Aspectos básicos del análisis multivariante

Julio Meneses

PID_00199866

Universitat Oberta de Catalunya

Índice

Introducción..... 5

1. El caso de la discriminación de género en la Universidad de Berkeley..... 7

2. Asociación, confusión y causalidad..... 11

3. Diseño de la investigación e inferencia estadística..... 16

4. ¿Qué es el análisis multivariante y para qué sirve?..... 21

5. Una clasificación de las técnicas de análisis multivariante.... 26

6. Una guía para la elección de las técnicas de análisis multivariante..... 30

7. El proceso de construcción de modelos multivariantes..... 35

Bibliografía..... 43

Introducción

Esta introducción a los aspectos básicos del análisis multivariante tiene como objetivo general proporcionar al lector algunas claves importantes para abordar los conocimientos, las habilidades y los valores vinculados con el desarrollo de modelos estadísticos que permiten llevar a cabo un análisis complejo de los datos obtenidos en la investigación cuantitativa. Tomando como punto de partida un estudio clásico sobre la discriminación por razón de género de la Universidad de Berkeley, expondremos este caso controvertido como un ejemplo de investigación en el que la omisión de una información relevante para el análisis puede conducir a una conclusión inadecuada. Esta discusión servirá para introducir algunos conceptos importantes, como la asociación, la confusión y la causalidad, así como reconocer la importancia del diseño de la investigación para poder extraer conclusiones no sesgadas, que, además, sean generalizables más allá de los límites de los estudios particulares. A continuación, presentaremos el análisis multivariante como el marco analítico general que permite modelar las múltiples relaciones existentes entre diversas variables de forma simultánea, describiremos sus objetivos principales y presentaremos una clasificación general de las diferentes técnicas disponibles, que nos permitirá ofrecer una guía para orientar a los investigadores en el momento de escoger aquella que mejor se ajuste a su investigación. Esta exposición servirá para situar este tipo de análisis en el contexto general de la investigación cuantitativa y, finalmente, presentar algunos de los principios que rigen las diferentes fases con que es posible estructurar el proceso de construcción de modelos multivariantes. De este modo, trataremos de poner las bases necesarias a partir de las cuales se desarrollará la exposición de las diferentes técnicas de análisis multivariante que abordaremos en detalle en el resto de módulos que siguen a esta introducción general.

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

1973 fue un año interesante para la discusión sobre la situación de las mujeres en el mundo universitario en Estados Unidos. Resueltas las solicitudes de acceso para el comienzo del curso aquel otoño, 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 sus estudiantes a los programas de posgrado. En este sentido, examinando los datos recogidos en los archivos de los diferentes departamentos, el profesor Hammel, entonces decano de estos estudios, se encontró con una situación, como mínimo, aparentemente paradójica (Bickel, Hammel y O'Connell, 1975).

Teniendo en cuenta el conjunto global de solicitudes, aquel curso se presentó 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. 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 incontestable 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 a los programas de posgrado de la Universidad de Berkeley (X^2=111,25, df=1, p=0,000). A pesar de ser estadísticamente significativa, no obstante, esta relación no muestra una intensidad o una magnitud importante (V de Cramér=0,09).

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).

Si asumimos, y no tenemos evidencias para no hacerlo así, que las mujeres y los hombres no difieren de manera significativa en sus capacidades, aptitudes y habilidades, la Universidad de Berkeley estaría prefiriendo a los hombres por delante de las mujeres como estudiantes de sus programas. Esta situación, sin embargo, resulta más compleja que la representación que ofrece el análisis de esta tabla de contingencia. Tal como muestran Bickel, Hammel y O'Connell (1975), la aparente discriminación por razón de género se produce únicamente cuando agregamos los datos para el conjunto de la universidad. A pesar de que en su trabajo no reproducen los datos proporcionados por cada uno de los ciento un departamentos que ofrecían estos estudios, su análisis sirve como una 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 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, bien 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.

Con todo, 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 a la minoría que lo hacían 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 contradicción. Tomando en consideración los datos proporcionados por los seis departamentos más grandes, que habían evaluado aproximadamente a un tercio de los candidatos de toda la universidad, registraron el número de solicitudes y calcularon sus respectivas tasas de admisión. La tabla 2 recoge estos datos, desagregando las solicitudes en función del género de los candidatos. Tal como se puede observar, los porcentajes de admisión son bastante similares a estos seis departamentos. La excepción más notable es el departamento A, que mostró una preferencia importante por las mujeres aceptando a un 82 % de ellas en comparación al 62 % de los hombres. En el sentido contrario, el departamento E mostró una preferencia más clara por los hombres aceptando un 28 % de ellos en comparación al 24 % de las mujeres. En cambio, si nos fijamos en las solicitudes a los seis departamentos en su conjunto, la relación entre el género de los candidatos y su aceptación a los programas de posgrado vuelve a ser evidente a favor de los hombres, con una tasa global del 44 % en comparación a la del 30 % en el caso de las mujeres.

Tabla 2. Datos de admisión en 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).

Una diferencia de 14 puntos entre los hombres y las mujeres en la tasa global de aceptación de los seis departamentos más grandes volvería a ser una evidencia incontestable 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 en esta aparente contradicción. Teniendo en cuenta sus respectivas tasas de aceptación, los departamentos A y B serían aquellos que más solicitudes aceptaron finalmente y, por lo tanto, aquellos en los que resultó más fácil acceder para los candidatos, fueran hombres o mujeres, que se presentaron. 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 aquellos que más dificultades pusieron a los candidatos, fueran hombres o mujeres, porque finalmente resolvieron favorablemente 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 a sus programas.

Los departamentos, por lo tanto, no mostraron un comportamiento similar en relación con la aceptación de sus estudiantes. Sin embargo, lo que es más importante para entender la aparente contradicción de este caso de discriminación por razón de género es que 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 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, lo que representa casi la práctica totalidad 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 a los 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 su elección del departamento como hacemos en la tabla 2, la relación entre el género de los candidatos y su aceptación final a 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 (Bickel, Hammel y O'Connell, 1975) se ha convertido en un ejemplo clásico de un fenómeno que se produce a menudo 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 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 de Pearson sobre la existencia de correlaciones espurias entre variables cuantitativas (Aldrich, 1995; David y Edwards, 2001). En este sentido, podemos definir este fenómeno, aparentemente paradójico, 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 ciencias sociales, particularmente en los estudios observacionales, y resulta especialmente sorprendente a los 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 sus solicitudes de acceso en el conjunto de los estudios que ofrece y, a la vez, no hacerlo o, incluso, discriminar ligeramente a los hombres en cada uno de los departamentos que la componen. En ningún caso, sin embargo, 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 estadísticamente significativas existen, son reales, tanto en el caso del conjunto de los candidatos valorados por la universidad como en el detalle de sus departamentos. En este sentido, lo que pone de manifiesto la contradicción no es la existencia de estas relaciones, sino el hecho de que las evidencias observadas de asociación entre las variables sean utilizadas como prueba para llevar a cabo juicios causales. Dado 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 *variable de confusión* es una variable extraña, no contemplada en la investigación, que puede alterar la relación entre dos variables 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 predictora o explicativa– y una *variable dependiente* –también conocida como resultado o explicada–, una tercera variable sería un factor de confusión si su incorporación al análisis comportara el incremento, el decremento, 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 forma 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 por lo que, como todos los investigadores tendrían que 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, el lector interesado puede encontrar una introducción general al 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 seminal 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 el que la omisión de una variable de confusión en el análisis agregado para el conjunto de los departamentos conduce a una conclusión sesgada. Tal 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 sería consecuencia de la preferencia de los hombres y las mujeres por aquellos más fáciles y más difíciles de acceder, respectivamente. No disponiendo 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, que el departamento cumple con la condición de estar asociado tanto con el género de los candidatos como con su aceptación final. 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 hacer evidente una relación que ni siquiera había sido observada inicialmente.

Imaginemos una universidad ficticia compuesta, para simplificar el análisis, únicamente por dos departamentos. Teniendo en cuenta el conjunto global de solicitudes, suponemos que se presentaron un total de mil candidatos, de los cuales 450 serían hombres y 450 mujeres, y que de estos candidatos finalmente un 60 % tanto de los hombres como de las 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. En este caso, teniendo en cuenta que la tasa global de aceptación en el conjunto de los departamentos habría sido del 60 % tanto para los hombres como para las mujeres, el hecho de que no haya ninguna diferencia sería una evidencia incontestable en contra de la existencia de una discriminación por razón de género. Si utilizamos los datos totales que se presentan en la última hilera 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).

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.

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

un 76 % y aproximadamente un 69 %, respectivamente. Una diferencia entre los hombres y las mujeres de 20 puntos en el departamento A y de 17 puntos en el departamento B sería una evidencia incontestable a favor de la existencia de una discriminación por razón de género. Los dos departamentos de esta universidad estarían, en realidad, prefiriendo a los hombres por delante de las mujeres como estudiantes de sus programas. De hecho, si utilizamos los datos que se presentan en la primera y en la segunda hilera para construir dos tablas de contingencia separadas, el análisis de su 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). La intensidad o magnitud de esta relación es, con todo, más importante en el 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 o una variable de confusión. Más allá de la inspección visual de las tasas de aceptación de la tabla 3, las tablas 4 y 5 presentan dos tablas de contingencia construidas a partir de los mismos datos que nos permitirán determinar hasta qué punto el departamento cumple con las condiciones exigidas a una variable de confusión y, 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, predictora o explicativa.

Tabla 4. Datos de admisión en 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.

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.

Por un lado, colapsando el género de los estudiantes, 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 candi-

datos que se habrían presentado. Así, el departamento A sería aquel que más dificultades habría puesto a los estudiantes, de forma que habría resuelto favorablemente solo un tercio de sus 300 solicitudes. En comparación, el departamento B sería aquel al que habría resultado más fácil acceder, aceptando algo más de dos tercios de las 700 solicitudes que habría valorado. Por otro lado, colapsando ahora los resultados de las resoluciones de los dos 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 habría escogido, de forma que sus 100 candidatas solo suponen un tercio de las solicitudes que habría valorado. En cambio, el departamento B sería al que más mujeres se habrían presentado, valorando 450 candidatas que representan casi dos tercios de sus solicitudes. En este sentido, utilizando estas dos tablas de contingencia para analizar su asociación, podemos afirmar que existe una relación estadísticamente significativa tanto con la aceptación final de los candidatos (X^2=126,98, df=1, p=0,000) como con su género (X^2=81,29, df=1, p=0,000) que, además, resulta comparativamente de una intensidad o magnitud más importante (V de Crámer=0,36 y 0,29, respectivamente). En efecto, tal como sugería la inspección preliminar de los datos desagregados, el departamento actuaría como un factor o una variable de confusión en nuestra universidad ficticia.

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. Tal como hemos podido ver, su incorporación al análisis puede comportar el incremento, el decremento, la desaparición o, incluso, la inversión de su relación, de forma que la mera evidencia de una asociación entre dos variables no implica, necesariamente, la existencia de una relación 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 cuál 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 de forma exhaustiva las relaciones entre sus variables y los potenciales factores de confusión relevantes para 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. Entendiendo el análisis estadístico como la culminación de un complejo proceso de planificación a través del cual se lleva a cabo cualquier investigación cuantitativa, es posible distinguir dos grandes tipos de diseños: la investigación *experimental* y la investigación *no experimental*. En los dos casos, la investigación parte del desarrollo o la adopción de una teoría como marco general de referencia a partir de la 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 poner a prueba su asociación mediante las pruebas estadísticas oportunas, y la consideración de cualquier variable extraña que pueda actuar como factor de confusión interfiriendo en estas relaciones y, por lo tanto, 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, el lector interesado puede 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, el investigador tiene el control sobre los diferentes niveles o condiciones de al menos una variable independiente –generalmente llamada tratamiento– de modo que puede decidir de acuerdo con su voluntad la manera como serán expuestos los participantes de la investigación. Mediante una asignación aleatoria, el investigador selecciona a los individuos que forman parte de cada uno de los grupos experimentales y, una vez administrado el tratamiento, mide sus efectos en una o más variables dependientes. De este modo, cuando dispone de una muestra suficientemente amplia, el investigador iguala los diferentes grupos experimentales en relación con cualquier factor o variable de confusión, de forma que su influencia en la variable dependiente queda neutralizada gracias a la asignación aleatoria de los participantes. A pesar de que de acuerdo con esta lógica general un experimento puede adoptar formas mucho más complejas, su rasgo característico radica en la capacidad que da al investigador para atribuir, más allá de las pequeñas diferencias entre los grupos debidas al azar, las variaciones observadas en la variable dependiente como una consecuencia necesaria de la manipulación de su variable independiente o tratamiento.

Por otro lado, es posible caracterizar la investigación no experimental como la que se produce cuando el investigador no tiene el control sobre los diferentes niveles o condiciones de una o más variables independientes. Este tipo de investigación puede adoptar muchas formas, pero la más frecuente es el *cuestionario* o la *encuesta*. En este contexto, el investigador define sus variables independientes y, en tanto no tiene la capacidad de manipularlas de acuerdo con su voluntad, se limita a observarlas a partir de las respuestas proporcionadas por una muestra generalmente amplia de participantes en una o más ocasiones a lo largo del tiempo. Una vez administrado el cuestionario, el investigador identifica a los individuos que forman parte de los diferentes grupos previamente existentes y mide sus diferencias en una o más variables dependientes. De este modo, con una cierta confianza, atribuye estas diferencias a las variaciones en la variable independiente, pero, a diferencia de la investigación experimental, no será posible evitar la intervención de potenciales factores o variables de confusión, de forma que resulta difícil excluir la posibilidad de que su influencia se convierta en una explicación alternativa a la que propone.

Estos dos tipos de investigación difieren en su *validez interna*, es decir, en su 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 dos variables. 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, les puede permitir 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 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 un experimento o una encuesta, es en el momento del diseño y la construcción de la muestra, cuando los investigadores tienen que elegir a los participantes que, finalmente, formarán parte de sus estudios.

Dado que, por razones prácticas, no siempre es posible obtener información sobre el conjunto de la población que se propone analizar una investigación, es frecuente que los investigadores lleven 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*. En este sentido, 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 escogidos para participar en la investigación. Partiendo de una definición clara y precisa de la población objeto de estudio, en condiciones ideales los investigadores tendrían que ser capaces de identificar a todos sus miembros –por ejemplo, a partir de un listado con sus nombres– y, a continuación, procederían a escoger al azar a 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 intencional según su conveniencia o su disponibilidad.

A pesar de que una muestra aleatoria puede adoptar formas mucho más complejas, es conveniente señalar que solo cuando el criterio de selección de los participantes es aleatorio dispondremos de las garantías suficientes para considerar que las muestras sean representativas. De este modo, los investigadores tendrán la confianza de que las relaciones observadas a partir de la asociación entre sus variables son extrapolables al conjunto de la población a partir de la cual han sido extraídas las muestras. Es por esta razón por lo que tanto la investigación experimental como la no experimental no solo difieren en su validez interna, sino que también pueden hacerlo en su *validez externa*, es decir, en su capacidad para proporcionar las evidencias necesarias que permitan determinar que la existencia de una relación es generalizable a otras situaciones u otros individuos que no han formado parte del estudio. La tabla 6 presenta de manera esquemática 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 del azar al proceso de inferencia estadística.

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

	Asignación aleatoria	**Asignación no aleatoria**

Fuente: Elaboración propia.

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. Por un lado, el cuadrante superior izquierdo representa la investigación que, a través de su diseño, lleva a cabo una selección y una asignación aleatorias de sus participantes. Sería el caso de un experimento desarrollado a partir de una muestra representativa, donde la validez interna y externa de la investigación sería óptima y, por lo tanto, los investigadores se encontrarían en las mejores condiciones para establecer una relación causal que también fuera generalizable en la población. A su vez, en los cuadrantes superior derecho e inferior izquierdo encontramos las investigaciones que solo llevan a cabo una asignación o una selección no aleatorias y que, por lo tanto, tendrían una validez interna o externa más baja, respectivamente. En el primer caso, se trataría de una encuesta administrada a una muestra representativa, que permitiría establecer relaciones generalizables en el conjunto de la población pero que, en ningún caso, proporcionaría evidencias suficientes para determinar su naturaleza. 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. Finalmente, en el peor de los escenarios desde el punto de vista de la validez, el cuadrante inferior derecho representa la investigación que no lleva a cabo ni una selección ni una asignación aleatorias y que, por lo tanto, como sería el caso de una encuesta dirigida a una muestra no representativa, no permitiría establecer ni una relación causal ni generalizar sus resultados en el conjunto de la población.

Estos cuatro tipos de investigación difieren fundamentalmente en su validez y, tal como hemos podido ver, la razón por la cual esto es así no es ninguna otra más que el papel que asignan a la aleatorización en su diseño. En este sentido, la diferente capacidad que tienen los investigadores para determinar la existencia de una relación causal generalizable al conjunto de la población sirve como una buena ilustración de la contribución del azar a 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 de una muestra, es necesario tener presente que este proceso únicamente es posible si la selección o la asignación de los participantes 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 depen-

diente 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 poblacional*. De este modo, siempre que se cumpla esta condición, la estadística inferencial proporciona un conjunto de procedimientos que permite a los investigadores evaluar las diferencias observadas y decidir, con un determinado nivel de confianza, hasta qué punto responden a una diferencia realmente existente en la población o, en cambio, pueden ser sencillamente explicadas como resultado del azar en la selección y/o en la asignación de los participantes.

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

Pese a la importancia del diseño de la investigación para poder 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. 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 a las diferentes condiciones experimentales. Esta situación es bastante frecuente en las ciencias sociales, y es especialmente evidente cuando los estudios se desarrollan, lejos de las condiciones controladas de los laboratorios, en los contextos naturales donde se produce la actividad cotidiana de las personas. Si, tal como planteábamos al inicio de este módulo, el objetivo es analizar fenómenos complejos, como la discriminación por razón de género en el acceso de los estudiantes a la 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 sus 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 su 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 llevar a cabo juicios de causalidad a partir de la observación de asociaciones entre variables, cuando el análisis multivariante se presenta como el marco analítico general que permite modelar las múltiples relaciones existentes entre las diferentes variables involucradas en una 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 diversas variables de forma 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 objeto de interés para la investigación. Este marco analítico general, por lo tanto, ofrece a los investigadores la oportunidad de llevar a cabo un control estadístico de cualquier tercera variable, que, como eventual factor de confusión, pudiera interferir en la relación entre las variables dependientes e independientes. Es importante tener presente, no obstante, que la elección de las técnicas estadísticas –y el análisis multivariante no es una excepción– no tiene ninguna relación con el diseño que haya sido 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 no experimentales. Tal como ya hemos podido discutir ampliamente, 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 para 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, a su vez, este como una extensión del análisis univariante. En este sentido, el *análisis univariante* es la forma más simple de análisis estadístico y se propone la descripción de 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 el cálculo de una medida de tendencia central –como la media, la mediana o la moda– o 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 es que solo toma en consideración una única variable con el objetivo de describir la muestra y, cuando es posible, establecer una inferencia sobre la población que representa. Obviamente, cuando los investigadores llevan a cabo sus estudios nunca concentran todos sus 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 sus 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 parte de la universidad.

Por otro lado, el *análisis bivariante* es una extensión del análisis univariante y, 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 bien el cálculo de una correlación en el caso de variables cuantitativas, este tipo de análisis tiene como objetivo 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. Tal 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 debido, en último término, a 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. De este modo, 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 relaciones entre dos variables que no habrían 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 resultado 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 existentes entre diversas variables, pero lo hace, en este caso, mediante la construcción de modelos complejos que permiten determinar su existencia de forma simultánea. De este modo, 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, todas aquellas 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 su interés. Controlando estadísticamente la contribución de todas estas variables en el sistema de relaciones, este tercer tipo de análisis permite mantener constante 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 los 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. Para acabar con el caso que nos ha servido de hilo conductor en esta introducción, la estadística multivariante permitiría conocer la contribución simultánea de las características de los estudiantes y de los departamentos a 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 los hombres y las mujeres en sus capacidades, aptitudes o habilidades, controlando, por ejemplo, su expediente académico previo o sus resultados en las pruebas de acceso, de forma que sería posible extraer una conclusión todavía más exacta sobre la existencia de una discriminación por razón de género en el acceso de los estudiantes a la universidad.

Sería conveniente tener presente, sin embargo, 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 aquel que se utiliza 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 es-

tadística 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 su análisis. Por otro lado, entienden la estadística bivariante como el estudio de las relaciones entre parejas de variables que habrían sido obtenidas en investigaciones no experimentales, de forma 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 independientes consideradas, los investigadores amplían el número de variables dependientes en la construcción de sus modelos.

Esta aproximación alternativa plantea algunos inconvenientes que hacen poco interesante su adopción. Por un lado, 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 de los datos obtenidos, de forma 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. Por otro lado, 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 sus 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 donde los investigadores se proponen el objetivo de determinar la contribución simultánea de diversas variables independientes en una única variable dependiente.

En cualquier caso, la clave del análisis multivariante como marco analítico general no se encuentra en el hecho de que los investigadores dispongan de múltiples variables porque, como ya hemos dicho, los estudios no son nunca diseñados con el objetivo de observar una única variable. El rasgo distintivo de este tipo de análisis, en cambio, se encuentra en su capacidad de modelar las múltiples relaciones existentes entre las diferentes variables involucradas en una investigación de forma 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 hacerlo, en función de los objetivos de su 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 sus modelos y, de este modo, llevar a cabo una evaluación de su contribución 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 el otro, 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

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 sus características y las condiciones en que pueden ser utilizadas y, de manera particular, ofrecer una guía que permita al lector interesado 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 no experimentales, es importante tener presente que la elección de la técnica depende de dos aspectos estrechamente vinculados con el diseño de la investigación: la pregunta u objetivo general que motiva su desarrollo y las características de los datos que proporciona para ofrecer una respuesta. En este sentido, tal como hemos podido ver, el uso de las técnicas de análisis multivariante es 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 forma simultánea. Sin embargo, 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 su 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 y otros, 2013, para una discusión más amplia), la medida no sería otra cosa que el establecimiento de una correspondencia entre las propiedades de los fenómenos objeto de interés y los números que las representan en una determinada escala. De este

modo, 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 reflejan el grado o la cantidad con que la característica está presente sino que, en cambio, únicamente permiten distinguir de manera discreta a los individuos que cumplen las condiciones para pertenecer a un determinado nivel de entre todos los posibles. En este sentido, las variables cualitativas pueden ser definidas a partir del uso de escalas nominales y ordinales, cuando sus niveles sirven para identificar, respectivamente, a 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 logrado pero que, en ningún caso, reflejan de manera precisa 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 forma que es posible crear una serie de nuevas variables –llamadas *ficticias* o *dummies*– que identifican a los individuos que pertenecen a cada uno de los grupos en comparación al resto.

Por otro lado, las *variables cuantitativas* o *métricas* son aquellas en las 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 a 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 forma 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 con test o cuestionarios 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 por lo tanto muchos autores consideran que en realidad su escala tendría que ser considerada como ordinal, lo cierto es que la investigación desarrollada en las ciencias sociales trata a menudo estas variables como si realmente fueran de intervalo siempre y cuando su distribución sea aproximadamente normal. Finalmente, en el segundo caso, utilizan una escala de razón variable 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 su 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 importantes implicaciones para el proceso de medida. En este sentido, los investigadores tienen que escoger aquellas que mejor reflejen la variabilidad observada en los fenómenos objeto de interés y, por lo tanto, sean capaces de recoger de manera adecuada la información relativa a la presencia o ausencia de unas determinadas características o, cuando sus estudios lo requieren, el grado o la cantidad con que están presentes en los participantes. Pero –lo que es más relevante para una introducción como esta al análisis multivariante– 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 forma simultánea. Por un lado, los investigadores tienen que 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 sus diferentes niveles no son más que etiquetas numéricas arbitrarias que sirven para identificar los diferentes grupos de participantes que, en ningún caso, reflejan el grado o la cantidad con que una determinada característica está presente en los individuos. Si bien es cierto que en algunas ocasiones es posible tratar como cuantitativas algunas variables que, en principio, tendrían una escala ordinal, los investigadores tendrán que examinar su 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 la elección de la técnica de análisis multivariante más adecuada para cumplir con 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 en función de la pregunta u objetivo general que motiva la investigación. Tal como nos hemos propuesto, esta clasificación nos permitirá ofrecer una panorámica general sobre sus características y las condiciones en que pueden ser utilizadas de forma que, en último término, pueda servir de guía para orientar a los investigadores en el momento de escoger la técnica que mejor se ajuste a su investigación. A pesar de que la diversidad de técnicas disponibles nos impide abordar todas y cada una de ellas, esta clasificación servirá para presentar algunas de las más utilizadas y, además,

introducir aquellas que serán tratadas con más detalle en los módulos posteriores de este texto. Para hacer esto, 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 forma simultánea y, de este modo, distinguir la contribución independiente de cada una de ellas 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 de manera esquemática 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 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 predicciones o explicaciones	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 ajustamiento de modelos concatenados	Diversas variables cuantitativas	Ecuaciones estructurales

Fuente: Elaboración propia.

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

En primer lugar, cuando no es posible distinguir entre variables dependientes e independientes, 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 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 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 como determinan la estructura de componentes o factores. En el primer caso, los investigadores no disponen de una teoría sólida sobre las relaciones para construir sus modelos y, por lo tanto, se limitan a determinar de forma empírica la existencia de los componentes que, de hecho, agrupan sus variables. En cambio, en el segundo caso, los investigadores parten de una teoría sobre los fenómenos 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. Es importante tener presente que, a pesar de que existen algunos procedimientos para poder tratar variables cualitativas, estas dos técnicas son generalmente aplicadas cuando las variables analizadas son cuantitativas. En caso de que las variables utilizadas sean cualitativas, los investigadores tienen a su alcance una técnica alternativa, el *análisis de correspondencias*, para lograr los mismos objetivos. Mediante la transformación de la información cualitativa para poder tratarla de forma cuantitativa, esta técnica procede de una manera comparable y, por lo tanto, permite obtener un conjunto de dimensiones –similares a los componentes o 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 sirve 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*, también conocida como *análisis de clúster*. Esta técnica ofrece un conjunto de procedimientos para 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 en sus características. Seleccionadas las variables que formarán parte de los modelos, que pueden ser cuantitativas o cualitativas, y siempre en función del número de casos que los investigadores se proponen clasificar, el análisis de conglomerados se basa en el análisis y la interpretación de la asociación observada entre los individuos, de forma que el cálculo de su distancia o proximidad sirve para conformar grupos homogéneos en relación con las características seleccionadas que, a la vez, sean el máximo de 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 su investigación, la técnica más adecuada es el *escalamiento multidimensional*. En este caso, a diferencia de lo que sucede con las otras técnicas de análisis de las relaciones de interdependencia que hemos introducido antes, la busca 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 conformadas a partir de un conjunto de objetos, de acuerdo con sus preferencias o sus percepciones de similitud. Como en el caso del análisis de conglomerados, el escalamiento multidimensional puede ser aplicado tanto a variables cuantitativas como cualitativas.

En segundo lugar, cuando es posible distinguir entre variables dependientes e independientes, los investigadores se mueven 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 forma que servirían para evaluar la contribución específica del cambio o la variación en los niveles de todas y cada una de las variables independientes consideradas en la construcción de los modelos. Además, a pesar de que las variables independientes 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 cuantitativas que permiten explicar. Cuando los investigadores se proponen analizar la variabilidad de los individuos en una característica, y por lo tanto centran su atención en una única variable dependiente, 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 en 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 hacer 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 para 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 forma que permitirían evaluar la contribución específica de su pertenencia a diferentes grupos –llamados factores– conformados 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, tal 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 que los individuos han sido asignados de acuerdo con su voluntad. 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 y, como en el caso de la regresión múltiple y la correlación canónica, difieren en el hecho de que permiten explicar la variabilidad de los grupos en una o más variables dependientes cuantitativas, 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 no experimental al permitir tener en cuenta la influencia de otras características importantes de los individuos 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 su pertenencia a diferentes grupos. En este sentido, cuando los investigadores se proponen analizar las relaciones simultáneas entre diversas variables con la intención de clasificar a los individuos en los diferentes grupos conformados 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 que pertenecen, de forma que permitirían evaluar la contribución específica de todas las variables independientes consideradas en la construcción de los modelos. Tanto el análisis discriminado como la regresión

logística se basan también en el análisis y la interpretación de las asociaciones observadas entre las variables, pero difieren, fundamentalmente, en el número de niveles que la variable dependiente cualitativa puede adoptar y 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 conformados por una variable dependiente cualitativa de dos o más niveles y, además, lo hacen tomando en consideración un conjunto de variables independientes cuantitativas, su técnica de elección es el análisis discriminante. Es importante tener presente, sin embargo, que esta técnica impone una restricción en relación con las variables independientes, de forma que solo puede ser aplicada cuando todas y cada una de ellas sigan una distribución normal. En cambio, a pesar de que la regresión logística es aplicable únicamente cuando la variable dependiente es dicotómica, 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, permite considerar variables independientes cuantitativas o cualitativas.

En tercer lugar, para cerrar esta panorámica general sobre las diferentes técnicas de análisis multivariante, en algunas ocasiones los investigadores no se mueven exclusivamente en el contexto de la interdependencia o de la dependencia, sino que lo hacen en la combinación de estos dos tipos de relaciones. En este sentido, cuando están interesados en analizar relaciones entre sus variables que pueden ser de dependencia y de interdependencia simultáneamente, la técnica más adecuada es la de *ecuaciones estructurales*. A diferencia de todas las anteriores, esta técnica tiene como objetivo general analizar simultáneamente las múltiples relaciones existentes entre diferentes grupos de variables, de forma que permitiría evaluar el ajustamiento de varios modelos multivariantes concatenados. Para hacer esto, las ecuaciones estructurales se basan en el análisis y la interpretación de las asociaciones observadas entre varias variables que, en términos generales, pueden ser organizadas en dos grandes tipos de modelos. Por un lado, de acuerdo con la lógica 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 el otro, de acuerdo con la lógica de las relaciones de interdependencia, 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 forma simultánea. En este sentido, 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, existen algunos procedimientos que permiten tratar también variables cualitativas. De este modo, las ecuaciones estructurales se presentan como la técnica de análisis más eficiente de que disponen los investigadores

interesados al 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

Pese a la diversidad de técnicas disponibles en función de la pregunta u objetivo general que motiva la investigación y las características de los datos que proporciona para ofrecer una respuesta, acabaremos esta introducción a los aspectos básicos del análisis multivariante abordando el proceso de construcción de los modelos estadísticos complejos, que permiten analizar e interpretar las múltiples relaciones existentes entre las diferentes variables involucradas en una investigación de forma simultánea. Antes de hacerlo, tal como hemos mostrado a lo largo de este módulo, es necesario recordar que el análisis multivariante únicamente adquiere todo su sentido en relación con el procedimiento establecido en la investigación cuantitativa que, en términos generales, podríamos resumir de la siguiente manera:

1) Formular una pregunta o un objetivo general que sirva para abordar un problema relevante.

2) Escoger el diseño de la investigación y especificar la muestra de participantes.

3) Definir adecuadamente todas las variables involucradas y especificar sus características.

4) Desarrollar o escoger los instrumentos necesarios para llevar a cabo las medidas.

5) Recoger las evidencias necesarias que permitan responder a los objetivos de la investigación.

6) Resumir y tratar estadísticamente los datos para evaluar las evidencias obtenidas y, cuando es posible, generalizar las conclusiones más allá de los límites de los estudios particulares.

De acuerdo con este procedimiento, el análisis multivariante proporciona el marco analítico general que permite a los investigadores describir, explicar o predecir los fenómenos objeto de interés mediante el desarrollo de los modelos estadísticos más adecuados para 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 a través del cual los investigadores obtienen una combinación lineal de variables que les permite estimar, a partir de los datos obtenidos de los participantes, el peso específico o la importancia relativa de todas y cada una de ellas y, de este modo, 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 las múltiples relaciones consideradas de forma simultánea en sus modelos y, siguiendo los procedimientos establecidos en función de la técnica aplicada, para decidir hasta qué punto estos modelos se ajustan o son una buena representación de la realidad. 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 diferentes técnicas disponibles siguen esta lógica, de forma 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, entre los principios generales que rigen el análisis multivariante, podemos señalar algunos de los más importantes:

- **La construcción de modelos multivariantes requiere de una fundamentación teórica de las relaciones**. Dada 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 su utilización 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 marco general de referencia a partir del cual sea razonable esperar que se puedan producir las relaciones 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 proporcionará la investigación y, en último término, les permita escoger la técnica más adecuada que será conveniente utilizar 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 se encuentra en la computación estadística de los modelos multivariantes, sino precisamente en todas las decisiones que los investigadores tienen que tomar para poder llegar a construirlos con éxito.

- **La exploración de los datos es una condición previa para el desarrollo del análisis multivariante**. Como extensión del análisis univariante y bivariante, los investigadores no tendrán que perder de vista la importante contribución que estos dos tipos de análisis hacen en el momento de tomar contacto con los datos obtenidos en la investigación. 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 forma 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 sus objetivos de investigación. Esta exploración de los datos es especialmente relevante en el contexto del

análisis de las relaciones de dependencia, en las que, como hemos podido ver, permite obtener los indicios necesarios para considerar el papel de cualquier factor o variable de confusión que pueda estar interfiriendo en las relaciones objeto de interés y, por lo tanto, controlar estadísticamente su influencia en la construcción de los modelos multivariantes para evitar así que se conviertan en una explicación alternativa.

- **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 a los investigadores comprobar hasta qué punto se cumplen los supuestos sobre los que se basa la técnica escogida para modelar las múltiples relaciones entre variables de manera simultánea. En este sentido, tal como hemos expuesto en nuestra clasificación de las diferentes técnicas disponibles, es importante que corroboren que tanto la naturaleza de las relaciones que se proponen analizar como las características de variables implicadas se ajustan a los requerimientos de la técnica seleccionada. Además, más allá de los requerimientos estadísticos particulares que tiene cada técnica, es importante tener presente que la inferencia estadística asume también algunos supuestos importantes como en relación a la distribución aproximadamente normal de las variables, la linearidad 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 garantizan 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 tendrán que decidir cuál de entre todas las variables de que disponen serán utilizadas en la especificación de sus modelos. En este sentido, es conveniente tener presente que tan importante resulta seleccionar todas aquellas variables que sean pertinentes desde el punto de vista teórico, y por lo tanto no dejar de tener en cuenta ninguna importante, como evitar incluir cualquiera otra variable que, en realidad, no sea relevante para analizar los fenómenos objeto de interés. Es lo que llamamos una especificación adecuada de los modelos que, en todo caso, no 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 van añadiendo y sacando variables en función de los resultados que les proporcionan. El objetivo de este proceso es, además, la obtención de unos modelos que sean parsimoniosos, es de-

cir, capaces de representar el máximo de la complejidad de los fenómenos 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 correcta especificación de los modelos multivariantes, y por lo tanto, no pueden ser nunca utilizados como una 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, por lo tanto, condicionan de forma necesaria el papel que juegan 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 que los investigadores evalúen hasta qué punto la combinación de variables que proponen se ajusta razonablemente bien a la variabilidad observada en los datos y, 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 poder interpretar las múltiples relaciones objeto de interés para la investigación.

- **El diseño de la investigación condiciona la inferencia estadística basada en el análisis multivariante**. Tal como hemos discutido ampliamente, 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 llevar a cabo la investigación. Como sucede con cualquier otra técnica estadística, tanto la inferencia causal como la inferencia en la población que permite el análisis multivariante solo es posible si la selección o la asignación de los participantes a los diferentes grupos ha sido aleatoria. 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 una cierta confianza, la existencia de una relación causal generalizable en la población que representa la muestra o si, en cambio, podrían ser explicadas simplemente como consecuencia del azar. Pese a 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 de que disponen para establecer sus inferencias.

Finalmente, para cerrar esta introducción, 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 procedi-

miento establecido en la investigación cuantitativa a partir del cual el análisis multivariante adquiere su sentido y, de manera particular, 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 discutiendo a lo largo de este módulo y, además, permiten poner en práctica todos los conocimientos, las habilidades y los valores vinculados con la construcción de modelos multivariantes. Dado su carácter general y, por lo tanto, con independencia de las especificidades de los procedimientos con que tienen que ser aplicadas las diferentes técnicas disponibles, estas diez fases fundamentales sirven para organizar de forma secuencial las diferentes decisiones que los investigadores tienen que tomar para llevar a cabo un análisis complejo de sus datos. De este modo:

1) Delimitación del propósito del análisis. La construcción de modelos multivariantes empieza siempre con una definición precisa de los objetivos particulares para los que los investigadores se proponen analizar e interpretar las múltiples relaciones entre diversas variables de forma simultánea. Tal como hemos podido ver, las diferentes técnicas de análisis multivariante disponibles pueden ser utilizadas con multitud de finalidades, que, de manera general, permiten a los investigadores describir, explicar o predecir los fenómenos objeto de interés para su investigación. 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. 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 proporcionados por su investigación. Esta es una decisión importante en el proceso de construcción de modelos multivariantes, puesto que la técnica finalmente escogida condiciona los procedimientos que los investigadores tienen que llevar a cabo durante las siguientes fases.

3) Exploración inicial de los datos. Una vez seleccionada la técnica de análisis multivariante, los investigadores tienen que familiarizarse con la distribución de las variables involucradas y, a continuación, examinar sus relaciones por parejas. Mediante la aplicación de técnicas de análisis univariante y bivariante, este primer contacto con los datos obtenidos en la investigación permite a los investigadores determinar la conveniencia de llevar a cabo un análisis multivariante para responder a sus objetivos particulares. Tal 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 tiene que 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 las 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 las variables implicadas 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 tendrá que ser evaluado y, si procede, revisado a lo largo de las siguientes fases.

6) Evaluación del ajustamiento 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 tendrán que llevar a cabo sus 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 ajustamiento de los modelos conduce a la séptima fase, en que los investigadores valoran la conveniencia de añadir o sacar 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 muy especificados que además sean parsimoniosos, de forma que los investigadores tienen que ser capaces de hacer un

balance adecuado entre el incremento del nivel de complejidad de sus modelos y los beneficios que esto tendría que comportar en relación con la mejora sustantiva en su ajustamiento a los datos.

8) Interpretación del sistema de relaciones. Una vez conseguido un ajustamiento global aceptable, llega el momento en que los investigadores pueden utilizar sus modelos multivariantes para analizar e interpretar la naturaleza de 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 su contribución independiente al sistema de relaciones. Cuando el objetivo del análisis es establecer inferencias causales o en la población, esta interpretación permite determinar la significación estadística de las asociaciones observadas en la muestra y, lo que 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 tendría que contemplar 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 tendrían que disponer de una muestra de participantes diferente de la que han utilizado para modelar sus 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 de utilidad 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 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 tiene que ir siempre acompañado de un esfuerzo especial por parte de los investigadores para transmitir y hacer accesibles los resultados de sus estudios. En este sentido, es especialmente relevante el uso de un lenguaje sencillo pero cuidadoso que permita reflejar de forma adecuada 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 en la población que representa la muestra.

Bibliografía

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

Bickel, P. J.; Hammel, E. A.; O'Connell, J. W. (1975). «Sex bias in graduate admissions: Data from Berkeley». *Science* (núm. 187, págs. 398-404).

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

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

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

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

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

Meneses, J.;Barrios, M.; Bonillo, A.; Cosculluela, A.; Lozano, L. M.; Turbany, J. I.; Valero, S. (2013). *Psicometría*. Barcelona: Editorial UOC.

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

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

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

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

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