Introducción al análisis multivariante

PID_00263801

Julio Meneses

Tiempo mínimo de dedicación recomendado: 4 horas

Universitat Oberta de Catalunya

Julio Meneses

Profesor agregado de Metodología de la investigación de los Estudios de Psicología y Ciencias de la Educación, investigador de Internet Interdisciplinary Institute (IN3) y responsable de la Unidad de Evaluación de Proyectos Institucionales del eLearn Center de la Universitat Oberta de Catalunya (UOC).

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Julio Meneses (2019)

Primera edición: septiembre 2019

Avda. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

«You can't fix by analysis what you bungled by design».

Light, Singer y Willet, 1990, p. viii.

El análisis multivariante puede contribuir a enriquecer el debate público sobre los fenómenos que son objeto de interés para los profesionales y los investigadores, gracias a la oportunidad que les ofrece para llevar a cabo un análisis complejo de los datos obtenidos en sus estudios. Al servicio de la investigación cuantitativa, y como extensión de las técnicas de análisis univariante y bivariante, el análisis multivariante tiene como objetivo principal modelar las múltiples relaciones existentes entre diversas variables de manera simultánea.

La construcción de modelos multivariantes ejerce, pues, un papel importante en el desarrollo de las diferentes disciplinas basadas en el análisis de datos cuantitativos y requiere, por lo tanto, una atención especial en la formación de futuros profesionales e investigadores. Conocer la lógica, las características específicas de las diferentes técnicas disponibles, los objetivos particulares que permiten lograr y las condiciones en que pueden ser utilizadas son algunos de los retos importantes a los que nos enfrentaremos en este material.

Para hacerlo, en este texto nos adentraremos en los aspectos básicos involucrados en el análisis multivariante de los datos como el marco analítico general que se propone analizar e interpretar las relaciones simultáneas entre diversas variables mediante la construcción de modelos estadísticos complejos que permiten distinguir la contribución independiente de cada una de ellas en el sistema de relaciones para, de este modo, describir, explicar o predecir los fenómenos que son objeto de interés.

La clave de este marco analítico general no se encuentra, por lo tanto, en el hecho de que los investigadores dispongan de múltiples variables, sino en la capacidad que las diferentes técnicas disponibles les ofrecen para estimar el peso específico o la importancia relativa de cada una de ellas en sus modelos. En este sentido, como veremos, el análisis multivariante puede proporcionar las evidencias necesarias que permitan establecer inferencias a partir de la observación de asociaciones entre las variables, de forma que sea posible extraer conclusiones no sesgadas que, además, sean generalizables más allá de los límites de los estudios particulares siempre que sea posible.

Este no es un objetivo menor y, de hecho, está íntimamente relacionado con la naturaleza del diseño utilizado en la investigación a partir de la que se han obtenido los datos. Es por esta razón que, teniendo en cuenta las palabras de Light, Stinger y Willet (1990), los investigadores no tienen que recurrir a las potencialidades que ofrece el análisis multivariante para intentar resolver

los problemas eventuales que puedan surgir en el supuesto de que la investigación no haya sido correctamente diseñada o desarrollada. Al contrario, la construcción de modelos multivariantes adquiere todo el sentido en relación con el procedimiento general establecido en la investigación cuantitativa que, en último término, es el que permite que los investigadores dispongan de las garantías suficientes para decidir si las múltiples asociaciones simultáneas observadas entre las variables son una evidencia adecuada para determinar, con una cierta confianza, la existencia de relaciones extrapolables al conjunto de la población que representa su muestra de participantes.

Teniendo en cuenta estas consideraciones, empezaremos la exposición tomando como punto de partida un estudio clásico sobre la discriminación por razón de género en la Universidad de Berkeley. La discusión de este caso controvertido nos servirá para introducir algunos conceptos importantes, como son la asociación, la confusión y la causalidad, reconocer explícitamente la importancia del diseño de la investigación para extraer conclusiones no sesgadas que sean generalizables, y ofrecer una definición formal que nos permita situar el análisis multivariante como el marco analítico general que permite modelar las múltiples relaciones existentes entre las diferentes variables involucradas en una investigación de manera simultánea.

Una vez establecidos estos fundamentos, desarrollaremos las implicaciones de la definición describiendo los objetivos principales y presentaremos una clasificación general de las diferentes técnicas disponibles que nos servirá para ofrecer una panorámica general sobre sus características y las condiciones en que pueden ser utilizadas. De este modo, los lectores interesados dispondrán de una guía que les permitirá escoger la técnica que mejor se ajuste a su investigación y, a continuación, ofreceremos una discusión de algunos de los principios que, en el contexto de la investigación cuantitativa, rigen las diferentes fases con que es posible estructurar el proceso de construcción de modelos multivariantes.

La recapitulación de estas fases nos servirá, en última instancia, para proporcionar una perspectiva de conjunto sobre las cuestiones más importantes introducidas a lo largo del texto. Finalmente, concluiremos esta introducción general con una bibliografía anotada que servirá de ayuda para complementar y ampliar nuestra aproximación a los aspectos básicos del análisis multivariante.

1. El caso de la discriminación de género en la Universidad de Berkeley

El año 1973 fue interesante para la discusión sobre la situación de las mujeres en el mundo universitario en los Estados Unidos. Resueltas las solicitudes de acceso para el comienzo del curso, la Universidad de Berkeley llevó a cabo una investigación interna para determinar si había indicios fundados sobre la existencia de una discriminación por razón de género en el acceso de los estudiantes a los programas de posgrado. En este sentido, examinando los datos recogidos a los archivos de los diferentes departamentos, el profesor Hammel, entonces decano de estos estudios, se encontró con una situación, cuando menos, aparentemente paradójica (Bickel, Hammel y O'Connell, 1975).

Teniendo en cuenta el conjunto global de solicitudes, en aquel curso se presentaron un total de 12.763 candidatos, de los cuales 8.442 fueron hombres y 4.321 mujeres. De estos candidatos, aproximadamente un 44 % de los hombres y un 35 % de las mujeres fueron finalmente admitidos para iniciar sus estudios de posgrado. La tabla 1 recoge estos datos, desagregando las candidaturas admitidas y rechazadas en función del género de los solicitantes, y permite ilustrar las conclusiones preliminares de esta investigación.

Tabla 1. Resolución sobre las solicitudes de acceso a los programas de posgrado de la Universidad de Berkeley según el género de los candidatos (otoño de 1973)

	Solicitudes	Admisiones	Rechazos	Porcentaje de admisión
Hombres	8.442	3.738	4.704	44,28 %
Mujeres	4.321	1.494	2.827	34,58 %
Total	12.763	5.232	7.531	40,99 %

Fuente: Bickel, Hammel y O'Connell (1975).

En efecto, teniendo en cuenta que la tasa global de aceptación en el conjunto de los departamentos fue de un 41 % aproximadamente, la diferencia de casi 10 puntos entre los hombres y las mujeres sería una evidencia a favor de la existencia de una discriminación por razón de género. De hecho, si utilizamos esta tabla de contingencia para analizar su asociación, podemos afirmar que existe una relación estadísticamente significativa entre el género de los candidatos y su aceptación final en los programas de posgrado de la Universidad de Berkeley (X^2 = 111,25, df = 1, p < 0,001). Pero a pesar de ser estadísticamente significativa, esta relación no muestra una intensidad o una magnitud importante (V de Cramér = 0,09).

Si asumimos, y no tenemos evidencias para no hacerlo así, que las mujeres y los hombres no difieren significativamente en sus capacidades, aptitudes y habilidades, la Universidad de Berkeley preferiría a los hombres antes que a las mujeres como estudiantes de los programas de posgrado. Pero esta situación resulta más compleja que la representación que ofrece el análisis de esta tabla de contingencia.

Tal como mostraron Bickel, Hammel y O'Connell (1975), la discriminación aparente por razón de género se produciría únicamente cuando agregamos los datos para el conjunto de la Universidad. A pesar de que en su trabajo no reproducen los datos proporcionados para cada uno de los ciento un departamentos que ofrecían estos estudios, su análisis sirve como interesante ilustración de una relación espuria entre el género de los candidatos y su aceptación final.

Descartando los registros de los departamentos que no recibieron ninguna solicitud por parte de ninguna mujer o que, finalmente, no rechazaron a ningún candidato, identificaron cuatro de los ochenta y cinco departamentos restantes que, efectivamente, mostraban una preferencia estadísticamente significativa por los hombres. En cambio, seis de estos mismos ochenta y cinco departamentos resolvieron sus solicitudes en el sentido contrario, mostrando una preferencia estadísticamente significativa por las mujeres. Es más, examinando las tablas de contingencia de estos diez departamentos que mostraban una preferencia, por los hombres o por las mujeres, su conclusión fue que la discriminación por razón de género en el acceso a los estudios de posgrado afectaba, en realidad, más a los hombres que a las mujeres.

Pero, dada una relación estadísticamente significativa entre el género de los candidatos y su aceptación en el conjunto de la universidad a favor de los hombres, ¿cómo es posible que una gran mayoría de los departamentos de Berkeley no mostrara ninguna preferencia y que, teniendo en cuenta la minoría que lo hacía por los hombres o por las mujeres, esta discriminación por razón de género afectara más a los hombres que a las mujeres? Freedman, Pisani y Purves (2007) ofrecen una aproximación complementaria que nos puede ayudar a entender esta aparente contradicción.

Tomando en consideración los datos proporcionados por los seis departamentos más grandes, que habían evaluado aproximadamente un tercio de los candidatos de toda la universidad, registraron el número de solicitudes y calcularon las respectivas tasas de admisión. La tabla 2 recoge estos datos, desagregando las solicitudes en función del género de los candidatos.

Tabla 2. Datos de admisión a los seis departamentos más grandes de la Universidad de Berkeley según el género de los candidatos (otoño de 1973)

	Hombres		**Mujeres**	
Departamento	Solicitudes	Porcentaje de admisión	Solicitudes	Porcentaje de admisión
A	825	62 %	108	82 %
B	560	63 %	25	68 %
C	325	37 %	593	34 %
D	417	33 %	375	35 %
E	191	28 %	393	24 %
F	373	6 %	341	7 %
Total	2.691	44 %	1.835	30 %

Fuente: Freedman, Pisani y Purves (2007).

Como puede observarse en la tabla, los porcentajes de admisión son bastante similares en estos seis departamentos. La excepción más notable es el departamento A, que mostró una preferencia importante por las mujeres y aceptó un 82 % en comparación con el 62 % de los hombres. En el sentido contrario, el departamento E mostró una preferencia más clara por los hombres y aceptó un 28 % en comparación con el 24 % de las mujeres. En cambio, si nos fijamos en las solicitudes a los seis departamentos en conjunto, la relación entre el género de los candidatos y su aceptación en los programas de posgrado vuelve a ser evidente a favor de los hombres, con una tasa global del 44 % en comparación con la del 30 % en el caso de las mujeres.

Una diferencia de 14 puntos entre hombres y mujeres en la tasa global de aceptación de los seis departamentos más grandes volvería a ser una evidencia a favor de la existencia de una discriminación por razón de género en la Universidad de Berkeley. Pero si observamos con detenimiento los datos desagregados para cada departamento que recoge la tabla 2, seremos capaces de encontrar una explicación intuitiva a esta aparente contradicción.

Teniendo en cuenta las tasas de aceptación respectivas, los departamentos A y B serían los que más solicitudes aceptaron finalmente y, por lo tanto, aquellos a los que los candidatos –fueran hombres o mujeres– que se presentaron les resultó más fácil acceder. Con unos porcentajes que varían entre el 82 % y el 62 %, esto supone que al menos dos terceras partes acabaron accediendo a los programas que ofrecían estos dos primeros departamentos. En cambio, los departamentos C, D, E y F serían los que más dificultades pusieron a los candidatos –fueran hombres o mujeres– porque finalmente resolvieron favo-

rablemente un número sensiblemente más bajo de las solicitudes que recibieron. Con unos porcentajes que oscilan entre el 37 % y el 6 %, al menos dos terceras partes de los candidatos no acabaron accediendo en sus programas.

Los departamentos, por lo tanto, no mostraron un comportamiento similar en relación con la aceptación de los estudiantes. Pero, y esto es lo más importante para entender la aparente contradicción de este caso de discriminación por razón de género, los estudiantes tampoco mostraron un comportamiento similar en relación con la elección del departamento para presentar sus candidaturas.

Teniendo en cuenta el número de solicitudes que recibieron, los departamentos A y B valoraron un total de 1.385 hombres, es decir, algo más de la mitad (51,47 %) de los 2.691 que se presentaron como candidatos en el conjunto de los seis departamentos. En cambio, los departamentos C, D, E y F valoraron 1.702 mujeres, que representan casi la práctica totalidad (92,75 %) de las 1.835 que se presentaron. De este modo, los hombres solicitaron el acceso a los departamentos más fáciles o, al menos, a aquellos que más candidatos aceptaron, mientras que las mujeres lo hicieron, contrariamente, a los más difíciles o que menos candidatos aceptaron. Por esta razón, a pesar de que de manera agregada podría parecer lo contrario, cuando controlamos las diferencias entre los hombres y las mujeres en la elección del departamento, como hacemos en la tabla 2, la relación entre el género de los candidatos y su aceptación final en los programas de posgrado a favor de los hombres prácticamente desaparece.

2. Asociación, confusión y causalidad

El caso de la discriminación de género en la Universidad de Berkeley que acabamos de explicar se ha convertido en un ejemplo clásico de un fenómeno que a menudo se produce en el análisis estadístico cuando el estudio de las relaciones entre dos variables omite o no tiene en cuenta adecuadamente alguna información relevante para el estudio. Es lo que se ha denominado la *paradoja de Simpson*, expresión acuñada por Blyth (1972) a partir de la exposición de Simpson (1951) para hacer referencia a un fenómeno que, en realidad, fue descrito originalmente unos cuantos años antes por Yule (1903) como extensión a las tablas de contingencia de la discusión que hizo antes Pearson sobre la existencia de correlaciones espurias entre variables cuantitativas (Aldrich, 1995; David y Edwards, 2001).

Podemos definir la paradoja de Simpson como el hecho de que una asociación observada entre dos variables cualitativas cambia su sentido si, en lugar de hacerlo de manera agregada, se analiza su relación en cada uno de los subgrupos que se conforman a partir de una tercera variable cualitativa.

La paradoja de Simpson no es un fenómeno infrecuente en las disciplinas basadas en el análisis de datos cuantitativos, particularmente en los estudios observacionales, y resulta especialmente sorprendente a ojos del público no especializado que no espera encontrarse este tipo de contradicciones. Una universidad no puede discriminar a las mujeres en la resolución de las solicitudes de acceso en el conjunto de los estudios que ofrece y a la vez no hacerlo o, incluso, discriminar ligeramente los hombres en cada uno de sus departamentos. Pero en ningún caso es adecuado interpretar esta aparente contradicción como el resultado de un artefacto estadístico o como un indicio de que la investigación haya sido incorrectamente diseñada o desarrollada. Las relaciones observadas existen, son reales, tanto en el caso del conjunto de los candidatos valorados por la Universidad de Berkeley como en el detalle de sus departamentos.

Lo que pone de manifiesto la contradicción no es la existencia de estas relaciones en los dos niveles de análisis, sino el hecho de que las evidencias observadas de asociación entre las variables sean empleadas para llevar a cabo juicios causales. Teniendo en cuenta que en el análisis agregado se estaría omitiendo o no teniendo en cuenta adecuadamente una información relevante para el estudio, la relación observada entre las variables resultaría una estimación sesgada y, por lo tanto, una evidencia inadecuada para la inferencia causal que persigue. Solo cuando se toman en consideración los resultados del análisis

desagregado, no sesgado en el supuesto que nos ocupa, es posible entender adecuadamente el fenómeno objeto de estudio en los diferentes subgrupos y, de este modo, la aparente contradicción se diluye.

En este sentido, podemos considerar la paradoja de Simpson como un caso particular, de hecho el más extremo, de **confusión**. Un **factor** o una **variable de confusión** es una variable extraña, no prevista o contemplada en la investigación, que puede alterar la relación entre dos variables que son objeto de interés y que, por lo tanto, puede afectar a los juicios de causalidad que hacen los investigadores a partir de la observación de su asociación.

Si, en el contexto de una investigación que tenga como objetivo poner a prueba una relación de causalidad, observamos una asociación entre una **variable independiente** –también llamada *variable predictora* o *explicativa*– y una **variable dependiente** –también conocida como *variable resultado* o *explicada*–, una tercera variable sería un factor de confusión si su incorporación al análisis comportara el incremento, el decrecimiento, la desaparición o, incluso, como hemos podido ver, la inversión de su relación.

Para hacerlo, el potencial factor de confusión tendría que cumplir necesariamente la condición de estar asociado tanto con la variable dependiente como con la independiente, de manera que su efecto o contribución específica en relación con la variable dependiente resultaría indistinguible del que tendría la variable independiente. Es precisamente por esta razón que, como todos los investigadores deberían tener siempre presente en su práctica, a pesar de que la determinación de una relación de causalidad implica la observación de una asociación entre dos variables, la mera evidencia de esta asociación desde el punto de vista estadístico no implica, necesariamente, la existencia de una relación causal. Más allá de estas nociones básicas, los lectores interesados pueden encontrar una introducción general en el estudio de las relaciones de causalidad en la investigación social en Russo (2009) y una discusión más amplia sobre el establecimiento de este tipo de inferencias en el trabajo pionero de Pearl (2000).

El estudio sobre la discriminación por razón de género en el acceso de los estudiantes a los programas de posgrado de la Universidad de Berkeley es, por lo tanto, un buen ejemplo de investigación en que la omisión de una variable de confusión en el análisis agregado para el conjunto de los departamentos conduce a una conclusión sesgada. Tal y como hemos podido ver, una sencilla inspección visual de la tabla 2, que recoge la distribución de los seis departamentos más grandes en función del número de solicitudes presentadas por los candidatos y de sus tasas de aceptación final, nos ha permitido esbozar una explicación intuitiva sobre su papel como potencial factor de confusión. Teniendo en cuenta que ni los departamentos ni los estudiantes se comportaron

de manera similar, el cambio de sentido en la relación entre el género de los candidatos y su aceptación era consecuencia de la preferencia de los hombres y las mujeres por los más fáciles y más difíciles de acceder, respectivamente.

En cualquier caso, al no disponer de los datos originales desagregados para la totalidad de los departamentos, no es posible ir más allá de esta explicación intuitiva y mostrar, mediante las pruebas estadísticas oportunas, de qué manera el departamento actúa en este caso como un factor de confusión y, por lo tanto, cumple la condición necesaria de estar asociado tanto al género de los candidatos (variable independiente) como a su aceptación final (variable dependiente). En cambio, podemos ilustrar este requerimiento con un ejemplo ficticio que, además, nos permitirá poner de manifiesto cómo la incorporación de un factor de confusión al análisis no solo puede alterar la relación observada entre dos variables, sino que, incluso, puede hacer evidente una relación que ni siquiera había sido observada inicialmente.

Imaginemos una universidad ficticia formada, para simplificar el análisis, únicamente por dos departamentos. Teniendo en cuenta el conjunto global de solicitudes, supongamos que se presentaron un total de 1.000 candidatos, de los cuales 450 habrían sido hombres y 550 mujeres. Supongamos también que de estos candidatos finalmente un 60 %, tanto de hombres como de mujeres, habrían sido aceptados para iniciar sus estudios. La tabla 3 recoge estos datos, desagregando las candidaturas admitidas y rechazadas en función del departamento escogido y del género de los solicitantes.

Tabla 3. Resolución sobre las solicitudes de acceso a una universidad ficticia según el departamento escogido y el género de los candidatos

		Solicitudes	Admisiones	Rechazos	Porcentaje de admisión
Departamento A	**Hombres**	200	80	120	40,00 %
	Mujeres	100	20	80	20,00 %
Departamento B	**Hombres**	250	190	60	76,00 %
	Mujeres	450	310	140	68,89 %
Total	**Hombres**	450	270	180	60,00 %
	Mujeres	550	330	220	60,00 %

Fuente: elaboración propia.

En este caso, teniendo en cuenta que la tasa global de aceptación en el conjunto de los dos departamentos habría sido del 60 %, tanto para los hombres como para las mujeres, el hecho de que no se observe ninguna diferencia sería una evidencia en contra de la existencia de una discriminación por razón de género. Si utilizamos los datos totales que se presentan en la última fila para construir una tabla de contingencia, el análisis de su asociación nos permite afirmar que, al menos de manera agregada, no existe ninguna relación entre

el género de los candidatos y su aceptación en esta universidad ficticia (X^2 = 0, df = 1, p = 1). Como es natural, tratándose de dos variables totalmente independientes entre sí, la intensidad o magnitud de su relación es nula (V de Cramér = 0).

Nuestra universidad ficticia no mostraría ninguna preferencia, ni por los hombres ni por las mujeres, en la resolución de las solicitudes de acceso de los estudiantes a sus programas. Pero, si en lugar de hacer un análisis agregado nos fijamos en los datos que corresponden a cada uno de los dos departamentos, la situación que nos encontramos resulta muy diferente. Teniendo en cuenta sus respectivas solicitudes, al departamento A se habrían presentado 200 hombres y 100 mujeres, de los cuales habrían sido finalmente aceptados un 40 % y un 20 %, respectivamente. En un sentido similar, al departamento B se habrían presentado 250 hombres y 450 mujeres, de los cuales habrían sido aceptados, respectivamente, un 76 % y aproximadamente un 69 %.

Una diferencia entre hombres y mujeres de 20 puntos en el departamento A y de 17 puntos en el departamento B sería una evidencia clara a favor de la existencia de una discriminación por razón de género. Los dos departamentos de esta universidad preferirían, en realidad, a los hombres antes que a las mujeres como estudiantes de sus programas.

De hecho, si utilizamos los datos que se presentan en la primera y en la segunda fila para construir dos tablas de contingencia separadas, el análisis de la asociación nos permitiría afirmar que existe una relación estadísticamente significativa entre el género de los candidatos y su aceptación a favor de los hombres, tanto en el departamento A (X^2 = 12, df = 1, p < 0,001) como en el departamento B (X^2 = 3,98, df = 1, p < 0,05). Aun así, la intensidad o magnitud de esta relación es más importante en el caso del primer departamento (V de Cramér = 0,2) que en el segundo (V de Cramér = 0,08).

En este sentido, el análisis de los datos desagregados para cada uno de los dos departamentos de nuestra universidad ficticia sugiere la existencia de un factor de confusión que debería ser tenido en cuenta. Más allá de la inspección visual de las tasas de aceptación de la tabla 3, a continuación presentamos dos tablas de contingencia construidas a partir de los mismos datos, que nos permitirán determinar hasta qué punto el departamento cumple la condición necesaria exigida a cualquier factor o variable de confusión y que, por lo tanto, está efectivamente relacionado tanto con la aceptación de los candidatos –es decir, la variable dependiente, resultado o explicada– como con su género –la variable independiente, predictiva o explicativa.

Por un lado, agrupando todos los candidatos independientemente de su género, la tabla 4 presenta los datos de admisión según el departamento escogido y muestra una importante diferencia en su comportamiento en relación con la aceptación de los estudiantes que se habrían presentado. Así, el departamento

A sería el que más dificultades habría puesto a los estudiantes, de forma que habría resuelto favorablemente solo un tercio (33,33 %) de sus 300 solicitudes. En comparación, habría sido más fácil acceder al departamento B, que habría aceptado algo más de dos tercios (71,43 %) de las 700 solicitudes que habría valorado.

Tabla 4. Datos de admisión a una universidad ficticia según el departamento escogido por los candidatos

Departamento	Solicitudes	Admisiones	Rechazos	Porcentaje de admisión
A	300	100	200	33,33 %
B	700	500	200	71,43 %
Total	1.000	600	400	60,00 %

Fuente: elaboración propia.

Por otro lado, agrupando ahora todos los candidatos independientemente de su aceptación final en los departamentos, la tabla 5 presenta las solicitudes de acceso según el género de los candidatos y muestra también una importante diferencia en su comportamiento en relación con la elección del departamento para presentar sus candidaturas. Así, el departamento A sería el que menos mujeres habrían escogido, de forma que sus 100 candidatas solo suponen un tercio (33,33 %) de las solicitudes que habría valorado. En cambio, en el departamento B se habrían presentado más mujeres, y habría valorado 450 candidatas que representan casi dos tercios (64,29 %) de sus solicitudes.

Tabla 5. Solicitudes de acceso a una universidad ficticia según el género de los candidatos

Departamento	Solicitudes	Hombres	Mujeres	Porcentaje de mujeres
A	300	200	100	33,33 %
B	700	250	450	64,29 %
Total	1.000	450	550	55,00 %

Fuente: elaboración propia.

En este sentido, utilizando estas dos tablas de contingencia para analizar la asociación del departamento con las dos variables, podemos afirmar que existe una relación estadísticamente significativa tanto con la aceptación final de los candidatos ($X^2 = 126{,}98$, df = 1, $p < 0{,}001$) como con su género ($X^2 = 81{,}29$, df = 1, $p < 0{,}001$) que, además, resulta comparativamente de una intensidad o magnitud más importante en el primer caso (V de Cramér = 0,36 y 0,29, respectivamente). En efecto, tal como sugería la inspección preliminar de los datos desagregados, el departamento estaría actuando como factor o variable de confusión y, por lo tanto, el análisis agregado en el caso de nuestra universidad ficticia nos habría llevado a una conclusión sesgada.

3. Diseño de la investigación e inferencia estadística

La lección que podemos extraer del caso de la discriminación de género de la Universidad de Berkeley, como ejemplo clásico de la paradoja de Simpson, es que la existencia de potenciales factores de confusión no considerados en el análisis es una de las amenazas más importantes para los investigadores que se plantean hacer juicios de causalidad a partir de la observación de asociaciones entre sus variables. Como hemos podido ver, la incorporación de estos factores al análisis puede comportar el incremento, el decrecimiento, la desaparición o, incluso, la inversión de las relaciones observadas, de forma que la mera evidencia de la existencia de una asociación entre dos variables no implica, necesariamente, que esta relación sea de naturaleza causal.

De hecho, la incorporación de un factor de confusión al análisis no solo puede alterar la relación observada entre dos variables, sino que también puede hacer evidente una relación que, como en el caso de nuestra universidad ficticia, ni siquiera había sido inicialmente observada. Por esta razón, sea cual sea el tipo de investigación, es obligación de los investigadores considerar la eventual influencia de cualquier tipo de variable extraña que pudiera interferir y, por lo tanto, examinar exhaustivamente las relaciones entre sus variables y los potenciales factores de confusión relevantes en el contexto particular de sus estudios.

En este sentido, es importante tener presente que la capacidad de los investigadores para establecer inferencias causales a partir del análisis de sus datos está muy relacionada con la naturaleza del diseño de la investigación empleado para obtenerlas. Si entendemos el análisis estadístico como la culminación de un complejo proceso de planificación a través del cual se lleva a cabo cualquier investigación cuantitativa, resulta conveniente distinguir dos grandes tipos de diseños: la **investigación experimental** y la **investigación observacional**.

En los dos casos, la investigación parte del desarrollo o la adopción de una teoría como el marco general de referencia a partir del cual sea razonable establecer una relación causal entre las variables, el planteamiento de algunas hipótesis sobre las relaciones entre las variables dependientes e independientes para poder poner a prueba su asociación mediante las pruebas estadísticas oportunas y, como decíamos, la consideración de cualquier variable extraña que pudiera actuar como factor de confusión, es decir, que interfiriera en las relaciones objeto del análisis y, por lo tanto, pudiera convertirse en una explicación alternativa.

La diferencia sustancial, como veremos a continuación, se encuentra en la capacidad de los investigadores para manipular las variables independientes de forma que sea posible atribuir adecuadamente las diferencias observadas en

las variables dependientes a las variaciones de las variables independientes. Más allá de la breve exposición que haremos a continuación, los lectores interesados pueden encontrar una discusión más profunda sobre el diseño de la investigación en los trabajos de Shadish, Cook y Campbell (2002), Coolican, (2014) o Cozby y Bates (2015).

De una manera sencilla, podemos caracterizar la **investigación experimental** describiendo la forma más simple que puede adoptar un **experimento**. En este contexto, los investigadores tienen el control sobre los diferentes niveles o las condiciones de al menos una variable independiente –generalmente denominada *tratamiento*–, de forma que pueden decidir de acuerdo con su voluntad cómo serán expuestos los participantes. Mediante una asignación aleatoria, los investigadores seleccionan los individuos que forman parte de cada uno de los grupos experimentales y, una vez administrado el tratamiento, miden sus efectos en una o más variables dependientes.

Así, cuando disponen de una muestra suficientemente amplia, los investigadores igualan los diferentes grupos experimentales en relación con cualquier factor o variable de confusión, de forma que su influencia en la variable dependiente quede neutralizada gracias a la aleatorización de los participantes. A pesar de que, de acuerdo con esta lógica general, un experimento puede adoptar formas mucho más complejas, su rasgo característico se encuentra en la capacidad que da a los investigadores para atribuir, más allá de las pequeñas diferencias entre los grupos debido al azar, las variaciones observadas en la variable dependiente como una consecuencia necesaria de la manipulación de la variable independiente o tratamiento.

Por otro lado, es posible caracterizar la **investigación observacional** como la que se produce cuando los investigadores no tienen control sobre los diferentes niveles o las condiciones de una o más variables independientes. Este tipo de investigación puede adoptar muchas formas, pero una de las más frecuentes se basa en la utilización de un **cuestionario** o una **encuesta**. En este contexto, los investigadores definen sus variables independientes y, como consecuencia de la imposibilidad de manipularlas de acuerdo con su voluntad, se limitan a observarlas a partir de las respuestas proporcionadas por una muestra generalmente amplia de participantes.

Una vez administrados sus cuestionarios, los investigadores identifican a los individuos que forman parte de los diferentes grupos previamente existentes y miden sus diferencias en una o más variables dependientes. De este modo, con una cierta confianza, atribuyen estas diferencias a las variaciones existentes en la variable independiente. Pero a diferencia de la investigación experi-

mental, en este escenario no será posible evitar la intervención de potenciales factores o variables de confusión en las relaciones observadas, de forma que les resultará difícil excluir la posibilidad de que su influencia se convierta en una explicación alternativa a la que proponen.

Estos dos tipos de investigación difieren en su **validez interna**, es decir, en la capacidad para proporcionar las evidencias necesarias que permitan determinar la existencia de una relación de causalidad a partir de la observación de una asociación entre las variables dependientes e independientes. Obviamente, los resultados de un único estudio no son nunca suficientes para dar por probada una relación de este tipo. Pero el hecho de que los investigadores utilicen, siempre que les resulte posible, la asignación aleatoria de los individuos a los diferentes grupos que caracteriza la metodología experimental, puede permitirles obtener evidencias más sólidas para llevar a cabo juicios causales a partir de sus resultados.

Este no es, sin embargo, el único momento en que el azar juega un papel importante en el diseño de la investigación. De hecho, resulta también determinante cuando los investigadores se proponen, como suele ser habitual, generalizar sus conclusiones más allá de los límites de sus estudios particulares. Con independencia del tipo de investigación, sea experimental u observacional, es en el momento del diseño y la construcción de la muestra que los investigadores tienen que seleccionar los participantes que, finalmente, acabarán formando parte de sus estudios.

Dado que, por razones prácticas, no siempre es posible obtener información sobre el conjunto de la población objeto de análisis en una investigación, a menudo los investigadores llevan a cabo un proceso de selección con el objetivo de escoger solo una fracción, un subconjunto, del total de individuos que la conforman. En este sentido, es posible identificar dos grandes tipos de estrategias para la elección de los participantes de cualquier investigación: la **selección aleatoria** o **probabilística** y la **selección no aleatoria** o **intencional**.

De manera sintética, consideramos que una **muestra es aleatoria** cuando todos y cada uno de los individuos que forman parte de la población tienen la misma probabilidad de ser seleccionados para formar parte de la investigación. Partiendo de una definición clara y precisa de la población que es objeto de estudio, en condiciones ideales, los investigadores deberían ser capaces de identificar a todos los miembros –por ejemplo, a partir de una lista con los nombres– y, a continuación, procederían a escoger al azar a aquellos que finalmente serán sus participantes. En cambio, una **muestra es no aleatoria** cuando los individuos no han sido escogidos usando esta estrategia, sino que, más bien, son sencillamente el producto accidental de una elección intencio-

nal según su conveniencia o disponibilidad. Es por esta razón que, de acuerdo con esta segunda estrategia, no todos los individuos que conforman la población de interés tienen, de hecho, la misma probabilidad de ser seleccionados.

Aunque una muestra aleatoria pueda adoptar formas mucho más complejas, es conveniente señalar que solo cuando el criterio de selección de los participantes es aleatorio tendremos las garantías suficientes para considerar que las muestras son representativas. De este modo, los investigadores tendrán la confianza de que las relaciones observadas a partir de la asociación entre sus variables serán extrapolables al conjunto de la población a partir de la que han sido extraídas las muestras. Es por esta razón que, tanto la investigación experimental como la observacional, no solo difieren en su validez interna, sino que también pueden hacerlo en su **validez externa**. Es decir, en la capacidad para proporcionar las evidencias necesarias que permitan concluir, con las garantías suficientes, que la existencia de una relación es generalizable a otras situaciones o a otros individuos que no han formado parte del estudio.

La tabla 6 presenta esquemáticamente la relación entre la selección y la asignación de los participantes en el diseño de la investigación que, a continuación, nos permitirá poner de relieve la importante contribución que tiene el azar en el proceso de inferencia estadística.

Tabla 6. La relación entre el diseño de la investigación y la inferencia estadística

	Asignación aleatoria	**Asignación no aleatoria**	
Selección aleatoria	Relación causal generalizable	Relación no causal generalizable	**Alta validez externa**
Selección no aleatoria	Relación causal no generalizable	Relación no causal no generalizable	**Baja validez externa**
	Alta validez interna	**Baja validez interna**	

Fuente: elaboración propia.

De acuerdo con esta tabla, el cruce de las diferentes formas con que pueden ser seleccionados y asignados los individuos a los diferentes grupos proporciona cuatro tipos básicos de investigaciones que difieren, fundamentalmente, en su validez. En primer lugar, el cuadrante superior izquierdo representa la investigación que, mediante su diseño, lleva a cabo una selección y una asignación aleatorias de los participantes. Sería el caso de un experimento desarrollado a partir de una muestra representativa, en la que la validez interna y externa de la investigación serían óptimas y, por lo tanto, los investigadores se encontrarían en las mejores condiciones para establecer una relación causal a partir de la observación de las relaciones entre sus variables que también fuera generalizable a la población.

A su vez, en los cuadrantes superior derecho e inferior izquierdo encontramos las investigaciones que únicamente llevan a cabo una selección o una asignación aleatorias y que, por lo tanto, tendrían una validez interna o externa,

respectivamente, más baja. En el primer caso, se trataría de una encuesta administrada a una muestra representativa, que permitiría establecer relaciones generalizables al conjunto de la población pero que, en ningún caso, proporcionaría evidencias suficientes para determinar la naturaleza causal. En el segundo, se trataría del caso de un experimento llevado a cabo a partir de una muestra no representativa, que proporcionaría evidencias sobre la naturaleza causal de la relación pero que, en cambio, no permitiría su generalización al conjunto de la población.

Finalmente, en el peor de los escenarios posibles desde el punto de vista tanto de la validez interna como de la externa, el cuadrante inferior derecho representa la investigación que no lleva a cabo ni una selección ni una asignación aleatorias de los participantes. Este sería el caso de una encuesta dirigida a una muestra no representativa en la que, por lo tanto, no sería posible establecer ni la naturaleza causal de las relaciones observadas ni generalizar las conclusiones obtenidas al conjunto de la población.

Estos cuatro tipos de investigación difieren fundamentalmente en su validez y, como hemos podido ver, la razón por la cual esto es así no es otra que el papel que juega el **azar** en el diseño. En este sentido, la distinta capacidad que tienen los investigadores para determinar la existencia de una relación causal generalizable al conjunto de la población a partir de la observación de relaciones entre sus variables sirve como una buena ilustración de la importante contribución del azar en la inferencia estadística.

Si entendemos la **inferencia estadística** como el proceso a través del cual podemos extraer conclusiones generales a partir del análisis de los datos obtenidos de una muestra, es necesario tener presente que este proceso únicamente es posible si la selección de los participantes o la asignación a los diferentes grupos han sido aleatorias. Es decir, solo cuando el azar interviene en al menos uno de estos dos momentos importantes para el diseño de la investigación es posible llegar a concluir si las diferencias observadas en la variable dependiente son consecuencia de la manipulación de la variable independiente o tratamiento –**inferencia causal**–, o si estas diferencias son generalizables más allá de la muestra –**inferencia a la población**.

De este modo, siempre que se cumpla esta condición, la estadística inferencial proporciona un conjunto de procedimientos que permite a los investigadores evaluar las asociaciones observadas y decidir, con un determinado nivel de confianza, hasta qué punto son realmente el producto de una relación causal existente en el conjunto de la población. O lo que es lo mismo, disponer de las

evidencias suficientes para ser capaces de excluir la posibilidad alternativa de que los resultados obtenidos puedan ser, en realidad, explicados como consecuencia de una selección y/o una asignación no aleatorias de los participantes.

4. ¿Qué es el análisis multivariante y para que sirve?

A pesar de la importancia del diseño de la investigación para extraer conclusiones no sesgadas que, además, sean generalizables más allá de los límites de los estudios particulares, lo cierto es que los investigadores no siempre pueden utilizar experimentos para desarrollar sus trabajos de campo. En este sentido, cuestiones de orden práctico o ético pueden desaconsejar –o incluso impedir– que se lleve a cabo una asignación aleatoria de los participantes en las diferentes condiciones experimentales. Esta situación es bastante frecuente en las disciplinas basadas en el análisis de datos cuantitativos y resulta especialmente evidente cuando los estudios se desarrollan, lejos de las condiciones controladas de los laboratorios, en los contextos naturales en que se produce la actividad cotidiana de las personas.

Si, como planteábamos al inicio de este texto, el objetivo es analizar fenómenos complejos como la eventual discriminación por razón de género en el acceso de los estudiantes a una universidad, resulta obvio que no será posible decidir el género de los candidatos ni, del mismo modo, tampoco se podrá escoger el departamento al que los candidatos tendrían que presentar las solicitudes. De hecho, incluso cuando se reúnen las condiciones idóneas para usar experimentos, los investigadores no siempre pueden prever o controlar adecuadamente, mediante el diseño de la investigación, todos y cada uno de los potenciales factores de confusión que podrían amenazar sus conclusiones.

Es en este contexto en que la manipulación de las variables no es una estrategia factible o suficiente para obtener evidencias sólidas que permitan sustentar juicios de causalidad a partir de la observación de asociaciones entre variables que el análisis multivariante se presenta como el marco analítico general que permite modelar las múltiples relaciones existentes entre las diferentes variables involucradas en una determinada investigación.

> En este sentido, podemos definir el **análisis multivariante** como el conjunto de técnicas estadísticas que tienen como objetivo analizar e interpretar las relaciones entre distintas variables de manera simultánea, mediante la construcción de modelos estadísticos complejos que permiten distinguir la contribución independiente de cada una de ellas en el sistema de relaciones y, de este modo, describir, explicar o predecir los fenómenos que son objeto de interés para la investigación.

Por lo tanto, este marco analítico general ofrece a los investigadores la oportunidad de llevar a cabo el **control estadístico** de cualquier variable extraña que, como eventual factor de confusión, pudiera interferir en la relación entre

las variables dependientes e independientes que son objeto de interés. Pero es importante tener presente que la elección de las técnicas estadísticas –y el análisis multivariante no es una excepción– no tiene ninguna relación con el diseño empleado en la investigación, de forma que estas técnicas pueden ser utilizadas para analizar los datos obtenidos tanto en los contextos experimentales como en los observacionales. Como ya hemos explicado, la única limitación se encuentra en el momento de la interpretación de los resultados y, especialmente, en el riesgo que los investigadores estén dispuestos a asumir en el momento de determinar la existencia de sus relaciones a partir de las evidencias de que disponen.

De una manera sencilla, podemos entender el análisis multivariante como una extensión del análisis bivariante y este, a su vez, como una extensión del análisis univariante.

En este sentido, el **análisis univariante** es la forma más simple de análisis estadístico y se propone describir la distribución de una única característica de los individuos que forman parte de la investigación. Mediante la construcción de una tabla de frecuencias en el caso de una variable cualitativa, o bien del cálculo de una medida de tendencia central –como la media, la mediana o la moda– y de su dispersión –como el rango, la desviación estándar o la varianza– cuando se trata de una variable cuantitativa, la clave de este tipo de análisis se encuentra en el hecho de que solo toma en consideración una única variable con el objetivo de realizar una descripción de la muestra y, cuando es posible, establecer una inferencia sobre la población a la que representa.

Obviamente, cuando los investigadores llevan a cabo sus estudios nunca concentran todos los esfuerzos en observar únicamente una variable, pero, sea cual sea el número de medidas registradas en la investigación, este primer tipo de análisis se limita a explorar cada una de las variables de manera independiente. Así, retomando el caso del estudio sobre la discriminación de género en el acceso a la universidad, la estadística univariante nos permite conocer la proporción de estudiantes de la muestra que serían hombres o mujeres, los departamentos que habrían escogido para presentar sus solicitudes, o la cantidad de candidatos que finalmente habrían sido aceptados o rechazados por la universidad.

Por otro lado, el **análisis bivariante** es una extensión del análisis univariante que, a pesar de mantener su naturaleza exploratoria, se propone, en cambio, determinar la relación existente entre dos características de los participantes de la investigación. Mediante la construcción de una tabla de contingencia cuando se trata de variables cualitativas, o del cálculo de una correlación en el caso de variables cuantitativas, este tipo de análisis tiene por objeto examinar la distribución de una variable dependiente, resultado o explicada en función de los niveles de otra variable independiente, predictora o explicativa. De este

modo, la observación de su asociación permite determinar la existencia de una relación en la muestra y, siempre que sea posible, establecer una inferencia sobre la población que representa.

Como ya hemos dicho, la mera evidencia de una asociación entre dos variables desde el punto de vista estadístico no implica, necesariamente, la existencia de una relación causal. Y esto es a causa, en última instancia, del hecho de que este segundo tipo de análisis permite a los investigadores tener en cuenta las relaciones entre todas y cada una de las posibles parejas de sus variables, pero lo hace, en cada ocasión, de manera independiente. Así, no es posible descartar que cualquier otra variable pueda interferir en estas relaciones actuando como un potencial factor de confusión y, por lo tanto, alterando o incluso haciendo evidentes las relaciones entre dos variables que podrían no haber sido observadas inicialmente. Siguiendo con nuestro caso, la estadística bivariante nos permitiría conocer la relación entre el género de los candidatos y su aceptación final a los programas de la universidad o, lo que ha sido más importante, la relación del departamento tanto con la aceptación como con el género de los candidatos.

En este sentido, como extensión del análisis bivariante, el análisis multivariante se presenta como el marco analítico general que se propone analizar e interpretar las relaciones entre diversas variables, pero lo hace, en este caso, mediante la construcción de modelos complejos que permiten determinar su existencia de manera simultánea. Así, más allá de la consideración de las variables dependientes e independientes, este tipo de análisis permite a los investigadores incorporar a sus estudios las **variables de control** que sean necesarias. Es decir, les permite tener en cuenta todas las variables extrañas que eventualmente podrían actuar como factores de confusión y que, por lo tanto, podrían interferir en las relaciones que son realmente objeto de interés.

Controlando estadísticamente la contribución de todas estas variables al sistema de relaciones, este tercer tipo de análisis permite mantener constantes sus efectos y obtener así una estimación más precisa de las relaciones realmente existentes entre las variables dependientes y las independientes. Por lo tanto, la observación de las asociaciones entre las diferentes variables consideradas en la construcción de estos modelos permite determinar la existencia de múltiples relaciones en la muestra de participantes y, cuando se reúnen las condiciones necesarias, establecer inferencias sobre el conjunto de la población. De hecho, como veremos más adelante, este marco analítico no solo permite analizar las relaciones de dependencia entre las diferentes variables involucradas en una investigación, sino que también sirve para analizar, teniendo en

cuenta su interdependencia, las relaciones entre las variables que no pueden ser consideradas ni dependientes ni independientes desde un punto de vista teórico.

Con objeto de acabar con el caso que nos ha servido de hilo conductor hasta ahora, la estadística multivariante permitiría conocer la contribución simultánea de las características de los estudiantes y de los departamentos a los que habrían presentado sus solicitudes que estarían implicadas en la aceptación final de los candidatos. Más allá del papel del departamento como potencial factor de confusión, esta investigación podría tener en cuenta también las diferencias entre hombres y mujeres en cuanto a sus capacidades, aptitudes o habilidades, controlando, por ejemplo, el expediente académico previo o los resultados en las pruebas de acceso, de forma que sería posible extraer una conclusión todavía más precisa sobre la existencia de una discriminación por razón de género en el acceso de los estudiantes a la universidad.

Sin embargo, resulta conveniente tener presente que no todos los autores comparten esta manera de entender el análisis multivariante. De hecho, una corriente alternativa considera que esta aproximación es poco restrictiva y, en cambio, define este tipo de análisis como el que se utiliza únicamente en investigaciones que consideran múltiples variables dependientes. En este sentido, entienden también el análisis multivariante como una generalización del análisis univariante y bivariante, pero lo hacen tomando como punto de partida definiciones diferentes de estos dos tipos de análisis.

Por un lado, definen la estadística univariante como aquella que, en contextos experimentales, se ocupa de una única variable dependiente y, por lo tanto, no excluye la posibilidad de que los investigadores consideren más de una variable independiente en el análisis. Por otro lado, entienden la estadística bivariante como el estudio de las relaciones entre parejas de variables que habrían sido obtenidas en investigaciones observacionales, de modo que, de acuerdo con esta argumentación, no sería posible distinguir entre variables dependientes e independientes. En este sentido, la estadística multivariante no sería más que una generalización del análisis univariante en que, sea cual sea el número de variables independientes consideradas, los investigadores amplían el número de variables dependientes en la construcción de sus modelos.

Pero esta aproximación alternativa plantea algunos inconvenientes que hacen que su adopción sea poco interesante. En primer lugar, establece una relación directa entre el diseño de la investigación y el tipo de análisis que es posible desarrollar. Estrictamente hablando, en cambio, el análisis estadístico no impone ningún requerimiento en relación con la naturaleza experimental u observacional de los datos obtenidos, de modo que, como ya hemos señalado, es responsabilidad de los investigadores valorar hasta qué punto las evidencias observadas de asociación entre sus variables son suficientes para determinar la existencia de relaciones de causalidad en sus estudios.

En segundo lugar, este planteamiento más restrictivo sobre el análisis multivariante focaliza la atención únicamente en las relaciones de dependencia entre las variables y, por lo tanto, excluye la posibilidad de que este marco analítico general sirva también para analizar relaciones de interdependencia. Finalmente, limita su alcance a las investigaciones que consideran como mínimo dos variables dependientes y, de este modo, omite otros escenarios igualmente interesantes en que los investigadores se proponen el objetivo de determinar la contribución simultánea de diversas variables independientes en una única variable dependiente.

En cualquier caso, es importante tener presente que la clave del análisis multivariante como el marco analítico general no es que los investigadores dispongan de múltiples variables, porque, como ya hemos dicho, los estudios no están diseñados con el objetivo de observar una única variable. El rasgo distintivo de este tipo de análisis, y la razón por la que resultan especialmente útiles para abordar problemas complejos, es la capacidad que tienen de modelar las múltiples relaciones existentes entre las diferentes variables involucradas en una investigación de manera simultánea. En este sentido, la construcción de modelos complejos, tanto de dependencia como de interdependencia, comparte una lógica común que se basa en la **combinación lineal de variables**.

Para hacer esto, en función de los objetivos de la investigación y, especialmente, del tipo de relaciones que se plantean estudiar desde un punto de vista teórico, los investigadores disponen de diferentes procedimientos para estimar, a partir de los datos obtenidos de sus participantes, el peso específico o la importancia relativa de cada una de las variables consideradas en los modelos y, de este modo, ser capaces de llevar a cabo una evaluación de su contribución específica o independiente al sistema de relaciones.

Por un lado, en el contexto de las **relaciones de dependencia**, la combinación lineal de variables en que se basa el análisis multivariante sirve para explicar o predecir las dependientes a partir de las independientes y, por lo tanto, ofrece la posibilidad de controlar el efecto de cualquier factor o variable de confusión que pudiera interferir en las relaciones que son realmente de interés para la investigación. Por otro lado, en el contexto del análisis de las **relaciones de interdependencia**, sirve para describir la estructura compartida por un conjunto de variables que no pueden ser identificadas como dependientes ni como independientes y, por lo tanto, ofrece la posibilidad de determinar la existencia de un tipo de supervariable o dimensión hipotética subyacente que, a pesar de no ser directamente observable, podría resultar interesante interpretar.

5. Una clasificación de las técnicas de análisis multivariante

Una vez definido el análisis multivariante como el marco analítico general que permite modelar las múltiples relaciones existentes entre las diferentes variables involucradas en una investigación, es el momento de presentar una clasificación de las diferentes técnicas disponibles. Esta clasificación general tiene como objetivo ofrecer una panorámica sobre las características y las condiciones en que pueden ser utilizadas y, de manera particular, servir de guía para que los lectores interesados puedan escoger la técnica que mejor se ajuste a su investigación.

A pesar de que, como hemos dicho, las técnicas de análisis multivariante pueden ser utilizadas para analizar los datos obtenidos tanto en contextos experimentales como observacionales, es importante tener presente que la elección de la técnica depende de dos aspectos estrechamente vinculados con el diseño de la investigación: la pregunta o el objetivo general que motiva su desarrollo y las características de los datos que proporciona para ofrecer una respuesta.

En este sentido, como hemos podido ver, el uso de las técnicas de análisis multivariante resulta conveniente cuando los investigadores se proponen responder preguntas que tienen que ver con el estudio de las múltiples relaciones existentes, ya sean de dependencia o de interdependencia, entre las diferentes variables involucradas en una investigación de manera simultánea. Pero antes de profundizar en los escenarios particulares en que se puede concretar el estudio de las relaciones en estos dos contextos, abordaremos brevemente la cuestión relativa a las características de los datos que proporciona la investigación.

Con independencia del objetivo general que se plantee, toda investigación cuantitativa se basa en la obtención de las evidencias necesarias que permitan a los investigadores establecer inferencias a partir de la observación de asociaciones entre sus variables. Para hacerlo, los investigadores no solo tendrán que planificar cómo se conducirá la investigación, sino que, además, tendrán que decidir cómo se codificará y registrará la información relativa a sus participantes, de forma que pueda ser tratada mediante las pruebas estadísticas oportunas. Es el momento de la **medida**, el proceso a través del cual los investigadores definen las variables de interés y establecen los diferentes niveles que pueden adoptar para reflejar adecuadamente la variabilidad observada en los fenómenos que se proponen estudiar.

A pesar de que puede ser un proceso complejo, especialmente en las investigaciones que se basan en la evaluación de atributos psicológicos no directamente observables (ved Meneses *et al.*, 2014, para una discusión más amplia), la medida no sería otra cosa que el establecimiento de una correspondencia entre las propiedades de los fenómenos que son objeto de interés y los números que las representan en una escala determinada. En este sentido, es posible distinguir dos grandes tipos de variables en función de la escala de medida que haya sido utilizada para definirlas: las variables cualitativas y las variables cuantitativas.

Por un lado, las **variables cualitativas** o **no métricas** son aquellas en las que la asignación de los números que representan sus diferentes niveles se corresponde con la presencia o ausencia de una determinada característica.

Este tipo de variables no refleja el grado o la cantidad con que la característica es presente, sino que, en cambio, únicamente permiten distinguir discretamente los individuos que cumplen las condiciones para pertenecer a un determinado nivel de entre todos los posibles. Para hacer esto, las variables cualitativas pueden ser definidas a partir del uso de escalas nominales y ordinales, cuando sus niveles sirven para identificar, respectivamente, individuos que pertenecen a grupos que son simplemente diferentes o que ocupan una posición relativa diferente en una serie ordenada.

En el primer caso, utilizan una **escala nominal** las variables que permiten codificar algunos atributos sociodemográficos clásicos como son, por ejemplo, el género, la ocupación o la religión y, en el contexto de la investigación experimental, el hecho de que los individuos hayan sido asignados o no a una de las condiciones experimentales. En el segundo caso, utilizan una **escala ordinal** las variables que también permiten tener en cuenta la existencia de un determinado orden entre sus niveles como, por ejemplo, el estatus socioeconómico o el nivel educativo alcanzado, pero que, en ningún caso, reflejan con precisión la cantidad o el grado con que la característica está presente.

Un caso particular de las variables cualitativas son las **dicotómicas**, que únicamente pueden tener dos niveles y que, en el contexto del desarrollo de modelos multivariantes, sirven para recodificar la información recogida en las variables cualitativas de tres o más niveles, de modo que es posible crear una serie de nuevas variables –llamadas *ficticias* o *dummies*– que identifican a todos los individuos que pertenecen a un determinado grupo por oposición al resto.

Por otro lado, las **variables cuantitativas** o **métricas** son aquellas en que la asignación de los números que representan sus diferentes niveles se corresponde exactamente con el grado o la cantidad con que una determinada característica está presente.

Este tipo de variables permite distinguir los individuos en función de la magnitud relativa con que se expresa la característica y, lo que es más importante, los valores que pueden adoptar se corresponden con unidades de medida constantes, de modo que cualquier diferencia entre ellos refleja una diferencia equivalente en relación con la característica representada. En este sentido, las variables cuantitativas pueden ser definidas a partir del uso de las escalas de intervalo y de razón, cuando entre sus niveles existe un punto cero arbitrario o, en cambio, cuando este punto cero es real y, por lo tanto, representa una ausencia absoluta de la característica.

En el primer caso, utilizan **escalas de intervalo** las variables que recogen información, por ejemplo, sobre el rendimiento en un examen, los resultados de una prueba de inteligencia o las puntuaciones obtenidas mediante tests diseñados para evaluar atributos psicológicos no directamente observables. A pesar de que no siempre es posible demostrar la existencia de una unidad de medida constante en todos estos casos, y que, por lo tanto, muchos autores consideran que en realidad su escala tendría que ser considerada como ordinal, lo cierto es que en la práctica a menudo se tratan estas variables como si realmente fueran de intervalo, siempre que su distribución sea aproximadamente normal.

Finalmente, en el segundo caso, utilizan una **escala de razón** variables como la edad, los ingresos o cualquier tipo de recuento en que la existencia de un valor cero significativo permite hacer comparaciones a partir de la magnitud y afirmar que un determinado valor es múltiple de otro.

Como hemos dicho, la distinción entre variables cualitativas y cuantitativas en función de la escala utilizada para definir sus niveles, tiene implicaciones importantes para el proceso de medida. En este sentido, los investigadores tienen que escoger siempre las que mejor reflejen la variabilidad observada en los fenómenos que son objeto de interés y, por lo tanto, aquellas que les permitan recoger adecuadamente la información relativa a la presencia o ausencia de unas características determinadas o, cuando sus estudios lo requieren, el grado o la cantidad con que estas características están presentes en los participantes.

Pero lo que es más relevante para una introducción al análisis multivariante como la que nos hemos propuesto en este texto es que la distinción entre variables cualitativas y cuantitativas tiene también algunas implicaciones importantes para la construcción de modelos complejos que permitan analizar e interpretar múltiples relaciones de manera simultánea.

Por un lado, los investigadores deben conocer, y tener siempre muy presente, la escala de medida de sus variables para incorporarlas adecuadamente en sus modelos. Esto es especialmente relevante cuando se utilizan variables cualitativas, puesto que los valores que representan los diferentes niveles no son más que etiquetas numéricas arbitrarias que sirven para identificar los diferentes grupos de participantes, pero en ningún caso reflejan el grado o la cantidad con que una determinada característica está presente en los individuos. Si bien es cierto que, a veces, es posible tratar como cuantitativas algunas variables que, en principio, tendrían una escala ordinal, los investigadores deberán examinar la distribución y comprobar que, al menos, es aproximadamente normal. Por otro lado, como veremos a continuación, la escala de medida de las variables dependientes e independientes es un condicionante importante en el momento de elección de la técnica de análisis multivariante más adecuada para lograr los objetivos de la investigación.

Una vez abordadas las implicaciones de las características de los datos que proporciona la investigación cuantitativa, estamos en disposición de clasificar las técnicas de análisis multivariante teniendo en cuenta, principalmente, la pregunta o el objetivo general que motiva el proceso de construcción de los modelos. Como nos hemos propuesto, esta clasificación nos permitirá ofrecer una panorámica general sobre sus características y las condiciones en que pueden ser utilizadas, de modo que pueda servir, en última instancia, de guía para orientar a los investigadores en el momento de escoger la técnica que mejor se ajuste a sus objetivos.

La diversidad de técnicas disponibles nos impide abordarlas todas, pero esta clasificación servirá para presentar algunas de las utilizadas más frecuentemente. Para hacerlo, organizaremos esta exposición a partir de los dos grandes contextos de dependencia e interdependencia en que, como hemos dicho, la construcción de modelos multivariantes permite analizar e interpretar las relaciones existentes entre las diferentes variables involucradas en una investigación de manera simultánea y, así, distinguir la contribución independiente de cada una de estas en el sistema de relaciones. A continuación, consideraremos los escenarios particulares en que este marco analítico general puede ser utilizado y propondremos algunas de las alternativas, presentadas esquemáticamente en la tabla 7, de que disponen los investigadores en función de las características de sus datos.

Tabla 7. Una clasificación de las técnicas de análisis multivariante en función de los objetivos de la investigación y de las características de los datos

Objetivo general	Escenario de aplicación	Características de los datos	Técnica multivariante
Analizar relaciones de interdependencia para describir la estructura de los datos	Identificación de grupos de características similares	Diversas variables cuantitativas	Análisis de componentes principales
			Análisis factorial
		Diversas variables cualitativas	Análisis de correspondencias
	Identificación de grupos de individuos similares	Diversas variables cuantitativas o cualitativas	Análisis de conglomerados
	Identificación de grupos de objetos similares	Diversas variables cuantitativas o cualitativas	Escalamiento multidimensional
Analizar relaciones de dependencia para hacer explicaciones o predicciones	Explicación de la variabilidad de los individuos	Una variable dependiente cuantitativa	Regresión múltiple
		Dos o más variables dependientes cuantitativas	Correlación canónica
	Explicación de la variabilidad de los grupos de individuos	Una variable dependiente cuantitativa	ANOVA de dos o más factores o ANCOVA
		Dos o más variables dependientes cuantitativas	MANOVA o MANCOVA
	Predicción de la pertenencia de los individuos a grupos	Una variable dependiente cualitativa	Análisis discriminante
			Regresión logística
Analizar relaciones de dependencia e interdependencia simultáneamente	Evaluación del ajuste de modelos concatenados	Diversas variables cuantitativas	Ecuaciones estructurales

Fuente: elaboración propia.

6. Una guía para la elección de las técnicas de análisis multivariante

De acuerdo con la clasificación presentada en la tabla 7, es posible establecer tres grandes grupos de técnicas en función del objetivo general al cual contribuye el análisis y la interpretación del sistema de relaciones mediante la construcción de modelos multivariantes. A continuación, nos ocuparemos de cada uno de estos tres grandes objetivos, identificaremos los diferentes escenarios de aplicación y presentaremos algunas de las alternativas utilizadas más frecuentemente en función de las características de las variables involucradas.

1) Cuando no es posible distinguir entre variables dependientes e independientes

En este caso, los investigadores se mueven en el contexto de la interdependencia y, por lo tanto, el objetivo general de su análisis es describir la estructura subyacente a sus datos. En este sentido, cuando su intención es analizar las relaciones simultáneas existentes entre diversas variables cuantitativas para identificar grupos de características similares, las técnicas más adecuadas son el **análisis de componentes principales** y el **análisis factorial**.

Las dos técnicas tienen como objetivo reducir la complejidad de los datos mediante la obtención de un conjunto limitado de componentes o factores que permitiría representar la variabilidad en las características de los individuos de una manera eficiente, es decir, conservando el máximo de la información recogida originalmente en las variables involucradas. Tanto el análisis de componentes principales como el análisis factorial se basan en el análisis y la interpretación de las asociaciones observadas entre las variables, pero difieren, básicamente, en la manera de determinar la estructura de componentes o factores.

En el caso del análisis de componentes principales, los investigadores no disponen de una teoría sólida sobre las relaciones para construir sus modelos y, por lo tanto, se limitan a determinar empíricamente la existencia de los componentes que, de hecho, emergen como agrupaciones de sus variables. En cambio, en el caso del análisis factorial, los investigadores parten de una teoría sobre los fenómenos que son objeto de su interés que les informa de los diferentes factores y, por lo tanto, utilizan estos modelos para poner a prueba la contribución de las diferentes variables de acuerdo con sus expectativas. Aunque es importante tener presente que a pesar de que existen algunos procedimientos para tratar variables cualitativas, estas dos técnicas son generalmente aplicadas cuando las variables analizadas son de naturaleza cuantitativa.

En caso de que las variables utilizadas sean cualitativas, los investigadores tienen a su disposición una técnica alternativa, el **análisis de correspondencias**, para lograr los mismos objetivos. Mediante la transformación de la información cualitativa para poder tratarla cuantitativamente, esta técnica procede de una manera comparable y, por lo tanto, permite obtener un conjunto de dimensiones –similares a los componentes o a los factores– que reflejarían una estructura compartida por las variables consideradas en la construcción de los modelos.

Por otro lado, el estudio de las relaciones de interdependencia con el objetivo de describir la estructura subyacente a los datos no solo puede servir para identificar grupos de características similares. Cuando los investigadores están interesados, en cambio, en identificar grupos de individuos, la técnica más adecuada es el **análisis de conglomerados** o **análisis de clúster**.

Esta técnica ofrece un conjunto de procedimientos que permiten reducir la complejidad de los datos mediante la obtención de un conjunto limitado de grupos, exhaustivos y mutuamente excluyentes, que permitiría representar la variabilidad de los individuos a partir de la similitud de sus características. Seleccionadas las variables que formarán parte de los modelos, que pueden ser cuantitativas o cualitativas, y siempre en función del procedimiento escogido por los investigadores, el análisis de conglomerados se basa en el análisis y la interpretación de la asociación observada entre los individuos, de modo que el cálculo de su distancia o proximidad sirve para conformar grupos homogéneos en relación con las características seleccionadas que, a la vez, sean tan heterogéneos entre ellos como sea posible.

Finalmente, cuando el propósito de los investigadores es identificar grupos de objetos similares a partir de las valoraciones que proporcionan los participantes de la investigación, la técnica más adecuada es el **escalamiento multidimensional**. En este caso, a diferencia de lo que sucede con el resto de técnicas de análisis de las relaciones de interdependencia que hemos introducido hasta ahora, la búsqueda de una estructura en los datos no se basa en el análisis y la interpretación de la asociación observada entre las características o los individuos, sino que parte de los juicios comparativos que hacen explícitamente los participantes sobre las parejas formadas a partir de un conjunto de objetos, de acuerdo con sus preferencias o las percepciones de similitud. Como sucede en el caso del análisis de conglomerados, el escalamiento multidimensional puede ser aplicado tanto a variables de naturaleza cuantitativa como cualitativa.

2) Cuando es posible distinguir entre variables dependientes e independientes

Los investigadores se mueven ahora en el contexto de la dependencia y, por lo tanto, el objetivo de su análisis es explicar o predecir las variables dependientes a partir de las independientes. En este sentido, cuando su intención es analizar

las relaciones simultáneas entre diversas variables cuantitativas para explicar la variabilidad de los individuos en una o más de sus características, las técnicas más adecuadas son la **regresión múltiple** y la **correlación canónica**.

Estas dos técnicas tienen como objetivo común determinar la intensidad o la magnitud de las relaciones entre las diferentes variables involucradas, de modo que servirían para evaluar la contribución específica del cambio o la variación en los niveles de todas las variables independientes consideradas en la construcción de los modelos. Además, a pesar de que las variables independientes consideradas en estos modelos suelen ser cuantitativas, las dos técnicas son suficientemente flexibles como para permitir incorporar variables cualitativas mediante la creación de las correspondientes variables ficticias o *dummies*.

Tanto la regresión múltiple como la correlación canónica se basan en el análisis y la interpretación de las asociaciones observadas entre las variables, pero difieren, básicamente, en el número de variables dependientes que permiten explicar. Cuando los investigadores se proponen analizar la variabilidad de los individuos en una característica y, por lo tanto, centran la atención en una única variable dependiente de naturaleza cuantitativa, su técnica de elección es la regresión múltiple. En cambio, podemos entender la correlación canónica como una extensión de la regresión múltiple que permite a los investigadores incorporar diversas variables dependientes cuantitativas a sus modelos y, de este modo, analizar la relación entre dos conjuntos diferenciados de características de los individuos.

Por otro lado, el estudio de las relaciones de dependencia con el objetivo de llevar a cabo explicaciones o predicciones no solo sirve para analizar la variabilidad de los individuos en una o más características. Cuando el propósito de los investigadores es, en cambio, analizar las relaciones simultáneas entre diversas variables con objeto de explicar la variabilidad de los grupos de individuos, las técnicas más adecuadas son el **análisis de la varianza (ANOVA)** de dos o más factores y el **análisis multivariante de la varianza (MANOVA)**.

En este sentido, las dos técnicas comparten el objetivo de determinar la existencia de diferencias entre los individuos de manera agregada, de modo que permitirían evaluar la contribución específica de su pertenencia a diferentes grupos –llamados *factores*– formados a partir de los niveles de una o más variables cualitativas. En este contexto, los factores actuarían como variables independientes en la construcción de los modelos y, como hemos podido ver en relación con el diseño de la investigación, pueden representar tanto grupos naturales, sobre los cuales los investigadores no tendrían ningún tipo de control, como diferentes condiciones experimentales a las que los individuos han sido asignados de manera aleatoria.

El ANOVA de dos o más factores y el MANOVA también se basan en el análisis y la interpretación de las asociaciones observadas entre las variables consideradas en los modelos y, como en el caso de la regresión múltiple y de la corre-

lación canónica, difieren en el hecho de que permiten explicar la variabilidad de los grupos en una o más variables dependientes de naturaleza cuantitativa, respectivamente.

Por otro lado, cuando los investigadores están interesados en considerar otras variables independientes cuantitativas –llamadas *covariantes*– con la intención de ajustar las diferencias entre los grupos en la construcción de sus modelos, las técnicas más adecuadas son el **análisis de la covarianza (ANCOVA)** y el **análisis multivariante de la covarianza (MANCOVA)**. Como extensión de las dos anteriores, estas técnicas resultan especialmente interesantes en el contexto de la investigación observacional, puesto que permiten tener en cuenta la influencia de la variabilidad de los individuos en otras características importantes cuando la asignación a los diferentes grupos no ha sido aleatoria.

Finalmente, más allá de permitir la explicación de la variabilidad en una o más características de los individuos, el estudio de las relaciones de dependencia puede servir también para predecir la pertenencia a diferentes grupos. En este sentido, cuando los investigadores se proponen analizar las relaciones simultáneas entre diversas variables con la intención de clasificar los individuos en los diferentes grupos formados a partir de los niveles de una variable cualitativa, las técnicas más adecuadas son el **análisis discriminante** y la **regresión logística**.

Las dos técnicas tienen como objetivo compartido determinar las características de los individuos que sirven para predecir con acierto los diferentes grupos a los cuales pertenecen, de forma que permitirían evaluar la contribución específica de todas las variables independientes consideradas en la construcción de los modelos. Tanto el análisis discriminante como la regresión logística se basan también en el análisis y la interpretación de las asociaciones observadas entre las variables involucradas, pero difieren, fundamentalmente, tanto en el número de niveles que la variable dependiente cualitativa puede adoptar como en el tipo de variables independientes que permiten considerar en los modelos para hacer las predicciones.

Cuando los investigadores se proponen clasificar con acierto los individuos en relación con los grupos formados por una variable dependiente cualitativa de dos o más niveles y, además, lo quieren hacer tomando en consideración un conjunto de variables independientes cuantitativas, su técnica de elección es el análisis discriminante. Sin embargo, es importante tener presente que esta técnica impone una restricción en relación con las variables independientes, de modo que solo debería ser aplicada cuando sigan una distribución normal.

En cambio, aunque la regresión logística es aplicable únicamente cuando la variable dependiente es dicotómica, y por lo tanto tiene solo dos niveles, el hecho de que haya sido desarrollada como una extensión de la regresión múltiple hace que no tenga que cumplir ninguna restricción y, por lo tanto, per-

mite considerar variables independientes cuantitativas o cualitativas, incorporando estas últimas mediante la creación de las variables ficticias o *dummies* correspondientes.

3) Cuando el contexto no es exclusivamente de interdependencia o de dependencia, sino que hay una combinación de estos dos tipos de relaciones

En este caso, cuando los investigadores están interesados en analizar las múltiples relaciones entre sus variables que pueden ser de dependencia y de interdependencia de manera simultánea, la técnica más adecuada es la de **ecuaciones estructurales**. A diferencia de todas las expuestas anteriormente, esta técnica tiene como objetivo general analizar simultáneamente las múltiples relaciones existentes entre diferentes grupos de variables, de modo que permitiría evaluar el ajuste de varios modelos multivariantes concatenados. Para hacer esto, las ecuaciones estructurales se basan en el análisis y la interpretación de las asociaciones observadas entre diversas variables que, en términos generales, pueden ser organizadas a partir de dos grandes tipos de modelos.

Por un lado, de acuerdo con la lógica del análisis de las relaciones de interdependencia, un «modelo de medida», que sirve para identificar variables latentes, similares a los factores que proporciona el análisis factorial, que representarían una estructura compartida entre diferentes características de los individuos. Por otro lado, de acuerdo con la lógica del análisis de las relaciones de dependencia, un «modelo estructural», que sirve para definir un conjunto de relaciones simultáneas entre variables dependientes e independientes que, por lo tanto, sería equivalente al desarrollo de varios análisis de regresión múltiple o de correlación canónica de manera simultánea.

No obstante, es importante tener presente que, a pesar de que esta técnica se aplica generalmente cuando las variables consideradas en la construcción de estos dos tipos de modelos son cuantitativas, disponemos de algunos procedimientos que permiten tratar también variables de naturaleza cualitativa. De este modo, las ecuaciones estructurales se presentan como la técnica de análisis más eficiente de que disponen los investigadores interesados en abordar fenómenos complejos y que, tomando como punto de partida las evidencias acumuladas en multitud de estudios previos, se proponen poner a prueba o contrastar marcos teóricos sólidos y muy bien definidos.

7. El proceso de construcción de modelos multivariantes

A pesar de la diversidad de técnicas disponibles en función de la pregunta o el objetivo general que motiva la investigación y las características de los datos utilizados para ofrecer una respuesta, acabaremos esta introducción al análisis multivariante abordando el proceso de construcción de los modelos estadísticos complejos que, como hemos dicho, permiten analizar e interpretar las múltiples relaciones existentes entre las diferentes variables involucradas en una investigación de manera simultánea.

Antes de hacerlo, como hemos ido mostrando a lo largo de este texto, es necesario recordar que el análisis multivariante únicamente adquiere todo su sentido en relación con el procedimiento general establecido en el contexto de la investigación cuantitativa que, brevemente, podríamos resumir de la siguiente manera:

- Formular una pregunta o un objetivo general que sirva para abordar un problema relevante.
- Escoger el diseño de la investigación y especificar la muestra de participantes.
- Definir adecuadamente todas las variables involucradas y especificar sus características.
- Desarrollar o escoger los instrumentos necesarios para llevar a cabo las medidas oportunas.
- Recoger las evidencias necesarias que permitan responder a los objetivos de la investigación.
- Resumir y tratar estadísticamente los datos con el objetivo de evaluar las evidencias obtenidas y, cuando sea posible, generalizar las conclusiones más allá de los límites del estudio en particular.

De acuerdo con este procedimiento, el análisis multivariante proporciona el marco analítico general que hace que los investigadores puedan describir, explicar o predecir los fenómenos que son objeto de interés mediante el desarrollo de los modelos estadísticos más adecuados que les permitan llevar a cabo un análisis complejo de sus datos.

En este sentido, con independencia de la técnica escogida, es posible caracterizar la construcción de modelos multivariantes como el proceso general con el que los investigadores pueden obtener una combinación lineal de variables que les permita estimar, a partir de los datos obtenidos de los participantes, el peso específico o la importancia relativa de cada una de ellas y, por lo tanto, evaluar su contribución independiente al sistema de relaciones.

Para hacer esto, los investigadores utilizan las asociaciones observadas entre sus variables como evidencia para determinar la existencia de múltiples relaciones simultáneas en sus modelos y, siguiendo los procedimientos específicos establecidos en función de la técnica escogida, pueden ser capaces de decidir hasta qué punto estos modelos se ajustan o son una buena representación de la realidad que se proponen analizar. Ya sea en el contexto del análisis de las relaciones de interdependencia, de dependencia o en la combinación de los dos, la diversidad de procedimientos que ofrecen las distintas técnicas disponibles siguen esta lógica básica, de modo que comparten unos principios generales y, como veremos a continuación, un conjunto de fases que estructuran el proceso de construcción de este tipo de modelos.

En este sentido, antes de esbozar las fases, resulta conveniente que nos detengamos brevemente en los principios generales que rigen el análisis multivariante, entre los que podemos señalar algunos de los más importantes:

- **La construcción de modelos multivariantes requiere una fundamentación teórica de las relaciones**. Teniendo en cuenta la gran diversidad de posibilidades que, como hemos podido ver, puede ofrecer el análisis multivariante de los datos, es importante tener presente que el punto de partida de cualquier investigación interesada en utilizarlo tiene que ser, necesariamente, la formulación de un problema relevante que permita identificar las relaciones entre las diversas variables involucradas de manera simultánea. En este sentido, resulta indispensable el desarrollo o la adopción de una teoría como el marco general de referencia a partir del cual sea razonable esperar que se produzcan las relaciones que son objeto de interés y que, por lo tanto, sirva de guía a los investigadores para definir sus objetivos particulares, determinar las características de los datos que requerirá la investigación y, en último término, les permita escoger la técnica más adecuada para llevar a cabo el análisis multivariante. En un momento en que el apoyo de los diferentes softwares estadísticos especializados disponibles facilita enormemente la ejecución de este tipo de análisis, el reto importante no es la computación estadística de los modelos multivariantes, sino precisamente todas las decisiones que los investigadores deben tomar para poder construirlos con éxito.

- **La exploración de los datos es una condición previa necesaria para el desarrollo del análisis multivariante**. Como extensión del análisis univariante y bivariante, los investigadores no deberán perder de vista la importante contribución de estos dos tipos de análisis en el momento de tomar contacto con sus datos. Si el hecho de disponer de un marco teórico sólido es una condición indispensable para construir modelos complejos que permitan analizar e interpretar múltiples relaciones de manera simultánea, no es menos cierto que, únicamente cuando los investigadores se hayan familiarizado con la distribución de las variables involucradas y hayan examinado sus relaciones por parejas, estarán en disposición de considerar la conveniencia de llevar a cabo un análisis multivariante para responder a los objetivos. Esta exploración de los datos es especialmente relevante en el contexto del análisis de las relaciones de dependencia que, como hemos podido ver, permite obtener los indicios necesarios para considerar el papel de cualquier factor o variable de confusión que pueda interferir en las relaciones que son objeto de interés y, por lo tanto, controlar estadísticamente su influencia en los modelos multivariantes con objeto de evitar que pueda convertirse una explicación alternativa a los resultados obtenidos.

- **El cumplimiento de los supuestos es un requerimiento importante para la aplicación de las técnicas de análisis multivariante**. La exploración inicial de los datos no solo sirve para determinar la conveniencia del análisis multivariante, sino que, además, permite comprobar hasta qué punto se cumplen los supuestos que asume la técnica escogida para modelar las múltiples relaciones entre variables de manera simultánea. En este sentido, tal como hemos expuesto en la clasificación de las diferentes técnicas disponibles, es importante que los investigadores corroboren que tanto la naturaleza de las relaciones que se proponen analizar como las características de las variables implicadas se ajustan a los requerimientos de la técnica escogida. Por otro lado, más allá de los requerimientos estadísticos particulares de cada una de las técnicas disponibles, es importante tener presente que la inferencia estadística asume también algunos supuestos importantes en relación con la distribución aproximadamente normal de las variables, la linealidad de las relaciones o la homogeneidad de las varianzas de las variables dependientes a lo largo de los diferentes niveles de las independientes. No es este el lugar para profundizar en esta cuestión, pero es importante tener presente que solo cuando se garantiza el cumplimiento de estos supuestos es posible generalizar los resultados de la investigación más allá de los límites de los estudios particulares.

- **La clave del éxito del análisis multivariante se encuentra en una especificación adecuada de los modelos**. Un cuarto principio importante para la construcción de modelos multivariantes es la selección de las variables que finalmente formarán parte del análisis. Una vez fundamentadas teóricamente las relaciones y, por lo tanto, de acuerdo con sus objetivos particulares, los investigadores deberán decidir cuáles, de entre todas

las variables disponibles, serán utilizadas en el momento de especificar los modelos. En este sentido, es conveniente tener presente que es tan importante seleccionar todas las variables que sean pertinentes desde el punto de vista teórico –y, por lo tanto, no dejarse ninguna importante–, como lo es evitar incluir cualquier otra variable que, en realidad, no sea relevante para analizar los fenómenos que son objeto de interés. Es lo que denominamos una *especificación adecuada de los modelos*, que en ningún caso se corresponde con una decisión única, sino que forma parte del proceso contingente e iterativo de construcción de los modelos multivariantes a través del cual los investigadores añaden y quitan variables de sus modelos en función de los resultados que les proporcionan. El objetivo último de este proceso es, además, la obtención de unos modelos que sean parsimoniosos, es decir, capaces de representar la máxima complejidad de los fenómenos que son objeto de interés con el número más pequeño posible de variables.

- **No es posible interpretar las relaciones entre las variables sin una evaluación previa de los modelos**. A pesar de que no es posible tener todas las garantías sobre la especificación correcta de los modelos multivariantes y teniendo en cuenta, por lo tanto, que no pueden ser nunca utilizados como prueba definitiva o concluyente para determinar si una teoría es o no correcta, lo cierto es que las decisiones que toman los investigadores durante todo este proceso afectan a los resultados de su análisis y, en consecuencia, condicionan necesariamente el papel que acaban jugando las diferentes variables implicadas en el sistema de relaciones. Incluir u omitir una determinada variable puede hacer que los modelos multivariantes se comporten de manera diferente y, precisamente por esta razón, es necesario evaluar hasta qué punto la combinación de variables escogida se ajusta razonablemente bien a la variabilidad observada en los datos y que, por lo tanto, sus modelos son una representación adecuada de la realidad. En este sentido, es importante tener presente que, como representación simplificada de la realidad, todos los modelos son incompletos y, por lo tanto, necesariamente incorrectos, pero cuando se centran en los aspectos sustanciales de los fenómenos se convierten en una herramienta muy útil para interpretar las múltiples relaciones simultáneas que son objeto de interés para la investigación.

- **El diseño de la investigación condiciona la inferencia estadística basada en el análisis multivariante**. Como hemos explicado, la capacidad de los investigadores para extraer conclusiones generales a partir del análisis de los datos de una muestra está estrechamente relacionada con el diseño utilizado para desarrollar la investigación. Como sucede con cualquier otra técnica estadística, tanto la inferencia causal como la inferencia a la población que permite el análisis multivariante solo es posible si la selección o la asignación de los participantes a los diferentes grupos han sido aleatorias. Es decir, únicamente cuando el azar interviene en al menos uno de estos dos momentos importantes para el diseño de la investigación, es

posible disponer de las garantías suficientes para decidir si las múltiples asociaciones simultáneas observadas entre las variables son una evidencia adecuada para determinar, con cierta confianza, la existencia de relaciones causales generalizables en la población. A pesar de la complejidad de los fenómenos que permite abordar, es importante tener siempre presente que la construcción de modelos multivariantes no exime a los investigadores de su responsabilidad en relación con la valoración de la adecuación de las evidencias que han obtenido para establecer sus inferencias.

Finalmente, estamos en disposición de recapitular las diferentes fases que, de manera general, permiten estructurar el proceso de construcción de modelos multivariantes. Teniendo en cuenta el procedimiento establecido en la investigación cuantitativa a partir del cual el análisis multivariante adquiere su sentido y, particularmente, tomando como punto de partida los principios que acabamos de presentar, estas fases ofrecen una perspectiva de conjunto sobre las cuestiones más importantes que hemos ido explicando a lo largo de este texto y, además, permiten poner en práctica todos los conocimientos, las habilidades y los valores vinculados a la construcción de modelos multivariantes.

Atendiendo a su carácter general y, por lo tanto, con independencia de las especificidades de los procedimientos particulares con que deben ser aplicadas las diferentes técnicas disponibles, estas diez fases fundamentales sirven para organizar secuencialmente las diferentes decisiones que los investigadores deben tomar para llevar a cabo un análisis complejo de sus datos mediante la construcción de modelos multivariantes. De este modo:

1) Delimitación del propósito del análisis. La construcción de modelos multivariantes empieza siempre con una definición precisa de los objetivos particulares a partir de los cuales los investigadores se proponen analizar e interpretar las múltiples relaciones entre diversas variables de manera simultánea. Como hemos podido ver, las diferentes técnicas de análisis multivariante disponibles pueden ser utilizadas con multitud de finalidades que, a todos los efectos, permiten a los investigadores describir, explicar o predecir los fenómenos que son objeto de su interés. Como consecuencia de la formulación de un problema relevante con una fundamentación teórica adecuada, un propósito muy definido es el primer paso para afrontar con éxito el proceso de construcción de modelos multivariantes.

2) Elección de la técnica de análisis. Una vez se ha definido el propósito del análisis multivariante, el segundo paso consiste en escoger la técnica más adecuada. De acuerdo con la clasificación de las técnicas presentada anteriormente, es necesario que los investigadores decidan si se mueven en el contexto del análisis de las relaciones de dependencia o de interdependencia, que identifiquen el escenario particular en que se puede concretar el estudio de las relaciones en estos dos contextos y, finalmente, que identifiquen las características de los datos de que disponen. Esta es una decisión importante en

el proceso de construcción de modelos multivariantes, puesto que la técnica escogida condiciona los procedimientos que los investigadores deben llevar a cabo a lo largo de las fases siguientes.

3) Exploración inicial de los datos. Una vez seleccionada la técnica de análisis multivariante más adecuada, los investigadores deben familiarizarse con la distribución de las variables involucradas y, a continuación, examinar las relaciones por parejas. Mediante la aplicación de técnicas de análisis univariante y bivariante, este primer contacto con los datos permite a los investigadores determinar la conveniencia de llevar a cabo un análisis multivariante para responder a sus objetivos particulares. Como hemos dicho, esta es una fase importante para el análisis de las relaciones de dependencia que permite considerar la existencia de potenciales factores o variables de confusión que sería conveniente tener en cuenta en el proceso de construcción de los modelos multivariantes.

4) Comprobación de los supuestos. La exploración de los datos debe servir, también, para determinar hasta qué punto es conveniente aplicar la técnica escogida. Por un lado, desde el punto de vista teórico, confirmando que sirve para analizar el sistema de relaciones que los investigadores se proponen abordar. Por otro lado, desde el punto de vista de las características de sus datos, asegurando que la distribución de las variables se ajusta a los requerimientos estadísticos particulares de la técnica. Finalmente, desde el punto de vista de la inferencia, garantizando que los datos cumplen también con los requerimientos estadísticos adicionales que implica, cuando los investigadores tienen este objetivo, la generalización de los resultados a la población que representa la muestra.

5) Estimación del modelo. La exploración de los datos y la comprobación de los supuestos para poder aplicar las técnicas dan paso, ahora sí, a la computación estadística de los modelos multivariantes. Siguiendo los procedimientos establecidos para la técnica escogida, y siempre con el apoyo del software estadístico adecuado, es el momento en que los investigadores obtienen la combinación lineal de variables que les permitirá estimar el peso específico o la importancia relativa de cada una de ellas en el sistema de relaciones. Como veremos a continuación, esta estimación no es más que un resultado inicial en el proceso de construcción de modelos multivariantes que deberá ser evaluado y, si procede, revisado a lo largo de las fases siguientes.

6) Evaluación del ajuste del modelo. Como hemos dicho, la interpretación de las relaciones entre las diferentes variables consideradas en los modelos requiere un análisis de su comportamiento global que, de acuerdo con los procedimientos específicos de cada técnica, permita determinar hasta qué punto se ajustan a la variabilidad observada en los datos y, por lo tanto, resulta razonable aceptar que son una representación adecuada de los fenómenos objeto de interés. Teniendo en cuenta estas evidencias, los investigadores deberán

llevar a cabo los juicios mediante la comparación del ajuste de las sucesivas variantes que, como aproximaciones complementarias o alternativas, puedan obtener en el proceso de construcción de sus modelos.

7) Revisión y mejora del modelo. La evaluación del ajuste de los modelos conduce a la séptima fase, en la que los investigadores valoran la conveniencia de añadir o quitar variables relevantes desde el punto de vista teórico teniendo en cuenta sus efectos en el comportamiento global de los modelos. Es importante recordar que, en cualquier caso, el objetivo final de este proceso es obtener modelos multivariantes bien especificados, que además sean parsimoniosos, de modo que los investigadores deben ser capaces de hacer un balance adecuado entre el incremento del nivel de complejidad de los modelos y los beneficios que esto comportaría en relación con la mejora sustantiva en su ajuste a los datos.

8) Interpretación del sistema de relaciones. Una vez logrado un ajuste global aceptable, llega el momento en que los investigadores pueden utilizar sus modelos multivariantes para analizar e interpretar las relaciones existentes entre las diversas variables implicadas. En este sentido, la combinación lineal de variables que proponen les sirve para estimar los pesos o las ponderaciones asociadas a cada una de ellas y, por lo tanto, evaluar la contribución independiente al sistema de relaciones. Cuando el objetivo del análisis es establecer inferencias causales o a la población, esta interpretación permite determinar la significación estadística de las asociaciones observadas en la muestra y, lo que todavía es más importante, la significación que estas relaciones tienen en la práctica.

9) Validación del modelo final. A pesar de que no siempre es posible, el proceso de construcción de modelos multivariantes debería considerar la conveniencia de poner a prueba la eventual generalización de las conclusiones más allá de los límites de los estudios particulares. De este modo, los investigadores deberían disponer de una muestra de participantes diferente de la que han utilizado para modelar las relaciones, o al menos dividir la muestra en dos partes, para poder ofrecer evidencias que permitan confiar en que los modelos no son una consecuencia de las especificidades de la muestra y que, en cambio, pueden ser útiles para analizar e interpretar las relaciones en el conjunto de la población.

10) Comunicación de los resultados del análisis. El proceso de construcción de modelos multivariantes concluye con la elaboración de un informe o una publicación científica que sirve para comunicar las principales conclusiones a las que ha llegado la investigación. Teniendo en cuenta la complejidad de los fenómenos que permite abordar, el análisis multivariante debe ir siempre acompañado de un esfuerzo especial por parte de los investigadores para transmitir y hacer accesibles los resultados obtenidos. En este sentido, es especialmente relevante el uso de un lenguaje sencillo pero cuidadoso que permita

reflejar adecuadamente la naturaleza de las relaciones observadas y, cuando es el objetivo, hasta qué punto es posible utilizar las evidencias obtenidas para establecer inferencias causales o a la población que representa la muestra.

8. Bibliografía anotada

Finalmente, entendiendo que por razones de espacio no es posible abordar el detalle de la ejecución y la interpretación de los resultados que proporcionan las diferentes técnicas disponibles en una introducción general, ni hacerlo con todo el detalle y la profundidad que pueden dedicarle otras contribuciones más avanzadas, los lectores interesados pueden encontrar varios manuales sobre análisis multivariante que les pueden ayudar a complementar y ampliar esta aproximación.

En este sentido, disponemos de algunos manuales publicados en nuestro entorno más cercano que sería interesante tener en cuenta como, por ejemplo, los de Martínez (1999), Díaz (2002), Peña (2002), Catena, Ramos y Trujillo (2003), Pérez (2004), Ximénez y San Martín (2004) o, más recientemente, de la Garza, Morales y González (2013), López-Roldán y Fachelli (2015) o Aldas y Uriel (2017).

De manera complementaria, a continuación ofrecemos una selección de algunas contribuciones relevantes desarrolladas en el contexto internacional que pueden ser muy útiles para adquirir una visión más amplia sobre la lógica general del análisis multivariante, las particularidades de las diferentes técnicas, sus fundamentos matemáticos o la utilización del software estadístico especializado disponible:

Harlow, L. L. (2014). *The essence of multivariate thinking. Basic themes and methods* (2.ª ed.). Nueva York: Routledge.

Como introducción general, este libro es una aproximación excelente a los aspectos básicos del análisis multivariante de los datos. Tomando como punto de partida una primera parte que, a modo de visión de conjunto, presenta las cuestiones transversales que comparten las diferentes técnicas, la autora dedica los capítulos siguientes a exponerlas una por una remarcando sus diferencias y similitudes, los objetivos particulares que permiten alcanzar, sus asunciones básicas y la manera en que deben ser interpretados los resultados que proporcionan. Reduciendo al mínimo la presencia de las cuestiones matemáticas relacionadas con la computación estadística de los modelos multivariantes, este manual propone una aproximación conceptual muy asequible que, además, está ilustrada con diferentes ejemplos prácticos desarrollados mediante la utilización de SPSS y SAS.

Miller, J. E. (2013). *The Chicago guide to writing about multivariate analysis* (2.ª ed.). Chicago: The University of Chicago Press.

A pesar de que este libro se centra principalmente en el desarrollo de las habilidades necesarias para llevar a cabo una presentación efectiva de los resultados proporcionados por las diferentes técnicas, su lectura es también muy recomendable para los lectores interesados en familiarizarse con algunas cuestiones básicas que van más allá de la computación estadística de los modelos multivariantes. En este sentido, adopta como principio la necesidad de hacer accesibles las interpretaciones de los investigadores sobre las relaciones entre las diversas variables involucradas en sus estudios, de manera que, en última instancia, puedan contribuir efectivamente a enriquecer el debate público sobre los fenómenos que son objeto de su interés adaptándose a las expectativas y necesidades de los diferentes tipos de audiencias.

Hair, J. F., Black, W. C., Babin, B. J., y Anderson, R. E. (2013). *Multivariate data analysis. Pearson new international edition* (7.ª ed.). Harlow: Pearson.

Esta es una referencia obligada para los lectores interesados en profundizar en el proceso de construcción de modelos multivariantes. Dado que se trata de un manual que ha sido revisado y actualizado regularmente durante las últimas tres décadas, los autores han llegado al punto de equilibrio necesario que permite combinar, de manera sencilla y accesible para un público que no tiene una formación estadística avanzada, los principales retos vinculados a la ejecución de cada una de las técnicas de análisis multivariante y su aplicabilidad en la práctica, tanto en contextos académicos como profesionales. A pesar de que los ejemplos que utiliza para ilustrar las técnicas son propios de la investigación desarrollada en el área de los estudios de marketing, todas las orientaciones, consejos y recomendaciones prácticas resultan igualmente interesantes para el resto de disciplinas basadas en el análisis de datos cuantitativos.

Tabachnick, B. G., y Fidell, L. S. (2019). *Using multivariate statistics* (7.ª ed.). Nueva York: Pearson.

Otra gran referencia en el campo de los manuales que se proponen ofrecer una introducción general a las diferentes técnicas de análisis multivariante disponibles ha cumplido también tres décadas de revisión y actualización, que han hecho que se convierta en uno de los libros de cabecera para muchos de los profesionales e investigadores que trabajan con datos cuantitativos. Con una orientación eminentemente práctica, este texto centra la atención en una discusión profunda de los tipos de preguntas u objetivos que las diferentes técnicas disponibles permiten responder, sus limitaciones y todas aquellas cuestiones importantes que es necesario tener en cuenta en su aplicación a los datos obtenidos en la investigación, la manera de hacerlo mediante el software estadístico especializado que ilustra con ejemplos para SPSS, SAS y SYSTAT, y las estrategias más convenientes para garantizar un análisis riguroso y honesto por parte de los investigadores.

Pituch, K. A., y Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences* (6.ª ed.). Nueva York: Routledge.

La última actualización de este manual nos ofrece una tercera aproximación que, con carácter general, se ocupa de las particularidades de las diferentes técnicas de análisis multivariante. A pesar de que dedica una parte de sus esfuerzos a introducir a los lectores en el álgebra matricial en que se basa la combinación lineal de variables, lo cierto es que la exposición que hace en el resto de capítulos sobre las técnicas, sus supuestos y las condiciones en que pueden ser aplicadas no obliga a profundizar en los formalismos matemáticos que, muy probablemente, podrían alejar a los lectores que no tienen una formación estadística avanzada. En este sentido, a pesar de que algunos de los ejercicios que propone se basan en el cálculo manual, también ofrece una buena ilustración de los diferentes capítulos con ejemplos prácticos para SPSS y SAS.

Hahs-Vaughn, D. L. (2017). *Applied multivariate statistical concepts*. Nueva York: Routledge.

Con un carácter marcadamente accesible, este manual tiene por objeto presentar los conceptos más importantes dejando de lado, en la medida de lo posible, toda la formulación matemática. Planteado como un texto introductorio, consigue de manera notable poner por delante la racionalidad del análisis multivariante, con el supuesto de que, teniendo en cuenta el nivel de sofisticación del software especializado disponible para ejecutar el análisis propiamente dicho, lo que es realmente importante es que los lectores interesados en el uso de las técnicas disponibles dominen la manera de implementarlas e interpretar los resultados que proporcionan. Sin embargo, cada capítulo proporciona un apartado específico orientado a la exploración de los datos y una revisión de los supuestos que es necesario comprobar, así como también incluye una pequeña panorámica matemática que permite conocer los aspectos más técnicos vinculados a cada técnica. Finalmente, el anexo también dispone de una introducción al álgebra matricial que sustenta todos los cálculos.

Everitt, B, y Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Nueva York: Springer.

Este es un buen ejemplo de los manuales que remarcan la utilización del software estadístico especializado que permite llevar a cabo el análisis multivariante. En este sentido, el texto adopta una aproximación eminentemente práctica al análisis multivariante poniendo en el centro de su atención los procedimientos disponibles en el marco del software estadístico de código abierto R. En cualquier caso, este es un manual asequible para los lectores que no tienen unos conocimientos avanzados en el uso de R y, además, proporciona ejemplos abundantes junto con todo el código necesario para llevar a cabo con éxito los ejercicios que, de manera práctica, complementan todas sus explicaciones.

Anderson, T. W. (2003). *An introduction to multivariate statistic analysis* (3.ª ed.). Nueva Jersey: Wiley-Interscience.

La tercera edición de este libro clásico, publicado por primera vez a finales de los años cincuenta como compilación de materiales docentes, es probablemente la referencia de elección para los lectores interesados en profundizar en los aspectos vinculados al tratamiento matemático de los datos en que se basan las diferentes técnicas de análisis multivariante. Este es, por lo tanto, un texto de un nivel avanzado que presenta, de manera estructurada, la teoría, los métodos y las demostraciones matemáticas necesarias para abordar rigurosamente los diferentes procedimientos estadísticos derivados de la distribución normal multivariante y sus características. A pesar de que exige un cierto dominio del álgebra matricial para seguir las explicaciones, incorpora un anexo que sirve como introducción básica a las definiciones y los teoremas que permiten desarrollar los cálculos de manera autónoma.

Johnson, R., y Wichern, D. (2019). *Applied multivariate statistic analysis* (6.ª ed.). Nueva Jersey: Pearson.

En un sentido similar, este libro pertenece a la categoría de los manuales que priorizan los aspectos matemáticos del análisis multivariante y, por lo tanto, está también dirigido a los lectores que tienen unos conocimientos estadísticos avanzados. A pesar de que, a diferencia del anterior, introduce de manera natural los aspectos básicos relacionados con el álgebra matricial, su objetivo es proporcionar las demostraciones matemáticas que permiten entender el funcionamiento de las diferentes técnicas, sus posibilidades y limitaciones, así como también los resultados que proporciona cada una. Para hacerlo, los autores presentan diferentes ejercicios que sirven para ilustrar las explicaciones, algunos basados en ejemplos ficticios simplificados para que sea posible resolverlos a mano, y otros, más complejos, a partir de datos reales que exigen la utilización de algún software estadístico especializado.

Brown, B. L., Hendrix, S. B., Hedges, D. W., y Smith, T. B. (2011). *Multivariate analysis for the biobehavioral and social sciences*. *A graphical approach*. Hoboken: John Wiley & Sons.

Finalmente, este texto adopta una perspectiva innovadora para abordar las diferentes técnicas de análisis multivariante. Poniendo la representación gráfica en el centro de la discusión, los autores desarrollan una aproximación que combina los aspectos matemáticos esenciales implicados en la computación de los modelos multivariantes con una exposición detallada del uso de SAS, STATA y SPSS para hacerlo. Más allá de la preceptiva introducción al álgebra matricial que caracteriza este tipo de manuales avanzados, cada capítulo viene acompañado de un ejemplo real extraído de una publicación científica que permite contextualizar los diferentes tipos de análisis y, en último término,

poner en evidencia cómo los gráficos son aliados poderosos que permiten representar adecuadamente el significado de los datos también en el contexto del análisis multivariante.

Bibliografía

Aldas, J., y Uriel, E. (2017). *Análisis multivariante aplicado con R* (2.ª ed.). Madrid. Paraninfo.

Aldrich, A. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science, 10*(4), 364-376.

Anderson, T. W. (2003). *An introduction to multivariate statistic analysis* (3.ª ed.). Nueva Jersey: Wiley-Interscience.

Bickel, P. J., Hammel, E. A., y O'Connell, J. W. (1975). Sexbias in graduate admisiones: Data from Berkeley. *Science*, 187, 398-404.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67*(338), 364-366.

Brown, B. L., Hendrix, S. B., Hedges, D. W., y Smith, T. B. (2011). *Multivariate analysis for the biobehavioral and social sciences. A graphical approach*. Hoboken: John Wiley & Sons.

Catena, A., Ramos, M. M., y Trujillo, H. M. (2003). *Análisis multivariado. Un manual para investigadores*. Madrid: Biblioteca Nueva.

Coolican, H. (2014). *Research methods and statistics in psychology* (6.ª ed.). Londres: Psychology Press.

Cozby, P. C., y Bates, S. C. (2015). *Methods in behavioral research* (12.ª ed.). Nueva York: McGraw-Hill.

David, H. A., y Edwards, A. W. F. (2001). *Annotated readings in the history of statistics*. Nueva York: Springer.

Díaz, V. (2002). *Técnicas de análisis multivariante para investigación social y comercial. Ejemplos prácticos utilizando SPSS versión 11*. Madrid: Ra-Ma.

Everitt, B., y Hothorn, T. (2011). *An introduction to applied multivariate analysis with R* . Nueva York: Springer.

Freedman, D., Pisani, R., y Purves, R. (2007). *Statistics* (4.ª ed.). Nueva York: W. W. Norton & Company.

de la Garza, J., Morales, B. N., y González, B. A. (2013). *Análisis estadístico multivariante*. México: McGraw-Hill.

Hahs-Vaughn, D. L. (2017). *Applied multivariate statistical concepts*. Nueva York: Routledge.

Hair, J. F., Black, W. C., Babin, B. J., y Anderson, R. E. (2013). *Multivariate data analysis. Pearson new international edition* (7.ª ed.). Harlow: Pearson.

Harlow, L. L. (2014). *The essence of multivariate thinking. Basic themes and methods* (2.ª ed.). Nueva York: Routledge.

Johnson, R., y Wichern, D. (2019). *Applied multivariate statistic analysis* (6.ª ed.). Nueva Jersey: Pearson.

Light, R. J., Singer, J. D., y Willett, J. B. (1990). *Bydesign. Planning research on higher education*. Cambridge: Harvard University Press.

López-Roldán, P., y Fachelli, S. (2015). Metodología de la investigación social cuantitativa. Barcelona: Universitat Autònoma de Barcelona. Disponible en: https://ddd.uab.cat/record/129382.

Meneses, J., Barrios, M., Bonillo, A., Cosculluela, A., Lozano, L. M., Turbany, J., y Valero, S. (2014). *Psicometría*. Barcelona: Editorial UOC.

Martínez, R. (1999). *El análisis multivariante en la investigación científica*. Madrid: La Muralla.

Miller, J. E. (2013). *The Chicago guide to writing about multivariate analysis* (2.ª ed.). Chicago: The University of Chicago Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Nueva York: Cambridge University Press.

Peña, D. (2002). *Análisis de datos multivariantes*. Madrid: McGraw-Hill.

Pérez, C. (2004). *Técnicas de analisis multivariante de datos. Aplicaciones con SPSS*. Madrid: Pearson.

Pituch, K. A., y Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences*. (6.ª ed.). Nueva York: Routledge.

Russo, F. (2009). *Causality and causal modelling in the social sciences*. Nueva York: Springer.

Shadish, W. R., Cook, T. D., y Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2.ª ed.). Boston: Houghton Mifflin.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13*(2), 238-241.

Tabachnick, B. G., y Fidell, L. S. (2019). *Using multivariate statistics* (7.ª ed.). Nueva York: Pearson.

Ximénez, M. C., y San Martín, R. (2004). *Fundamentos de las técnicas multivariantes*. Madrid: UNED.

Yule, G. U. (1903). Notes on the theory of association of attributes of statistics. *Biometrika, 2*(2), 121-134.