrograma de clasificación para el ejemplo de parlamentos (vinculación completa).

Fuente: elaboración propia con base en el paquete SPSS.

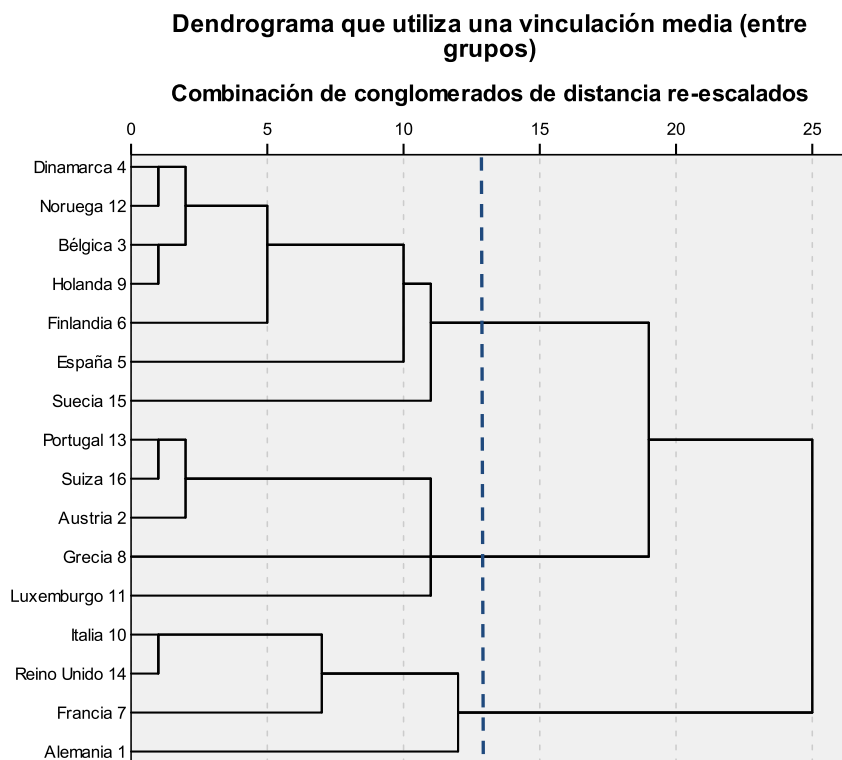


Figura 9.10. Dendrograma de clasificación para el ejemplo de parlamentos (vinculación intergrupos).

Fuente: elaboración propia con base en el paquete SPSS.

Comentarios finales

- El análisis de agrupamiento jerárquico no constituye precisamente un modelo estadístico (véase que no hay inferencia, cálculo de error ni intervalos de confianza); más bien, se basa en un algoritmo matemático, un procedimiento que calcula distancias entre objetos según reglas determinadas pero sin estimaciones.
- Se habrá notado que el método de aglomerativo jerárquico es más adecuado para análisis con pocos casos, sobre todo porque la interpretación del dendrograma se complica conforme aumenta el número de casos. Con muchas observaciones (*v. g.* una encuesta), preferiblemente se recurre a otros métodos de conglomeración como el k medias.

- El método, además, admite tanto variables métricas (como las utilizadas en los ejemplos) como categóricas, pero no permite combinar ambos tipos de variables (SPSS Inc., 2007).

Ejercicios

Con los siguientes datos de países latinoamericanos sobre la tasa de reelección de diputados y el número efectivo de partidos parlamentarios (OIR, 2014a y 2014b), utilice el análisis de conglomerados en SPSS para clasificar los casos según las dos variables. Pruebe los distintos enlaces para encontrar los dendrogramas con la representación más adecuada. Interprete los grupos resultantes. Sugerencia: recuerde estandarizar las variables si se considera necesario.

Cuadro 9.5. Reelección legislativa y número efectivo de partidos

País	Tasa de reelección en la Cámara Baja (promedio 1995- 2008) (%)	Número efectivo de partidos parlamentarios (diferentes años)
Argentina	52.01	3.65
Bolivia	9.99	1.40
Brasil	50.97	9.29
Chile	63.33	5.81
Colombia	32.12	7.20
Ecuador	12.69	5.85
El Salvador	43.44	2.94
Guatemala	13.50	4.78
Honduras	24.28	2.41
Nicaragua	27.16	3.24
Panamá	35.21	2.92
Paraguay	31.25	3.42
Perú	20.27	3.78
R. Dominicana	30.04	2.52
Uruguay	24.48	2.40
Venezuela	23.03	1.93

Fuente: OIR (2014a y 2014b).

CAPÍTULO 10

ANÁLISIS DE FACTORES

Introducción

El análisis de factores, inventado por Charles Spearman (1863-1945), es una técnica multivariada cuyo principal objetivo es describir una serie de datos con muchas variables mediante un número menor de variables, denominadas factores. Para ejemplificar: si originalmente se tenían diez variables para describir un conjunto de datos, con el análisis factorial se puede lograr resumir ese mismo conjunto de datos con dos factores. Adicionalmente, los factores que se construyen pueden tener poca o ninguna correlación entre sí (Hernández, 2013).

Existen dos enfoques en el análisis de factores. El primero se denomina el *análisis factorial exploratorio*, en el cual se busca un resultado que se ajuste mejor a los datos. El segundo es llamado *análisis factorial confirmatorio* y permite imponer un modelo factorial particular a los datos. En términos metodológicos, la diferencia más importante en cuanto a su utilidad es que el exploratorio permite construir teoría cuando el conocimiento teórico es escaso, no se tienen hipótesis y precisamente se quieren explorar datos. El confirmatorio, por su parte, se utiliza para probar teorías e hipótesis previamente desarrolladas (Bryan y Yarnold, 1995, p. 109).

Entre los usos que se le dan al análisis factorial, se destacan los siguientes:

- Reducción de variables en dimensiones teóricas (*v. g.* diferenciar dimensiones políticas, económicas y culturales).
- Construcción de índices, según el cual la misma estructura de datos genera una ponderación más refinada de cada variable (en lugar de realizar promedios en los que la ponderación es igual para todas).

- Solución a problemas de regresión como la multicolinealidad (cuando las variables independientes están muy correlacionadas entre sí, ver capítulo 7), ya que se producen nuevas variables no correlacionadas entre sí que pueden utilizarse como independientes. También, al reducir el número de variables, se puede corregir el exceso de variables respecto al número de casos.
- Comprobación de teorías e hipótesis (previamente desarrolladas), en el caso de análisis factorial confirmatorio.

Existen numerosas aplicaciones en ciencia política (al igual que en sociología y psicología). Por ejemplo, Verba y Nie (1972, pp. 60-65) aplicaron análisis factorial a 13 variables sobre participación política provenientes de datos de encuesta (como si persuade a personas para votar, si trabaja activamente para un partido o candidato, si contribuye con dinero, frecuencia del voto, si trabaja con otras personas en problemas locales, si contacta políticos estatales y nacionales, entre otras). Con base en este conjunto de variables, obtuvieron cuatro factores en los que identificaron distintas modalidades de participación política: actividad en campañas, participación electoral, actividad cooperativa y contacto a políticos.

Por su parte, Putnam (1993), en su estudio de tradiciones cívicas en Italia, construyó un índice de desempeño institucional para las 20 regiones italianas a través del análisis factorial basándose en datos de políticas sociales, innovación legislativa, estabilidad de los gabinetes, rapidez en la políticas de salud, gasto público en el sector agrícola, grado de respuesta de la burocracia y otras (12 en total).

Asimismo, Altman *et al.* (2009) aplicaron el análisis factorial con datos de una encuesta de expertos. Con base en la caracterización de los partidos latinoamericanos, lograron identificar tres factores (teniendo inicialmente 20 variables): Estado-mercado, tradición y democracia.

Como se observa en los ejemplos anteriores, el análisis factorial permite una representación parsimoniosa de los datos, reduciendo un gran número de variables a unas pocas dimensiones o factores, los cuales pueden ser interpretados según los marcos conceptuales y teóricos de cada materia.

En las siguientes páginas, se estudiarán algunos principios básicos del análisis factorial exploratorio con datos métricos o continuos a través de dos ejemplos.

Se enfocará en la interpretación de los factores resultantes según la solución rotada tipo Varimax.³⁶

Conceptos básicos

El análisis de factores plantea un modelo estadístico para explicar la variabilidad total de los datos. A diferencia de los modelos de regresión, no establece un relación entre variables independientes que explican una dependiente, sino que produce nuevas variables no observadas (o latentes) a partir de los datos disponibles, pero sin establecer relaciones causales.

Se recordará que en el análisis de variancia (capítulo 4) se dividía la variabilidad total en variabilidad entre grupos y dentro de grupos. En el análisis de factores, también se descompone la variabilidad total, pero, en este caso, entre variancia compartida y variancia específica. La variancia compartida, denominada *comunalidad*, para una variable X , se refiere a la variabilidad que comparte con otras variables a través de factores comunes. La variancia específica es exclusiva de la variable X y se llama variancia específica o *unicidad*. En resumen:

$$\text{Variabilidad total de } X = \text{comunalidad} + \text{unicidad}.$$

El punto clave del análisis es determinar factores que contengan una proporción importante de comunalidad, es decir, esperar encontrar una amplia variancia compartida entre variables para expresarlas en unos pocos factores y que cuanto no se pueda explicar por los factores comunes (la unicidad) sea menor. Por ejemplo, si se tienen cuatro variables, se podría generar un factor que explica el 70% de la variabilidad (es decir, es variancia compartida); el otro 30% corresponde a la unicidad de cada variable. Con ello, se gana en eficiencia y parsimonia: no hace falta utilizar cuatro variables cuando un único factor explica un 70% de lo que ellas originalmente contenían. El análisis de factores calcula

³⁶ Aunque no se abordará en el capítulo, el análisis de factores confirmatorio es un enfoque más sofisticado y potente que el exploratorio, pues resulta más útil para la investigación guiada teóricamente, sirve para validar constructos o conceptos teóricos y se extiende hacia los llamados modelos de ecuaciones estructurales que permiten modelar relaciones más complejas (ver Brown, 2006).

cargas factoriales que corresponden a la correlación entre cada variable X y cada factor. Al sumar las cargas factoriales al cuadrado, se obtiene la comunalidad:

$$\text{Comunalidad de } X = (\text{carga}_1)^2 + (\text{carga}_2)^2 + (\text{carga}_3)^2 + \dots$$

Por lo tanto, la variancia total de la variable X es la suma de las cargas factoriales al cuadrado más la unicidad (y esta suma de variancia compartida y específica es igual a 1, pues son variables estandarizadas):

$$\text{Variabilidad de } X = (\text{carga}_1)^2 + (\text{carga}_2)^2 + (\text{carga}_3)^2 + \dots + \text{unicidad} = 1.$$

Adicionalmente, se pueden aplicar las llamadas rotaciones, las cuales permiten una mejor distribución de las cargas factoriales para que estas se concentren solo en algunos factores y se facilite la interpretación.

Ejemplo 1: satisfacción con los servicios públicos

En la encuesta del CIEP realizada entre octubre y noviembre de 2012, uno de sus objetivos era explorar la satisfacción con políticas públicas, entre ellas los servicios públicos de distinta índole, en concreto los siguientes: abastecimiento de agua, seguridad ciudadana, presencia de fuerza pública o policía, estado de las calles, mantenimiento de carreteras, recolección de basura, suministro de energía eléctrica, limpieza de alcantarillas, atención en EBAIS,³⁷ clínicas y hospitales de la CCSS. En total, 11 servicios.

En concreto se preguntó “¿Cómo califica [nombre del servicio público]? con la escala ‘muy malo’, ‘malo’, ‘regular’, ‘bueno’, ‘muy bueno’, o ‘no lo ha utilizado’.” Con base en esas respuestas, se tiene una escala métrica de 1 a 5 (mayor nota, mejor calificación). Se propuso utilizar el análisis factorial para reducir las 11 preguntas a un menor número de dimensiones, esperando que cada una tenga alguna interpretación conceptual o teórica (p. g. tipo de servicio).

El análisis de factores se realiza en SPSS a través del siguiente procedimiento.

³⁷ Los Equipos Básicos de Atención Primaria en Salud (EBAIS) son centros de atención médica en Costa Rica.

Recuadro 10.1

Resumen del procedimiento para el análisis de factores en SPSS

Analizar → Reducción de dimensiones → Factor

Trasladar las variables.

En Extracción, en Método indicar Componentes principales, seleccionar Matriz de correlaciones, Solución factorial sin rotar. Dejar Autovalores mayores que 1 y 25 como número máximo de iteraciones para convergencia. Continuar.

En Rotación, indicar Método Varimax y marcar Solución rotada. Continuar.

En Puntuaciones, seleccionar Guardar como variables y en Método marcar Regresión.

Aceptar.

En el menú Analizar de SPSS, se busca Reducción de dimensiones y Factor, con lo cual se abre la principal ventana para definir el análisis factorial (figura 10.1). En ella se trasladan las variables de interés, en este caso las referidas a la calificación de servicios públicos.

En la opción de Extracción (figura 10.2), se define como método Componentes Principales, analizar Matriz de Correlaciones, mostrar Solución factorial sin rotar y precisar Autovalores mayores que 1 y 25 como número máximo de iteraciones para convergencia. Luego, Continuar.

En la ventana de Rotación, marcar Varimax y Solución Rotada (figura 10.3). Continuar. En Puntuaciones, se pide que se guarden como variables bajo el Método Regresión (figura 10.4). Continuar y luego Aceptar.

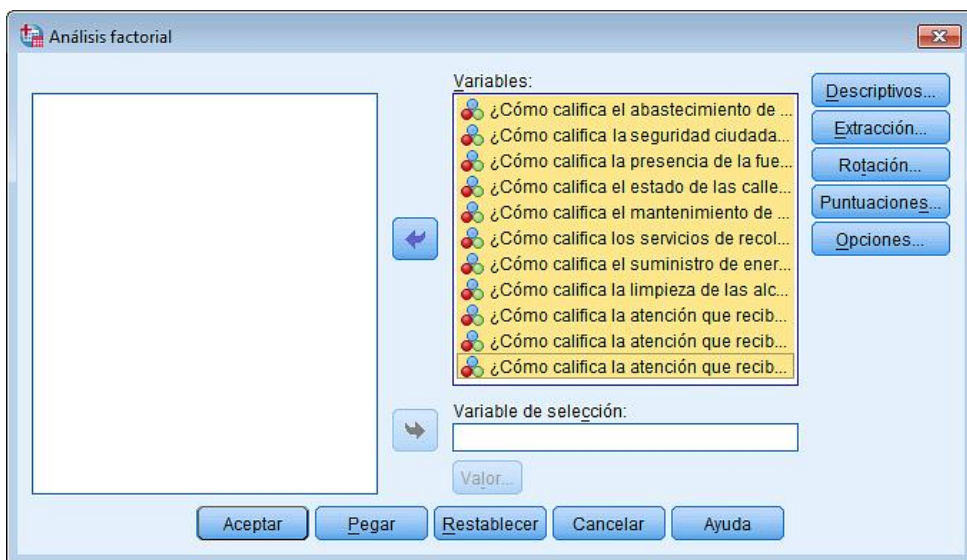


Figura 10.1. Ventana de Análisis Factorial en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

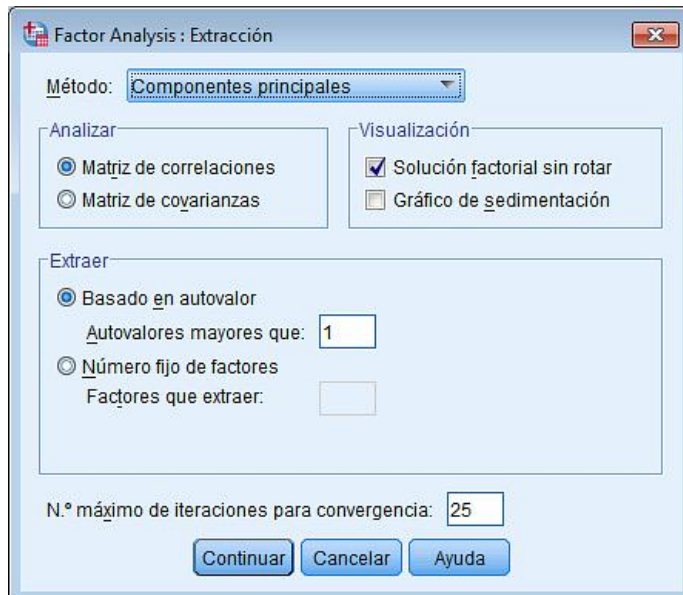


Figura 10.2. Ventana de Extracción para análisis factorial en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

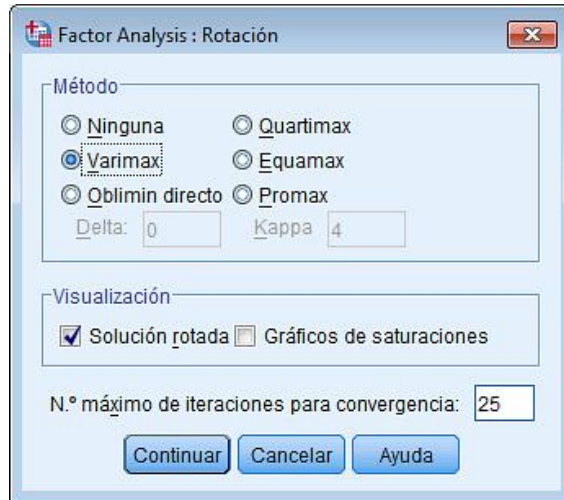


Figura 10.3. Ventana de Rotación para Análisis Factorial en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

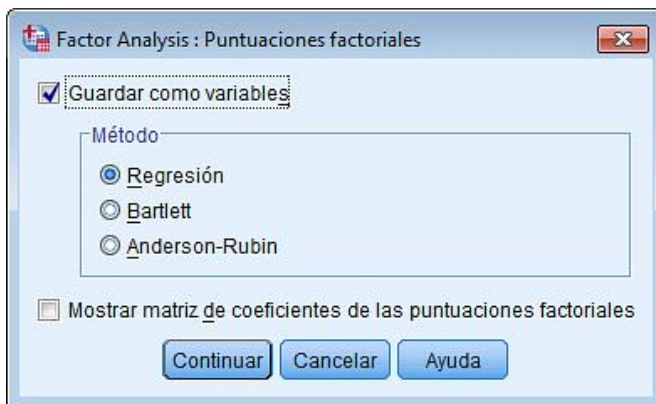


Figura 10.4. Ventana de Puntuaciones para Análisis Factorial en SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Entre las numerosas salidas que ofrece el paquete, se examinarán principalmente dos. La primera se refiere al total de variancia explicada (figura 10.5). Puesto que se reduce el número de variables al utilizar nuevos factores, no siempre se explica el 100% de la información o variabilidad original que presentaban los datos (es decir, la parsimonia tiene un costo). Siguiendo la tabla, la primera columna indica el número de cada componente o factor. Por lo general, se escogen únicamente

los factores con un autovalor (*eigenvalue*) mayor a 1. En este ejemplo se retienen, por lo tanto, solo los primeros cuatro componentes, pues son los únicos con autovalores mayores a 1.

Varianza total explicada									
Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3.322	30.197	30.197	3.322	30.197	30.197	2.027	18.430	18.430
2	1.575	14.320	44.517	1.575	14.320	44.517	1.919	17.444	35.874
3	1.118	10.160	54.677	1.118	10.160	54.677	1.546	14.056	49.930
4	1.020	9.275	63.952	1.020	9.275	63.952	1.542	14.022	63.952
5	.821	7.467	71.419						
6	.763	6.932	78.351						
7	.627	5.700	84.051						
8	.520	4.728	88.779						
9	.461	4.191	92.970						
10	.443	4.026	96.997						
11	.330	3.003	100.000						

Método de extracción: Análisis de Componentes principales.

Figura 10.5. Resultados de componentes (ejemplo de satisfacción con los servicios públicos).

Fuente: elaboración propia con base en el paquete SPSS.

De acuerdo con los resultados según la rotación, el primer factor explica un 18.4% de la variancia, el segundo factor un 17.4%, el tercero un 14.1% y el cuarto un 14.0%. Al sumar los porcentajes, se concluye que los factores capturan un 64.0% de la variancia total (ver también la columna final de porcentaje acumulado). El aporte de los factores restantes (con autovalores menores a 1) es menos significativo, por lo que no se muestra su variancia explicada y no se retienen (factores del 5 al 11).

El segundo resultado que se analizará es la matriz de componentes rotados (figura 10.6). Se habrá visto que el paquete ofrecía también una matriz de componentes sin rotar (que no se muestra aquí). Sin embargo, la solución rotada –en este caso mediante rotación Varimax, como se definió en el procedimiento– provee mejor interpretación de los componentes y la asignación de los nombres

para cada uno (Hernández, 2013, p. 137). Con la rotación no cambian ni los factores retenidos ni la comunalidad de cada variable; lo que sí se reajusta es el porcentaje de variancia explicada por cada uno de los factores (ver en la figura 10.5 cómo varían entre la columna de extracción y la de rotación, pero la variancia total explicada es la misma).

Esta matriz de componentes rotados está formada por las correlaciones entre las variables originales y los nuevos factores contruidos, que son las llamadas cargas factoriales. Al igual que el coeficiente de correlación lineal de Pearson (capítulo 5), las cargas más altas son las cercanas a 1 y -1 .

Para interpretar, se examina en cada factor cuáles son sus variables con cargas más altas. En el caso del primer componente, se observan cargas altas para las variables de atención en EBAIS, clínicas y hospitales. El segundo componente tiene cargas altas para el estado de calles, mantenimiento de carreteras y limpieza de alcantarillas. El tercer componente carga más en las variables de seguridad ciudadana y presencia de la fuerza pública. Finalmente, el cuarto componente muestra una carga alta en abastecimiento de agua, recolección de basura y suministro de energía eléctrica.

Siguiendo los conceptos ya vistos, es posible calcular la comunalidad de cada variable con base en las cargas factoriales. Por ejemplo, para la variable de agua,

$$\text{Comunalidad de agua} = (-0.111)^2 + (0.173)^2 + (0.071)^2 + (0.728)^2 = 0.577.$$

Las salidas de SPSS confirman este valor, pues proveen los valores de comunalidad para las once variables (figura 10.7).

Matriz de componentes rotados^a

	Componente			
	1	2	3	4
agua ¿Cómo califica el abastecimiento de agua en su comunidad?	-.111	.173	.071	.728
seguridad ¿Cómo califica la seguridad ciudadana en su comunidad?	.088	.009	.837	.256
policia ¿Cómo califica la presencia de la fuerza pública o policía en su comunidad?	.270	.271	.707	-.101
calles ¿Cómo califica el estado de las calles en su comunidad?	.072	.834	.026	.271
carreteras ¿Cómo califica el mantenimiento de carreteras que llevan a su comunidad?	.062	.861	.057	-.044
basura ¿Cómo califica los servicios de recolección de basura en su comunidad?	.382	.201	-.104	.592
electricidad ¿Cómo califica el suministro de energía eléctrica en su comunidad?	.241	-.011	.370	.649
alcantarillas ¿Cómo califica la limpieza de las alcantarillas en la comunidad?	.074	.570	.356	.256
ebais ¿Cómo califica la atención que recibe en el EBAIS?	.716	.088	.208	.100
clínicas ¿Cómo califica la atención que recibe en las clínicas de la Caja?	.783	.065	.124	.118
hospitales ¿Cómo califica la atención que recibe en los hospitales de la Caja?	.768	.020	.053	.006

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 5 iteraciones.

Figura 10.6. Matriz de componentes rotados (ejemplo de satisfacción con los servicios públicos).

Fuente: elaboración propia con base en el paquete SPSS.

Comunalidades

	Inicial	Extracción
agua ¿Cómo califica el abastecimiento de agua en su comunidad?	1.000	.577
seguridad ¿Cómo califica la seguridad ciudadana en su comunidad?	1.000	.775
policia ¿Cómo califica la presencia de la fuerza pública o policía en su comunidad?	1.000	.656
calles ¿Cómo califica el estado de las calles en su comunidad?	1.000	.776
carreteras ¿Cómo califica el mantenimiento de carreteras que llevan a su comunidad?	1.000	.750
basura ¿Cómo califica los servicios de recolección de basura en su comunidad?	1.000	.548
electricidad ¿Cómo califica el suministro de energía eléctrica en su comunidad?	1.000	.617
alcantarillas ¿Cómo califica la limpieza de las alcantarillas en la comunidad?	1.000	.522
ebais ¿Cómo califica la atención que recibe en el EBAIS?	1.000	.574
clínicas ¿Cómo califica la atención que recibe en las clínicas de la Caja?	1.000	.646
hospitales ¿Cómo califica la atención que recibe en los hospitales de la Caja?	1.000	.593

Método de extracción: Análisis de Componentes principales.

Figura 10.7. Comunalidades (ejemplo de satisfacción con los servicios públicos).

Fuente: elaboración propia con base en el paquete SPSS.

Obsérvese que las variables con cargas más altas (calles y carreteras) son aquellas que tienen también mayores comunalidades, es decir, variancia compartida. Además, menor comunalidad implica mayor unicidad (variancia específica). Así, siguiendo el caso de la variable de agua, la unicidad es $1 - 0.577 = 0.423$; por lo que, para esta variable, el 57.7% es variancia compartida y el 42.3% variancia específica. Resulta inmediato concluir que si una carga factorial fuese perfecta o igual a 1, entonces la comunalidad sería 1 (o del 100%) y la unicidad 0.

Una vez identificadas las cargas más altas de cada factor, es posible darles una interpretación sustantiva o conceptual. Así, el primer factor, correspondiente a clínicas, hospitales y EBAIS hace referencia a “servicios de salud”. El segundo, que cargó en estado de calles, mantenimiento de carreteras y limpieza de alcantarillas, se relaciona con “obras públicas”. El tercero se identifica fácilmente como “seguridad”. El cuarto y último factor indicó cargas altas en servicios de

electricidad, agua y recolección de basura, los cuales reciben las personas en sus propios hogares; por ello se le podría denominar “servicios domiciliarios”.

De esta forma, se logró pasar de once variables originales (una para cada pregunta sobre satisfacción con servicios públicos) a cuatro factores que resumen un 64% la variabilidad o información sobre el fenómeno de calificación de servicios públicos (cuadro 10.1).

Cuadro 10.1. Solución rotada y nombres de los factores

Variable	Factores			
	Servicios de salud	Obras públicas	Seguridad	Servicios domiciliarios
Clínicas de la Caja	0.783	0.065	0.124	0.118
Hospitales de la Caja	0.768	0.020	0.053	0.006
EBAIS	0.716	0.088	0.208	0.100
Mantenimiento de carreteras	0.062	0.861	0.057	-0.044
Estado de las calles	0.072	0.834	0.026	0.271
Limpieza de las alcantarillas	0.074	0.570	0.356	0.256
Seguridad ciudadana	0.088	0.009	0.837	0.256
Fuerza pública policía	0.270	0.271	0.707	-0.101
Energía eléctrica	0.241	-0.011	0.370	0.649
Abastecimiento de agua	-0.111	0.173	0.071	0.728
Recolección de basura	0.382	0.201	-0.104	0.592

Fuente: elaboración propia con base en la información de la figura 10.6.

Adicionalmente, SPSS guardó en la base de datos cuatro nuevas variables correspondientes a los cuatro factores creados, donde los valores se denominan “puntuaciones factoriales” (figura 10.8). Es decir, subsiguientes análisis, como de regresión, se pueden efectuar con base en estos factores, en lugar de utilizar las once variables originales.

	FAC1_1	FAC2_1	FAC3_1	FAC4_1
5	.29196	-1.22861	-.45292	1.09787
6	-.67124	.28297	.48233	-.92661
7	.53831	-.78982	-1.01086	.25138
8	1.02785	-1.32619	-.66775	.25332

Figura 10.8. Puntuaciones factoriales guardadas en la base de datos de SPSS.

Fuente: elaboración propia con base en el paquete SPSS.

Como se indicó anteriormente, gracias a la rotación Varimax los factores creados tienen correlación nula entre sí, lo cual es fácil de comprobar al generar una matriz de correlaciones de Pearson según el procedimiento visto en el capítulo 5 (ver figura 10.9).

Correlaciones					
		FAC1_1 REGR factor score 1 for analysis 1	FAC2_1 REGR factor score 2 for analysis 1	FAC3_1 REGR factor score 3 for analysis 1	FAC4_1 REGR factor score 4 for analysis 1
FAC1_1 REGR factor score 1 for analysis 1	Correlación de Pearson	1	,000	,000	,000
	Sig. (bilateral)		1,000	1,000	1,000
	N	166	166	166	166
FAC2_1 REGR factor score 2 for analysis 1	Correlación de Pearson	,000	1	,000	,000
	Sig. (bilateral)	1,000		1,000	1,000
	N	166	166	166	166
FAC3_1 REGR factor score 3 for analysis 1	Correlación de Pearson	,000	,000	1	,000
	Sig. (bilateral)	1,000	1,000		1,000
	N	166	166	166	166
FAC4_1 REGR factor score 4 for analysis 1	Correlación de Pearson	,000	,000	,000	1
	Sig. (bilateral)	1,000	1,000	1,000	
	N	166	166	166	166

Figura 10.9. Matriz de correlaciones entre los factores creados.

Fuente: elaboración propia con base en el paquete SPSS.

Ejemplo 2: las democracias según Lijphart

Arend Lijphart, en su ya clásico *Patterns of Democracy* (1999), comparó 36 democracias con base en distintos datos institucionales para examinar empíricamente si los modelos denominados mayoritario (o Westminster) y consensuales producían mejores resultados económicos, sociales y en calidad de la democracia; para ello contaba con 10 variables y argumentaba que podían simplificarse en dos dimensiones: (a) federal-unitaria y (b) ejecutivos-partidos (figura 10.9).

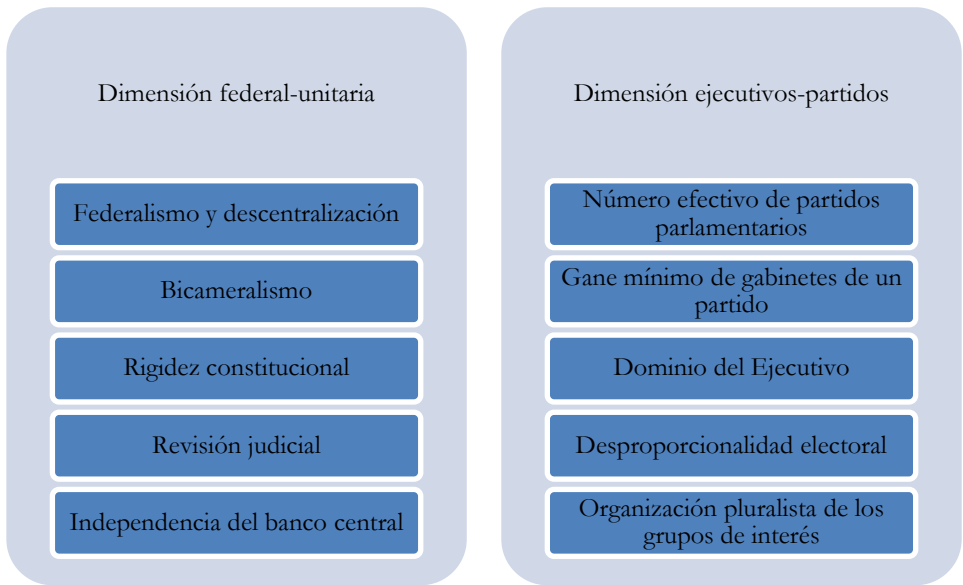


Figura 10.9. Dimensiones de las democracias según Lijphart.

Fuente: elaboración propia con base en Lijphart (1999).

Cuadro 10.2. Base de datos de Lijphart

PAÍS	PART	GAB	EJEC	DESPR	PLUR	FED	BICAM	RIG	REV	INDBC
ARG	3.15	82.40	8.00	17.98	2.70	4.50	4.00	4.00	2.70	0.39
AUL	2.22	80.70	9.10	9.44	2.12	5.00	4.00	4.00	3.00	0.42
AUT	2.68	43.30	8.07	2.51	0.38	4.50	2.00	3.00	3.00	0.55
BAH	1.69	100.00	9.44	16.48	3.00	1.00	2.00	3.00	2.00	0.41
BAR	1.68	100.00	8.87	17.27	2.20	1.00	2.00	2.00	2.00	0.38
BEL	4.72	37.30	2.57	3.35	1.15	3.50	2.80	3.00	1.80	0.27
BOT	1.38	100.00	9.90	14.61	2.60	1.00	2.50	2.00	2.00	0.33
CAN	2.52	88.40	8.10	11.56	3.25	5.00	3.00	4.00	3.40	0.52
CR	2.67	85.80	3.00	14.38	2.20	1.00	1.00	3.00	2.70	0.37
DEN	4.57	23.60	3.23	1.71	0.78	2.00	1.20	2.00	2.00	0.46
FIN	5.04	10.00	1.55	2.96	0.85	2.00	1.00	3.00	1.00	0.28
FRA	3.26	54.80	8.00	20.88	2.90	1.30	3.00	1.70	2.40	0.35
GER	3.09	37.80	3.80	2.67	0.88	5.00	4.00	3.50	4.00	0.69
GRE	2.27	98.10	4.45	7.88	3.12	1.00	1.00	2.00	2.00	0.38
ICE	3.72	46.30	3.20	3.85	2.20	1.00	1.40	1.00	2.00	0.34
IND	4.80	30.50	3.33	9.60	2.15	4.50	3.00	3.00	4.00	0.34
IRE	2.89	49.50	4.16	3.93	2.55	1.00	2.00	2.00	2.00	0.41
ISR	5.18	14.00	1.46	2.60	1.15	3.00	1.00	1.00	1.00	0.41
ITA	4.84	11.70	1.49	3.61	2.42	1.30	3.00	2.00	2.10	0.28
JAM	1.67	100.00	9.64	15.66	3.00	1.00	2.00	3.00	2.00	0.30
JPN	3.62	40.10	3.37	7.00	1.48	2.00	3.00	4.00	2.00	0.25
KOR	2.85	86.00	8.00	21.97	2.90	1.50	1.00	4.00	3.00	0.27
LUX	3.48	45.40	5.87	3.43	0.88	1.00	1.00	3.00	1.00	0.33
MAL	1.99	100.00	8.85	2.07	3.00	1.00	1.00	3.00	2.00	0.44
MAU	2.85	15.30	2.39	15.61	1.30	1.00	1.00	3.00	3.00	0.40
NET	4.87	26.80	2.91	1.21	0.98	3.00	3.00	3.00	1.00	0.48
NOR	3.64	55.30	4.04	4.53	0.38	2.00	1.50	3.00	2.00	0.17
NZ	2.28	81.40	4.54	9.25	2.68	1.00	1.10	1.00	1.00	0.21
POR	3.13	53.40	3.26	4.43	2.62	1.00	1.00	3.00	2.00	0.32
SPA	2.66	69.30	8.26	7.28	3.04	3.00	3.00	3.00	3.00	0.29
SWE	3.47	48.10	5.61	2.04	0.35	2.00	1.70	1.50	1.00	0.29
SWI	5.20	4.00	1.00	2.55	0.88	5.00	4.00	4.00	1.00	0.61
TRI	1.87	94.30	6.95	11.33	3.00	1.30	2.00	3.00	2.00	0.35
UK	2.16	97.30	8.12	11.70	3.02	1.20	2.50	1.00	1.00	0.31
URU	4.40	80.30	4.00	6.05	1.70	1.00	3.00	1.00	2.50	0.19
US	2.39	80.40	4.00	14.28	3.02	5.00	4.00	4.00	4.00	0.56

Notas: PART es el número efectivo de partidos; GAB el porcentaje de gabinetes con un partido ganador mínimo; EJEC el índice de dominio del ejecutivo; DESPR el índice de desproporcionalidad; PLUR el índice de pluralismos de los grupos de interés; FED el índice de federalismo; BICAM el índice de bicameralismo; RIG el índice de rigidez constitucional; REV el índice de revisión judicial; INDBC el índice de independencia del banco central. Para comprender la elaboración de estas escalas y su sustento teórico, ver Lijphart (1999).

Fuente: Lijphart (2013).

Una forma de identificar si las variables efectivamente corresponden a las dimensiones teóricas es realizando un análisis factorial.³⁸ Utilizando los datos actualizados para la segunda versión del libro (1945-2010)³⁹ que se presentan en el cuadro 10.2, se procede a ejecutar el análisis con las mismas especificaciones que el ejemplo 1, excepto que en la ventana de extracción (figura 10.10) ahora se declara un número fijo de factores por extraer: dos, correspondientes a las dimensiones teóricas propuestas por Lijphart.

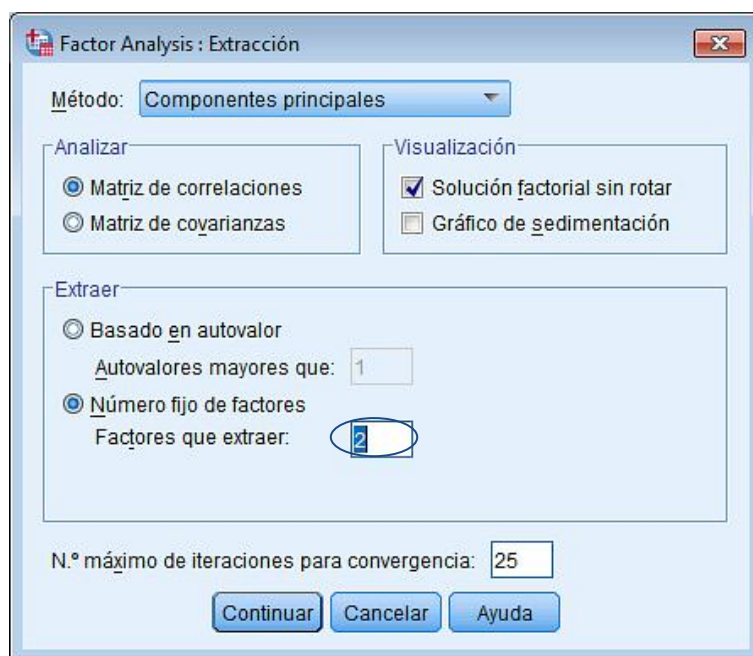


Figura 10.10. Ventana de Extracción para análisis factorial en SPSS (ejemplo de las democracias según Lijphart).

Fuente: elaboración propia con base en el paquete SPSS.

³⁸ Aunque en el ejemplo se prueba una teoría, conceptualmente no es propiamente un análisis confirmatorio pues este requiere no solo fijar el número de factores, sino también, el patrón de cargas factoriales y las relaciones de covariancia o independencia entre factores (Brown, 2006).

³⁹ Datos disponibles en la página oficial de Lijphart: <http://polisci.ucsd.edu/faculty/lijphart.html>.

En la figura 10.11, se observa que automáticamente el programa retiene solo los dos primeros componentes, ambos explican un 67% de la variabilidad total. Además, son los únicos factores con autovalores mayores a 1, lo cual indica que son factores relevantes.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3.802	38.023	38.023	3.802	38.023	38.023	3.802	38.022	38.022
2	2.917	29.174	67.197	2.917	29.174	67.197	2.918	29.175	67.197
3	.803	8.027	75.224						
4	.683	6.826	82.050						
5	.583	5.826	87.877						
6	.413	4.133	92.010						
7	.368	3.678	95.688						
8	.188	1.876	97.563						
9	.158	1.577	99.140						
10	.086	.860	100.000						

Método de extracción: Análisis de Componentes principales.

Figura 10.11. Resultados de componentes (ejemplo de las democracias según Lijphart).

Fuente: elaboración propia con base en el paquete SPSS.

Pasando directamente a la matriz de componentes rotados (figura 10.12), se encuentra que el primer componente tiene cargas altas para número efectivo de partidos parlamentarios, gane mínimo de gabinetes de un partido, dominio del Ejecutivo, desproporcionalidad electoral y organización pluralista de los grupos de interés. Es decir, este componente corresponde a la dimensión prevista por Lijphart como “ejecutivos-partido”. Por otro lado, las variables federalismo y descentralización, bicameralismo, rigidez constitucional, revisión judicial e independencia del banco central indican cargas altas en el segundo componente, es decir, en la dimensión “federal-unitaria”. De esta manera, se evidencian las dos dimensiones teóricas definidas por Lijphart y se confirman las correlaciones esperadas entre variables y factores.

Matriz de componentes rotados^a

	Componente	
	1	2
partidos	-.899	.026
gabinete	.920	-.096
ejecutivo	.839	.069
despropor	.779	.091
pluralismo	.807	-.019
federalismo	-.248	.893
bicameralismo	.019	.766
rigidezconsti	.095	.734
revisionjud	.322	.691
indepbancentral	-.107	.703

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 3 iteraciones.

Figura 10.12. Matriz de componentes rotados (ejemplo de las democracias según Lijphart).

Fuente: elaboración propia con base en el paquete SPSS.

Comentarios finales

- El capítulo trató el análisis de factorial de manera sucinta. Por ejemplo, se utilizaron componentes principales como método de extracción, pero también existe el método de máxima verosimilitud, entre otras alternativas. Asimismo, se ejecutó la rotación Varimax como una forma de mejorar la interpretación de los factores. Esta produce factores sin correlación entre sí, pero hay otras técnicas disponibles que sí admiten cierta correlación, denominada rotación oblicua.
- Así como los modelos de regresión se examinan con medidas de bondad de ajuste (coeficiente de determinación y tabla de clasificación), en análisis factorial se puede analizar el contraste de esfericidad de Bartlett y el coeficiente KMO para evaluar el ajuste de los resultados (ver Hernández, 2013). Este último es fácil de interpretar dado que toma valores entre 0 y 1, donde un KMO menor a 0.7 es insatisfactorio y cuanto más cercano a 1 es mejor el ajuste.

- Puede verse que el análisis produce los factores, pero no les da sentido. En el enfoque exploratorio, la interpretación, o ponerles “nombres” a los factores, debe estar fundamentada en la teoría o marcos conceptuales preexistentes (o al menos en el sentido o conocimiento común). El confirmatorio está, por definición, basado en teoría previa.
- Se debe tener un cuidado importante en cuanto a los valores perdidos. En el ejemplo 1, aunque la muestra total es de 545, esta se reduce a 166 unidades de observación, puesto que hay personas que no utilizan algún servicio, no lo califican y por ello son valores perdidos. Con el análisis factorial, basta con que falte un solo valor para una variable y esa observación ya se descarta del análisis. Se recomienda, por tanto, ahondar en métodos de sustitución de valores perdidos (el más simple consiste en reemplazar por la media, opción disponible en SPSS automáticamente).
- Finalmente, nótese que se utilizaron variables métricas o continuas, pero existen otros modelos factoriales para datos categóricos (o también, se puede recurrir a otras técnicas con similares objetivos como el análisis de correspondencias).

Ejercicios

1. Con los datos de noviembre de 2013 del CIEP, realice un análisis de factores con las calificaciones ciudadanas de las instituciones. Utilice las siguientes variables: *notagobierno*, *notaasamblea*, *notatse*, *notacontraloria*, *notasalacuarta*, *notapoderjudicial*, *notadefensoria*, *notaotj*, *notaguardiavil*, *notaucr*, *notauniversidadespub*, *notaiglesiakat* y *notaotraiglesias*. Interprete los factores según la solución rotada, póngales nombre e indique el porcentaje total de variancia explicada.
2. En el estudio *Respuestas ciudadanas ante el malestar con la política: salida, voz y lealtad* (Raventós *et al.*, 2012), se preguntó sobre la eficacia con que se perciben distintas formas de participación política (p. 87). Además, los autores realizan un análisis de factores. Con las cargas factoriales para cada variable presentadas a continuación (cuadro 10.2), identifique los factores e interpréte los.

Cuadro 10.2. Cargas factoriales para las formas de participación

Formas de participación	Factor 1	Factor 2	Factor 3
Ayudar en la campaña de un político	0.79	0.04	0.13
Bloquear carreteras en protestas	0.02	0.03	0.88
Denuncia en la Defensoría de los Habitantes	0.18	0.84	0.03
Firmar una carta a políticos planteando un problema	0.65	0.23	0.05
Participación en manifestaciones	0.21	0.16	0.79
Presentar un recurso ante la Sala IV	0.13	0.89	0.16
Reunirse con un político	0.82	0.04	0.06
Reunirse con una autoridad del gobierno	0.64	0.40	0.10

Fuente: Raventós *et al.* (2012).

Opcional. Calcule las comunalidades y unicidades para cada variable con los datos del ejercicio 2.

APÉNDICE A

FUENTES PARA PROFUNDIZAR

Como se señaló en los comentarios finales de cada capítulo, muchos temas quedaron por fuera de este material y los contenidos tratados se pueden incluso profundizar. A continuación, se exponen algunos de estos aspectos con el fin de motivar su estudio ulterior, sea a través de cursos o de forma autodidacta, para lo cual se precisan algunas fuentes accesibles, muchas de ellas con aplicaciones en ciencias sociales y políticas.

Primero, sobre la ciencia estadística en general se puede encontrar una introducción amena y llena de ejemplos en el libro *Naked Statistics* de Charles Wheelan (2013); para la historia de la disciplina y los aportes de eminentes estadísticos, examínese Salsburg (2001). Otra referencia general, que discute el tema del error, las predicciones y las consecuencias de las aplicaciones estadísticas es el éxito editorial de Nate Silver (2012), *The Signal and the Noise*.

Respecto a la inferencia estadística (capítulo 2), *Statistics. The Art and Science of Learning from Data* (Agresti y Franklin, 2013) ofrece una sólida y amigable introducción a la inferencia estadística y a la aplicación de numerosas pruebas, pero si se quiere un mayor rigor sobre la teoría estadística, puede consultarse a Wackerly, Mendenhall, y Scheaffer (2002) o cualquier texto de estadística matemática. También puede verse Hernández (2010) para aspectos de teoría de probabilidad e inferencia.

Los métodos de asociación y pruebas de hipótesis (capítulos 3, 4 y 5) conforman un enorme catálogo de coeficientes para relaciones entre variables con distintos niveles de medición. Por ejemplo, para evaluar la relación entre dos variables ordinales (donde hay un orden en las categorías), puede recurrirse a las llamadas estadísticas no paramétricas como el coeficiente de Spearman, la *tau* de Kendall y

otras. En Gutiérrez-Espeleta (2010) y en Sánchez (2005), se exponen las fórmulas de muchas de estas, pero el gran clásico en la materia es el de Agresti (2007).

Los modelos de regresión gaussiano (mínimos cuadrados ordinarios) y logístico (capítulos 6, 7 y 8) conllevan una gran serie de diagnósticos, medidas correctivas y especificaciones no vistas, para ello sírvase consultar Gujarati y Porter (2010) y Hosmer, Lemeshow y Sturdivant (2013). Pero, además, es posible modelar tanto variables métricas como categóricas asumiendo otras distribuciones de probabilidad; el modelo de Poisson para conteos (como pueden ser leyes aprobadas en una legislatura) es un ejemplo; pero desde el enfoque de los modelos lineales generalizados (consultar Dobson y Barnett, 2008), las opciones se expanden aún más (ver apéndice B).

Fenómenos con estructuras de datos particulares invitan a la utilización de otros modelos. Si las unidades de análisis están agrupadas en conglomerados (como cantones, provincias o países), es preferible utilizar modelos multinivel o jerárquicos (ver Gelman y Hill, 2007). Cuando lo que interesa es la ocurrencia de eventos y el tiempo para que sucedan, se entra al campo del análisis de sobrevivencia o modelos de historia de eventos (Box-Steffensmeier y Jones, 1997).

En el análisis multivariado (capítulos 9 y 10), existe una amplísima gama de técnicas con sus distintos algoritmos y opciones. Algunas de las no abarcadas son el análisis de correspondencias, el escalamiento multidimensional, el análisis discriminante y los árboles de segmentación (ver Díaz y Morales, 2012; Grimm y Yarnold, 1995; Hernández, 2013; Escobar, 2007).

Finalmente, se debe recordar que tan solo se trataron datos de tipo transversal. Para el modelaje de series de tiempo, Hernández (2011) ofrece una introducción a diversos métodos –incluidos los famosos modelos ARIMA–, mientras que para los datos de panel y longitudinales puede consultarse Frees (2004), así como el muy citado artículo de Beck y Katz (1995) que presenta una propuesta útil y simple para estimar modelos con datos transversales y temporales combinados.

APÉNDICE B

LOS MODELOS LINEALES GENERALIZADOS

Los métodos vistos en el libro pueden dar la idea de estar dispersos y desconectados entre sí. Sin embargo, este apéndice pretende sintetizar en un marco teórico la mayoría de técnicas antes vistas. La perspectiva que se adopta es la de los modelos lineales generalizados (MLG), introducidos por Nelder y Wedderburn (1972), precisamente con la intención de brindar un enfoque unificado a una diversidad de modelos para variables cuantitativas y cualitativas.

Para comprender el enfoque de los MLG, se debe tener en mente que los datos se pueden comportar según distintas distribuciones de probabilidad (en palabras simples, la forma o figura que adoptan al graficarlos en un histograma); la más conocida constituye la distribución normal o gaussiana para variables métricas o continuas. Con datos categóricos, pueden tenerse distribuciones como la binomial y la Poisson. Los MLG no solo aplican para estas distribuciones mencionadas, sino para muchas más, todas ellas dentro de una familia llamada exponencial.

Entre los MLG cuya variable dependiente sigue una distribución normal, se tiene el modelo de regresión lineal estimado por mínimos cuadrados ordinarios (capítulos 6 y 7) y el análisis de variancia (ANOVA) (capítulo 4). Asimismo, se puede vincular la prueba t (capítulo 3) con los MLG, en tanto esta es un caso particular del ANOVA con dos medias. Por su parte, el coeficiente de correlación de Pearson (capítulo 5) está íntimamente ligado con la regresión por mínimos cuadrados ordinarios (recuérdese que el coeficiente de determinación en regresión no es otra cosa sino el r de Pearson al cuadrado).

La regresión logística (capítulo 8) forma parte de los MLG en tanto se puede expresar como un modelo lineal con la función logito que enlaza las variables independientes con la dependiente; sus datos siguen una distribución binomial

(de 0 y 1). La prueba *chi* cuadrado (capítulo 5) tampoco es lejana de la perspectiva teórica de los MLG, ya que para una tabla de contingencia se puede asumir una distribución Poisson, la cual forma parte de la familia exponencial.

Finalmente, el análisis factorial (capítulo 10), 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 variables explicativas latentes que corresponden a los factores (Cooper, 1983).

La técnica de aglomeración jerárquica (capítulo 9) queda fuera del esquema de los MLG, pues, como se explicó en su respectivo capítulo, corresponde más a un algoritmo matemático que a un modelo estadístico.

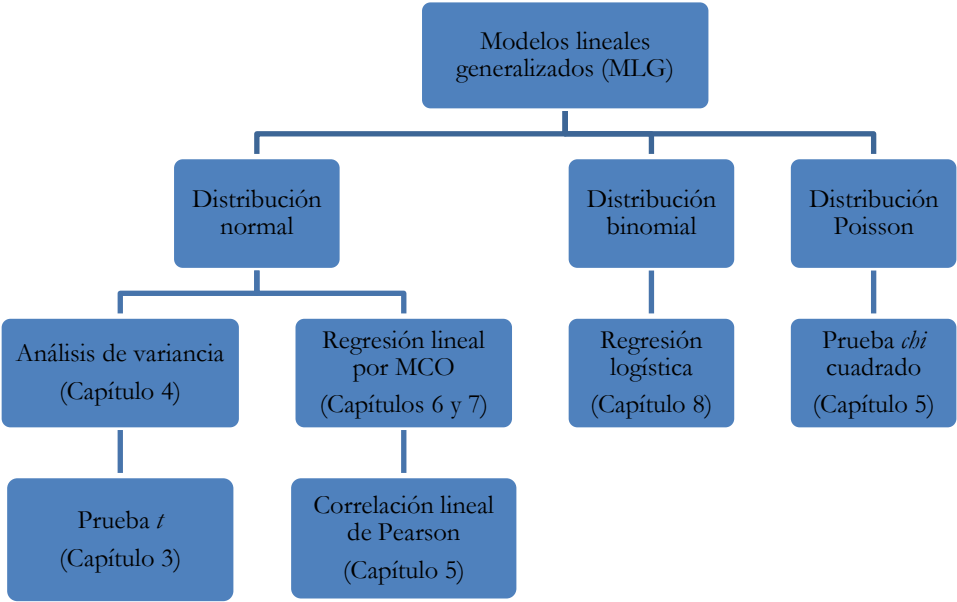


Figura B.1. Contenidos vistos desde el enfoque de los modelos lineales generalizados.

Fuente: elaboración propia.

RESPUESTAS A LOS EJERCICIOS

Capítulo 2

1. Con una confianza de 95%, el intervalo [310692 , 346684] contiene el valor real del ingreso per cápita promedio en Costa Rica en 2013. Es decir, que si se obtienen 100 muestras iguales, 95 de los 100 intervalos de confianza contendrían el valor real.
2. Con una confianza de 95%, el intervalo [19.72% , 21.68%] contiene el valor real del porcentaje de hogares pobres en Costa Rica. En otras palabras, con 100 muestras iguales, 95 de los 100 intervalos de confianza contendrían el valor real.
3. En el censo no se calculan márgenes de error porque no se realiza inferencia estadística. Los datos del censo son poblacionales (parámetros), mientras que los de la encuesta son estimaciones basadas en una muestra aleatoria.

Capítulo 3

1. Con un nivel de significancia (α) del 5%, la nota promedio dada a Laura Chinchilla por los hombres no es estadísticamente diferente de la nota otorgada por las mujeres (el valor p de la prueba es 0.055 y el intervalo de confianza para la diferencia contiene al cero).
2. Los promedios entre personas con primaria o menos y con universitaria son estadísticamente diferentes al 5% (el valor p es 0.001 y el intervalo de confianza para la diferencia no contiene al cero).
3. Las calificaciones promedio son estadísticamente diferentes entre jóvenes y adultos (el valor p de la prueba es 0.002 y el intervalo de confianza para la diferencia no contiene al cero).

Opcional. Se construye el intervalo $[-4.86 , 0.54]$ y se muestra que la diferencia de edades entre hombres y mujeres no es estadísticamente distinta de cero.

Capítulo 4

2. Con un 5% de significancia, se puede rechazar la hipótesis nula de que las medias son iguales (el valor p es 0.143). En otras palabras, personas en distintos grupos etarios muestran promedios similares en su valoración a la Asamblea Legislativa. Esto indica que la legitimidad política del parlamento no varía significativamente entre rangos de edades.
3. Con una significancia del 5%, el uso promedio de fuentes de información no es igual entre personas con distinto nivel educativo (el valor p es 0.000). Sin embargo, con el ANOVA no se puede decir en cuál nivel se usa más ni cuáles promedios por grupo son diferentes entre sí.
4. El resultado es el mismo porque se puede comprobar desde la teoría estadística que la prueba t es un caso particular del análisis de variancia.

Capítulo 5

1. Se rechaza la hipótesis nula de independencia entre las variables ($p = 0.000$), es decir, que el voto en 2010 y la intención de voto en 2014 están relacionados. El coeficiente V de Cramer muestra una asociación moderada (0.249). De manera que, aunque hay una relación entre el comportamiento electoral previo y la intencionalidad futura del sufragio, no es determinística (no todos piensan repetir su comportamiento).
2. Entre mayor edad, mayor nota otorgada a Johnny Araya. La relación es baja pero significativa. Entre menor edad, mayor nota a José María Villalta (relación negativa pero bajo). La relación entre edad y nota a Luis Guillermo Solís es muy baja y prácticamente nula.

Capítulo 6

El modelo estimado es:

$$TAS\widehat{A}EXITO_i = 16.337 + 1.095LEGIS_i.$$

Por cada punto porcentual de legisladores con que cuenta el Presidente, su tasa de éxito legislativo aumenta 1.1 puntos porcentuales, en promedio. Este

coeficiente es significativo al 5%. Se explica un 38% de la variabilidad de la tasa de éxito ($R^2 = 0.377$).

Capítulo 7

1. El modelo estimado es el siguiente:

Variable	Coeficiente	Significancia
Constante	52.655	0.000
EDAD	0.001	0.987
SEXO	1.012	0.534
OCUPADO	0.680	0.693
EDUCACION	-1.558	0.002
SIMPATIZA	4.706	0.006
VOTOLAURA2010	4.093	0.013

2. Las variables nivel educativo, simpatía partidaria y votó por Laura Chinchilla son estadísticamente significativas al 5%. Sexo, edad y estado de ocupación no son significativamente distintas de cero.

Por cada nivel educativo alcanzado, la valoración de los poderes disminuye 1.5 puntos en promedio, con todas las variables constantes. Entre los simpatizantes de algún partido político, la calificación de los poderes es 4.7 puntos mayor en promedio, con las demás variables constantes. Si votó por Laura Chinchilla en 2010, la calificación de los poderes aumenta 4.1 puntos en promedio, constantes las demás variables.

La constante se interpreta como la nota promedio entre hombres con cero años de edad, sin estudios, que no tienen simpatías partidarias y que no votaron por Laura Chinchilla.

Se explica un 3.5% de la variabilidad en las calificaciones a los poderes estatales (R^2 ajustado = 0.035).

Capítulo 8

1. El modelo estimado es el siguiente:

Variable	Coefficiente	Significancia
SEXO	-0.135	0.592
EDAD	-0.005	0.546
EDUCACION	0.175	0.038
SIMPATIZA	2.501	0.000
VOTO2010	1.673	0.000
Constante	-0.401	0.436

2. Las variables educación, simpatía partidaria y voto en 2010 son significativas al 5%. Sexo y edad no son significativas.

Al aumentar el nivel de educación, el chance de pensar ir a votar en 2014 se incrementa. Entre personas con simpatía partidaria, el chance de pensar ir a votar es mayor que entre personas sin simpatías. Entre personas que votaron en 2010, el chance de pensar ir a votar en 2014 es mayor que entre personas que no votaron en 2010.

El porcentaje global de clasificación correcta es 84.7%.

3. Se estima que una persona con esas características tiene una probabilidad de 0.89 de pensar ir a votar.

Capítulo 9

Como la tasa de reelección y el número efectivo de partidos tienen distintas escalas de medición, el análisis de conglomerados se hizo con variables estandarizadas. El enlace completo forma los grupos más homogéneos entre sí.

- *Grupo 1:* Honduras, Uruguay, Venezuela, Perú, Nicaragua, Paraguay, República Dominicana, Panamá y Bolivia. Países con niveles bajos o intermedios de reelección, pero con pocos partidos políticos en el parlamento.
- *Grupo 2:* Ecuador y Guatemala. Son países con baja reelección, pero mayor número de partidos respecto al primer grupo.

- *Grupo 3:* Colombia y Brasil. Son países con altos porcentajes de reelección y el mayor número partidos políticos en el parlamento entre el conjunto de casos.
- *Grupo 4:* El Salvador, Argentina y Chile. Se caracteriza por alto porcentaje de reelección parlamentaria, pero un número bajo de partidos políticos en el parlamento.

Capítulo 10

1. Se producen cuatro factores que explican el 69% de la variancia.

- *Primer factor:* notagobierno, notaasamblea, notatse, notacontraloria y notasalacuarta.
- *Segundo factor:* notapoderjudicial, notadefensoria, notaotj y notaguardiacivil.
- *Tercer factor:* notaucr y notauniversidadespub.
- *Cuarto factor:* notaiglesiakat y notaotraiglesias.

BIBLIOGRAFÍA

Acock, Alan C. y Stavig, Gordon R. (1979). A Measure of Association for Nonparametric Statistics. *Social Forces*, 57(4), 1381-1386.

Agresti, Alan. (2007). *An Introduction to Categorical Data Analysis*. New Jersey: Wiley.

Agresti, Alan y Franklin, Christine. (2013). *Statistics. The Art and Science of Learning from Data*. Boston: Pearson.

Aldenderfer, Mark S. y Blashfield, Roger K. (1984). *Cluster Analysis. Series: Quantitative Applications in the Social Sciences*. California: Sage.

Almond, Gabriel A. (1999). *Una disciplina segmentada. Escuelas y corrientes en las ciencias políticas*. México: 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.

Beck, Nathaniel y Katz, Jonathan N. (1995). What to do (and not to do) with Time-Series Cross-Section Data. *The American Political Science Review*, 89(3), 634-647.

Benoit, Kenneth. (2004). Models of electoral system change. *Electoral Studies*, 23, 363-389.

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.

Beach, Derek y Pedersen, Rasmus Brun. (2013). *Process-Tracing Methods. Foundations and Guidelines*. Ann Arbor: The University of Michigan Press.

Brady, Henry E. (2008). Causation and Explanation in Social Science. En Janet Box-Steffensmeier, Henry E. Brady y David Collier (eds.), *The Oxford Handbook of Political Methodology*. New York: Oxford University Press.

Brady, Henry E. y Collier, David (editores). (2010). *Rethinking Social Inquiry. Diverse Tools, Shared Standards*. Lanham: Rowman & Littlefield Publishers.

Brown, Timothy A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press.

Bryant, Fred B. y Yarnold, Paul R. (1995). Principal-Components Analysis and Exploratory and Confirmatory Factor Analysis. En Laurence G. Grimm y Paul R. Yarnold (eds.), *Reading and Understanding Multivariate Statistics*. Washington D. C.: American Psychological Association.

Bunge, Mario. (1999). *Buscar la filosofía en las ciencias sociales*. México D. F.: Siglo XXI.

Catellani, Patrizia y Alberici, Augusta Isabella. (2012). Does the Candidate Matter? Comparing the Voting Choice Early and Late Deciders. *Political Psychology*, 33(5), 619-634.

CIEP. (2012-2014). Encuestas del proyecto Estudios de Opinión Sociopolítica. Centro de Investigación y Estudios Políticos, Universidad de Costa Rica.

Cooper, John C. B. (1983). Factor Analysis: An Overview. *The American Statistician*, 37(2), 141-147.

Cox, David R. (1982). Statistical Significance Tests. *British Journal of Clinical Pharmacology*, 14, 325-331.

Crespi, Irving. (2000). *El proceso de opinión pública. Cómo habla la gente*. Barcelona: Ariel.

Creswell, John W. (2009). *Research Design. Qualitative, Quantitative, and Mixed Methods Approaches*. California: Sage.

Davidian, Marie y Louis, Thomas A. (2012). Why Statistics? *Science*, 6, 12.

della Porta, Donatella y Keating, Michael (eds.). (2008). *Approaches and Methodologies in the Social Sciences. A Pluralist Perspective*. New York: Cambridge University Press.

Díaz Monroy, Luis Guillermo y Morales Rivera, Mario Alfonso. (2012). *Análisis estadístico de datos multivariados*. Bogotá: Universidad Nacional de Colombia.

Dobson, Annette J. y Barnett, Adrian G. (2008). *An Introduction to Generalized Linear Models*. Boca Raton: Chapman & Hall/CRC Press.

Duverger, Maurice. (1957). *Los partidos políticos*. México: FCE.

Escobar Mercado, Modesto. (2007). *El análisis de segmentación: técnicas y aplicaciones de los árboles de clasificación*. Madrid: CIS.

Franzese, Robert J. (2007). Multicausality, Context-Conditionality, and Endogeneity. En Carles Boix y Susan C. Stokes (eds.), *The Oxford Handbook of Comparative Politics*. New York: Oxford University Press.

Frees, Edward W. (2004). *Longitudinal and Panel Data. Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press.

García Montero, Mercedes. (2009). *Presidentes y Parlamentos: ¿quién controla la actividad legislativa en América Latina*. Madrid: CIS.

Geddes, Barbara. (2003). *Paradigms and Sand Castles. Theory Building and Research Design in Comparative Politics*. Ann Arbor: The University of Michigan Press.

Gelman, Andrew y Hill, Jennifer. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, Andrew; Carlin, John B.; Stern, Hal S. y Rubin, Donald, B. (2004). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall/CRC.

Geys, Benny. (2006). Explaining voter turnout: A review of aggregate-level research. *Electoral Studies*, 25, 637-663.

Glasgow, Garrett y Alvarez, R. Michael. (2008). Discrete Choice Methods. En Janet Box-Steffensmeier, Henry E. Brady y David Collier (eds.), *The Oxford Handbook of Political Methodology*. New York: Oxford University Press.

Green, Donald P. y Gerber, Alan S. (2003). The Underprovision of Experiments in Political Science. *Annals of the American Academy of Political and Social Science*, 589, 94-112.

Grimm, Laurence G. y Yarnold, Paul R. (eds.). (1995). *Reading and Understanding Multivariate Statistics*. Washington D. C.: American Psychological Association.

- Gujarati, Dadomar y Porter, Dawn C. (2010). *Econometría*. México: McGraw-Hill.
- Gutiérrez-Espeleta, Édgar. (2010). *Métodos estadísticos para las ciencias biológicas*. Heredia: UNA.
- Hernández, Óscar. (2004). Costa Rica. En John G. Geer (ed.), *Public Opinion and Polling Around the World. A Historical Encyclopedia*. California: ABC-CLIO.
- Hernández, Óscar. (2010). *Elementos de probabilidades e inferencia estadística para Ciencias Sociales*. San José: Editorial UCR.
- Hernández, Óscar. (2011). *Introducción a las Series Cronológicas*. San José: Editorial UCR.
- Hernández, Óscar. (2012). *Estadística elemental para Ciencias Sociales*. San José: Editorial UCR.
- Hernández, Óscar. (2013). *Temas de análisis estadístico multivariante*. San José: Editorial UCR.
- Hosmer, David W.; Lemeshow, Stanley y Sturdivant, Rodney X. (2013). *Applied Logistic Regression*. New Jersey: Wiley.
- IDEA. (2014). Voter turnout [en línea]. International Institute for Democracy and Electoral Assistance [citado el 22/3/2014]. Disponible en: <http://www.idea.int/vt/index.cfm>
- INEC. (2011). X Censo Nacional de Población y VI de Vivienda: Resultados Generales. Instituto Nacional de Estadística y Censos. San José: INEC.
- INEC. (2013). Encuesta Nacional de Hogares Julio 2013: Resultados Generales. Instituto Nacional de Estadística y Censos. San José: INEC.
- IPU. (2013). Woman in national parliaments [en línea]. Inter-Parliamentary Union [citado el 23/6/2013]. Disponible en: <http://www.ipu.org/wmn-e/arc/classif010113.htm>
- Jackman, Simon. (2000). Estimation and Inference via Bayesian Simulation. *American Journal of Political Science*, 44(2), 375-404.
- Jackman, Simon. (2004). Bayesian Analysis for Political Research. *Annual Review of Political Science*, 7, 483-505.
- Kahneman, Daniel. (2012). *Pensar rápido, pensar despacio*. México, D. F.: 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*. New Jersey: Princeton University Press.
- Laakso, Markus y Taagepera, Rein. (1979). "Effective" Number of Parties. A Measure with Application to West Europe. *Comparative Political Studies*, 12(1), 3-27.
- Lawson, Chappell; Lenz, Gabriel S.; Baker, Andy y Myers, Michael. (2010). Looking Like a Winner: Candidate Appearance and Electoral Success in New Democracies. *World Politics*, 62(4), 561-593.
- Lieberman, Evan S. (2007). Nested Analysis as a Mixed-Method Strategy for Comparative Research. *The American Political Science Review*, 99(3), 435-452.
- Lindley, Dennis. (2000). The Philosophy of Statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 49(3), 293-337.
- Lijphart, Arend. (1999). *Patterns of Democracy. Government Forms and Performances in Thirty-Six Countries*. New York: Yale University Press.
- Lijphart, Arend. (2013). Detailed (disaggregated) data [en línea]. UC San Diego [consultada 29/6/2013]. Disponible en: <http://polisci.ucsd.edu/faculty/lijphart.html>
- Lipset, Seymour Martin. (1959). Some Social Requisite of Democracy: Economic Development and Political Legitimacy. *The American Political Science Review*, 53(1), 69-105.
- Mahoney, James y Goertz, Gary. (2006). A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research. *Political Analysis*, 14, 227-249.
- Maoz, Zeev y Abdolali, Nasrin. (1989). Regime Types and International Conflict, 1816-1976. *The Journal of Conflict Resolution*, 33(1), 3-35.
- Martínez Franzoni, Juliana. (2008). *Domesticar la incertidumbre en América Latina. Mercado laboral, política social y familias*. San José: Editorial UCR.
- Marshall, Monty G. y Cole, Benjamin R. (2011). *Global Report 2011. Conflict, Governance, and State Fragility*. Virginia: Center for Systemic Peace.

Massey, Douglas S. (1987). The Ethnosurvey in Theory and Practice. *International Migration Review*, 21(4), 1498-1522.

Moore, Barrington. (1966). *Social Origins of Dictatorship and Democracy. Lord and Peasant in the Making of the Modern World*. Boston: Beacon Press.

Monroe, Kristin Renwick. (2007). The Perestroika Movement, its Methodological Concerns, and the Professional Implications of These Methodological Issues. *Qualitative Methods*, 5(1), 2-6.

Mora Salas, Minor y Pérez Sáinz, Juan Pablo. (2009). *Se acabó la Pura Vida. Amenazas y desafíos sociales en la Costa Rica del Siglo XXI*. San José: FLACSO.

Morton, Rebecca B. y Williams, Kenneth C. (2008). Experimentation in Political Science. En Janet Box-Steffensmeier, Henry E. Brady y David Collier (eds.), *The Oxford Handbook of Political Methodology*. New York: Oxford University Press.

Nelder, John A. y Wedderburn, Robert W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370-384.

Norris, Pippa. (2004). *Electoral Engineering. Voting Rules and Political Behavior*. Cambridge: Cambridge University Press.

OIR. (2014a). Observatorio del Poder Legislativo en América Latina [en línea]. Observatorio de Instituciones Representativas de América Latina, Universidad de Salamanca [citado el 11/1/2014]. Disponible en: <http://americo.usal.es/oir/opal/>

OIR. (2014b). Observatorio de Partidos Políticos de América Latina [en línea]. Observatorio de Instituciones Representativas de América Latina, Universidad de Salamanca [citado el 11/1/2014]. Disponible en: <http://americo.usal.es/oir/legislatina/>

Pignataro, Adrián. (2012). *La participación electoral en países de América Latina: un modelo desde la teoría de la elección racional*. Tesis de Licenciatura en Ciencias Políticas. Universidad de Costa Rica.

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.

Przeworski, Adam. (2007). Is the Science of Comparative Politics Possible? En Carles Boix y Susan C. Stokes (eds.), *The Oxford Handbook of Comparative Politics*. New York: Oxford University Press.

Przeworski, Adam; Álvarez, Michael E.; Cheibub, José Antonio y Limongi, Fernando. (2000). *Democracy and Development. Political Institutions and Well-Being in the World, 1950-1990*. New York: Cambridge University Press.

Putnam, Robert D. (1993). *Making Democracy Work. Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Ragin, Charles C. (1987). *The Comparative Method. Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press.

Ramírez Moreira, Olman (ed.). (2010). *Comportamiento del electorado costarricense. Elecciones del 2006*. San José: Editorial UCR.

Raventós Vorst, Ciska; Fournier Facio, Marco Vinicio; Fernández Montero, Diego y Alfaro Redondo, Ronald. (2012). *Respuestas ciudadanas ante el malestar con la política: salida, voz y lealtad*. San José: Instituto de Formación y Estudios en Democracia.

Riquelme, José C., Ruiz, Roberto y Gilbert, Karina. (2006). Minería de Datos: Conceptos y Tendencias. *Inteligencia Artificial*, 10(29), 11-18.

Salsburg, David. (2001). *The Lady Tasting Tea. How Statistics Revolutionized Science in the Twentieth Century*. New York: A W. H. Freeman/ Holt Paperback.

Sánchez Ramos, Miguel Ángel. (2005). Uso metodológico de las tablas de contingencia en la Ciencia Política. *Espacios Públicos*, 8(16), 60-84.

Sartori, Giovanni. (2004). ¿Hacia dónde va la ciencia política? *Política y Gobierno*, 11(2), 349-354.

Schmitter, Philippe C. (2009). The nature and future of comparative politics. *European Political Science Review*, 1(1), 33-61.

Shively, W. Phillips. (2011). *The Craft of Political Research*. Boston: Longman.

Silver, Nate. (2012). *The Signal and the Noise. Why So Many Predictions Fail – but Some Don't*. New York: Penguin Press.

Stigler, Stephen M. (1978). Mathematical Statistics in the Early States. *The Annals of Statistics*, 6(2), 239-265.

Souva, Mark y Rohde, David. (2007). Elite Opinion Differences and Partisanship in Congressional Foreign Policy, 1975-1996. *Political Research Quarterly*, 60(1), 113-123.

SPSS Inc. (2007). *SPSS Statistics Base 17.0. User's Guide*. Chicago: SPSS Inc.

Stepan, Alfred y Skach, Cindy. (1994). Presidentialism and Parliamentarism Compared. En Juan Linz y Arturo Valenzuela (eds.), *The Failure of Presidential Democracy. Volume 1. Comparative Perspectives*. Baltimore: The Johns Hopkins University Press.

Treminio Sánchez, Ilka. (2010). Ensayo y error: la puesta en práctica de la democracia directa en Costa Rica. *Revista Centroamericana de Ciencias Sociales*, 7(1), 123-153.

Vargas-Cullell, Jorge y Rosero-Bixby, Luis. (2006). *Cultura política de la democracia en Costa Rica: 2006*. San José: LAPOP.

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*. Chicago y London: The University of Chicago Press.

Wackerly, Dennis D.; Mendenhall, William y Scheaffer, Richard L. (2002). *Estadística matemática con aplicaciones*. México: Thomson.

Wasserstein, Ronald L. y Lazar, Nicole A. (2016). The ASA's statement on p-values: contexts, process, and purpose. *The American Statistician*, 70(2), 129-133.

Wheelan, Charles. (2013). *Naked Statistics. Stripping the Dread from the Data*. New York: W.W. Norton & Company.

Wolfinger, Raymond E. y Rosenstone, Steven J. (1980). *Who Votes?* New Haven y London: Yale University Press.

Wooldridge, Jeffrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press

Wright, Raymond E. (1995). Logistic Regression. En Laurence G. Grimm y Paul R. Yarnold (eds.), *Reading and Understanding Multivariate Statistics*. Washington D. C.: American Psychological Association.

Yates, Frank y Mather, Kenneth. (1963). Ronald Aylmer Fisher. 1890-1962. *Biographical Memoirs of Fellows of the Royal Society*, 9, 91-129.

ACERCA DEL AUTOR

Profesor de la Escuela de Ciencias Políticas de la Universidad de Costa Rica e investigador del Centro de Investigación y Estudios Políticos (CIEP) de la misma institución. Correo electrónico: adrian.pignataro@gmail.com.