Aspectes bàsics de l'anàlisi multivariant

Julio Meneses

PID_00199845

UOC
Universitat Oberta de Catalunya

Índex

Introducció

Aquesta introducció als aspectes bàsics de l'anàlisi multivariant té com a objectiu general proporcionar al lector algunes claus importants per a abordar els coneixements, les habilitats i els valors vinculats amb el desenvolupament de models estadístics que permeten portar a terme una anàlisi complexa de les dades obtingudes en la investigació quantitativa. Prenent com a punt de partida un estudi clàssic sobre la discriminació per raó de gènere a la Universitat de Berkeley, exposarem aquest cas controvertit com un exemple d'investigació en què l'omissió d'una informació rellevant per a l'anàlisi pot conduir a una conclusió inadequada. Aquesta discussió servirà per a introduir alguns conceptes importants com són l'associació, la confusió i la causalitat, així com per a reconèixer la importància del disseny de la investigació per a poder extreure conclusions no esbiaixades, que, a més a més, siguin generalitzables més enllà dels límits dels estudis particulars. A continuació, presentarem l'anàlisi multivariant com el marc analític general que permet modelar les múltiples relacions existents entre diverses variables de manera simultània, en descriurem els objectius principals i presentarem una classificació general de les diferents tècniques disponibles, que ens permetrà oferir una guia per a orientar els investigadors en el moment d'escollir la que millor s'ajusti a la seva investigació. Aquesta exposició servirà per a situar aquest tipus d'anàlisi en el context general de la investigació quantitativa i, finalment, presentar alguns dels principis que regeixen les diferents fases amb què és possible estructurar el procés de construcció de models multivariants. D'aquesta manera, tractarem de posar les bases necessàries a partir de les quals es desenvoluparà l'exposició de les diferents tècniques d'anàlisi multivariant que abordarem en detall a la resta de mòduls que segueixen aquesta introducció general.

1. El cas de la discriminació de gènere a la Universitat de Berkeley

L'any 1973 va ser un any interessant per a la discussió sobre la situació de les dones en el món universitari als Estats Units. Resoltes les sol·licituds d'accés per al començament del curs aquella tardor, la Universitat de Berkeley va portar a terme una investigació interna per a determinar si hi havia indicis fundats sobre l'existència d'una discriminació per raó de gènere en l'accés dels seus estudiants als programes de postgrau. En aquest sentit, examinant les dades recollides als arxius dels diferents departaments, el professor Hammel, llavors degà d'aquests estudis, es va trobar amb una situació, si més no, aparentment paradoxal (Bickel, Hammel i O'Connell, 1975).

Tenint en compte el conjunt global de sol·licituds, aquell curs es van presentar un total de 12.763 candidats, dels quals 8.442 van ser homes i 4.321 dones. D'aquests candidats, aproximadament un 44% dels homes i un 35% de les dones van ser finalment admesos per a iniciar els seus estudis de postgrau. La taula 1 recull aquestes dades, desagregant les candidatures admeses i rebutjades en funció del gènere dels sol·licitants, i permet il·lustrar les conclusions preliminars d'aquesta investigació. En efecte, tenint en compte que la taxa global d'acceptació en el conjunt dels departaments va ser d'un 41% aproximadament, la diferència de gairebé 10 punts entre els homes i les dones seria una evidència incontestable a favor de l'existència d'una discriminació per raó de gènere. De fet, si utilitzem aquesta taula de contingència per a analitzar la seva associació, podem afirmar que existeix una relació estadísticament significativa entre el gènere dels candidats i la seva acceptació final als programes de postgrau de la Universitat de Berkeley (X^2 = 111,25, df = 1, p = 0,000). Tot i ser estadísticament significativa, però, aquesta relació no mostra una intensitat o una magnitud important (V de Cramér = 0,09).

Taula 1. Resolució sobre les sol·licituds d'accés als programes de postgrau de la Universitat de Berkeley segons el gènere dels candidats (tardor de 1973)

	Sol·licituds	Admissions	Rebutjos	Percentatge d'admissió
Homes	8.442	3.738	4.704	44,28%
Dones	4.321	1.494	2.827	34,58%
Total	12.763	5.232	7.531	40,99%

Font: Bickel, Hammel i O'Connell (1975).

Si assumim, i no tenim evidències per a no fer-ho així, que les dones i els homes no difereixen significativament en les seves capacitats, aptituds i habilitats, la Universitat de Berkeley estaria preferint els homes per davant de les dones com a estudiants dels seus programes. Aquesta situació, però, resulta

més complexa de la representació que ofereix l'anàlisi d'aquesta taula de contingència. Tal com mostren Bickel, Hammel i O'Connell (1975), l'aparent discriminació per raó de gènere es produeix únicament quan agreguem les dades per al conjunt de la universitat. Tot i que en el seu treball no reprodueixen les dades proporcionades per cadascun dels cent un departaments que oferien aquests estudis, la seva anàlisi serveix com una interessant il·lustració d'una relació espúria entre el gènere dels candidats i la seva acceptació final. Descartant els registres dels departaments que no van rebre cap sol·licitud per part de cap dona o que finalment no van rebutjar cap candidat, van identificar quatre dels vuitanta-cinc departaments restants que efectivament mostraven una preferència estadísticament significativa pels homes. En canvi, sis d'aquests mateixos departaments van resoldre les seves sol·licituds en el sentit contrari, mostrant una preferència estadísticament significativa per les dones. És més, examinant les taules de contingència d'aquests deu departaments que mostraven una preferència, pels homes o per les dones, la seva conclusió va ser que la discriminació per raó de gènere en l'accés als estudis de postgrau afectava, en realitat, més els homes que les dones.

Però, donada una relació estadísticament significativa entre el gènere dels candidats i la seva acceptació en el conjunt de la universitat a favor dels homes, com és possible que una gran majoria dels departaments de Berkeley no mostrés cap preferència i que, tenint en compte la minoria que ho feien pels homes o per les dones, aquesta discriminació per raó de gènere afectés més els homes que no pas les dones? Freedman, Pisani i Purves (2007) ofereixen una aproximació complementària que ens pot ajudar a entendre aquesta contradicció. Prenent en consideració les dades proporcionades pels sis departaments més grans, que havien avaluat aproximadament un terç dels candidats de tota la universitat, van registrar el nombre de sol·licituds i van calcular-ne les respectives taxes d'admissió. La taula 2 recull aquestes dades, desagregant les sol·licituds en funció del gènere dels candidats. Tal com es pot observar, els percentatges d'admissió són bastant similars a aquests sis departaments. L'excepció més notable és el departament A, que va mostrar una preferència important per les dones i en va acceptar un 82% en comparació amb el 62% dels homes. En el sentit contrari, el departament E va mostrar una preferència més clara pels homes i en va acceptar un 28% en comparació amb el 24% de les dones. En canvi, si ens fixem en les sol·licituds als sis departaments en el seu conjunt, la relació entre el gènere dels candidats i la seva acceptació als programes de postgrau torna a ser evident a favor dels homes, amb una taxa global del 44% en comparació amb la del 30% en el cas de les dones.

Taula 2. Dades d'admissió als sis departaments més grans de la Universitat de Berkeley segons el gènere dels candidats (tardor de 1973)

	Homes		**Dones**	
Departament	**Sol·licituds**	**Percentatge d'admissió**	**Sol·licituds**	**Percentatge d'admissió**
A	825	62%	108	82%

Font: Freedman, Pisani i Purves (2007).

	Homes		Dones	
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%

Font: Freedman, Pisani i Purves (2007).

Una diferència de 14 punts entre els homes i les dones en la taxa global d'acceptació dels sis departaments més grans tornaria a ser una evidència incontestable a favor de l'existència d'una discriminació per raó de gènere a la Universitat de Berkeley. Però si observem amb deteniment les dades desagregades per a cada departament que recull la taula 2 serem capaços de trobar una explicació intuïtiva a aquesta contradicció. Tenint en compte les seves respectives taxes d'acceptació, els departaments A i B serien els que més sol·licituds van acceptar finalment i, per tant, aquells a què va resultar més fàcil accedir per als candidats, fossin homes o dones, que s'hi van presentar. Amb uns percentatges que varien entre el 82% i el 62%, això suposa que almenys dues terceres parts van acabar accedint als programes que oferien aquests dos primers departaments. En canvi, els departaments C, D, E i F serien els que més dificultats van posar als candidats, fossin homes o dones, perquè finalment van resoldre favorablement un nombre sensiblement més baix de les sol·licituds que van rebre. Amb uns percentatges que oscil·len entre el 37% i el 6%, almenys dues terceres parts dels candidats no van acabar accedint als seus programes.

Els departaments, per tant, no van mostrar un comportament similar en relació amb l'acceptació dels seus estudiants. Però, el que és més important per a entendre l'aparent contradicció d'aquest cas de discriminació per raó de gènere: els estudiants tampoc no van mostrar un comportament similar en relació amb l'elecció del departament per a presentar les seves candidatures. Tenint en compte el nombre de sol·licituds que van rebre, els departaments A i B van valorar un total de 1.385 homes, és a dir, una mica més de la meitat dels 2.691 que es van presentar com a candidats en el conjunt dels sis departaments. En canvi, els departaments C, D, E i F van valorar 1.702 dones, que representa gairebé la pràctica totalitat de les 1.835 que s'hi van presentar. D'aquesta manera, els homes van sol·licitar l'accés als departaments més fàcils o, almenys, a aquells que més candidats van acceptar, mentre que les dones ho van fer, contràriament, als més difícils o que menys candidats van acceptar. Per aquesta raó, tot i que de manera agregada podria semblar el contrari, quan controlem les diferències entre els homes i les dones en la seva elecció del departament, com fem a la taula 2, la relació entre el gènere dels candidats i la seva acceptació final als programes de postgrau a favor dels homes pràcticament desapareix.

2. Associació, confusió i causalitat

El cas de la discriminació de gènere a la Universitat de Berkeley (Bickel, Hammel i O'Connell, 1975) s'ha convertit en un exemple clàssic d'un fenomen que es produeix sovint a l'anàlisi estadística quan l'estudi de les relacions entre dues variables omet o no té en compte adequadament alguna informació rellevant per a l'estudi. És la *paradoxa de Simpson*, expressió encunyada per Blyth (1972) a partir de l'exposició de Simpson (1951) per fer referència a un fenomen que, en realitat, va ser descrit originalment uns quants anys abans per Yule (1903) com a extensió a les taules de contingència de la discussió de Pearson sobre l'existència de correlacions espúries entre variables quantitatives (Aldrich, 1995; David i Edwards, 2001). En aquest sentit, podem definir aquest fenomen, aparentment paradoxal, com el fet que una associació observada entre dues variables qualitatives canvia el seu sentit si, en lloc de fer-ho de manera agregada, s'analitza la seva relació a cadascun dels subgrups que es conformen a partir d'una tercera variable qualitativa.

La paradoxa de Simpson no és un fenomen infreqüent a les ciències socials, particularment als estudis observacionals, i resulta especialment sorprenent als ulls del públic no especialitzat que no espera trobar-se aquest tipus de contradiccions. Una universitat no pot discriminar les dones en la resolució de les seves sol·licituds d'accés en el conjunt dels estudis que ofereix i, a la vegada, no fer-ho o, fins i tot, discriminar lleugerament els homes a cadascun dels departaments que la componen. En cap cas, però, és adequat interpretar aquesta aparent contradicció com el resultat d'un artefacte estadístic o com un indici que la investigació hagi estat incorrectament dissenyada o desenvolupada. Les relacions estadísticament significatives existeixen, són reals, tant en el cas del conjunt dels candidats valorats per la universitat com en el detall dels seus departaments. En aquest sentit, el que posa de manifest la contradicció no és l'existència d'aquestes relacions, sinó el fet que les evidències observades d'associació entre les variables siguin utilitzades com a prova per a portar a terme judicis causals. Com que, en l'anàlisi agregada, s'estaria ometent o no tenint en compte adequadament una informació rellevant per a l'estudi, la relació observada entre les variables resultaria una estimació esbiaixada i, per tant, una evidència inadequada per a la inferència causal que persegueix. Només quan es prenen en consideració els resultats de l'anàlisi desagregada, no esbiaixada en el cas que ens ocupa, és possible entendre adequadament el fenomen objecte d'estudi als diferents subgrups, i, d'aquesta manera, l'aparent contradicció es dilueix.

En aquest sentit, podem considerar la paradoxa de Simpson com un cas particular, de fet el més extrem, de confusió. Un *factor* o *variable de confusió* és una variable estranya, no prevista en la investigació, que pot alterar la relació entre dues variables objecte d'interès i que, per tant, pot afectar els judicis de

causalitat que fan els investigadors a partir de l'observació de la seva associació. Si, en el context d'una investigació que tingui com a objectiu posar a prova una relació de causalitat, observem una associació entre una *variable independent* –també anomenada predictora o explicativa– i una *variable dependent* –també coneguda com a resultat o explicada–, una tercera variable seria un factor de confusió si la seva incorporació a l'anàlisi comportés l'increment, el decrement, la desaparició o, fins i tot, com hem pogut veure, la inversió de la seva relació. Per a fer-ho, el potencial factor de confusió hauria de complir necessàriament la condició d'estar associat tant amb la variable dependent com amb la independent, de manera que el seu efecte o contribució específica en relació amb la variable dependent resultaria indistingible del que tindria la variable independent. És precisament per aquesta raó que, com tots els investigadors haurien de tenir sempre present en la seva pràctica, tot i que la determinació d'una relació de causalitat implica l'observació d'una associació entre dues variables, la mera evidència d'aquesta associació des del punt de vista estadístic no implica, necessàriament, l'existència d'una relació causal. Més enllà d'aquestes nocions bàsiques, el lector interessat pot trobar una introducció general a l'estudi de les relacions de causalitat en la investigació social a Russo (2009) i una discussió més ampla sobre l'establiment d'aquest tipus d'inferències al treball pioner de Pearl (2000).

L'estudi sobre la discriminació per raó de gènere en l'accés dels estudiants als programes de postgrau de la Universitat de Berkeley és, per tant, un bon exemple d'investigació en què l'omissió d'una variable de confusió en l'anàlisi agregada per al conjunt dels departaments condueix a una conclusió esbiaixada. Tal com hem pogut veure, una senzilla inspecció visual de la taula 2, que recull la distribució dels sis departaments més grans en funció del nombre de sol·licituds presentades pels candidats i de les seves taxes d'acceptació final, ens ha permès esbossar una explicació intuïtiva sobre el seu paper com a potencial factor de confusió. Tenint en compte que ni els departaments ni els estudiants es van comportar de manera similar, el canvi de sentit en la relació entre el gènere dels candidats i la seva acceptació seria conseqüència de la preferència dels homes i les dones pels més fàcils i més difícils d'accedir-hi, respectivament. En no disposar de les dades originals desagregades per a la totalitat dels departaments, no és possible anar més enllà d'aquesta explicació intuïtiva i mostrar, mitjançant les proves estadístiques oportunes, com el departament compleix la condició d'estar associat tant amb el gènere dels candidats com amb la seva acceptació final. En canvi, podem il·lustrar aquest requeriment amb un exemple fictici que, a més a més, ens permetrà posar de manifest com la incorporació d'un factor de confusió a l'anàlisi no només pot alterar la relació observada entre dues variables, sinó que també pot fer evident una relació que ni tan sols havia estat observada inicialment.

Imaginem una universitat fictícia formada, per tal de simplificar l'anàlisi, únicament per dos departaments. Tenint en compte el conjunt global de sol·licituds, suposem que s'hi van presentar un total de 1.000 candidats, dels quals 450 serien homes i 450 dones, i que d'aquests candidats finalment un

60% tant dels homes com de les dones haurien estat acceptats per a iniciar els seus estudis. La taula 3 recull aquestes dades, desagregant les candidatures admeses i rebutjades en funció del departament escollit i del gènere dels sol·licitants. En aquest cas, tenint en compte que la taxa global d'acceptació en el conjunt dels departaments hauria estat del 60% tant per als homes com per a les dones, el fet que no hi hagi cap diferència seria una evidència incontestable en contra de l'existència d'una discriminació per raó de gènere. Si utilitzem les dades totals que es presenten a la darrera filera per a construir una taula de contingència, l'anàlisi de la seva associació ens permet afirmar que, almenys de manera agregada, no existeix cap relació entre el gènere dels candidats i la seva acceptació en aquesta universitat fictícia (X^2 = 0, df = 1, p = 1). Com és natural, tractant-se de dues variables totalment independents entre si, la intensitat o magnitud de la seva relació és nul·la (V de Cramér = 0).

Taula 3. Resolució sobre les sol·licituds d'accés a una universitat fictícia segons el departament escollit i el gènere dels candidats

		Sol·licituds	Admissions	Rebutjos	Percentatge d'admissió
Departament A	**Homes**	200	80	120	40,00%
	Dones	100	20	80	20,00%
Departament B	**Homes**	250	190	60	76,00%
	Dones	450	310	140	68,89%
Total	**Homes**	450	270	180	60,00%
	Dones	550	330	220	60,00%

Font: Elaboració pròpia.

La nostra universitat fictícia no mostraria cap preferència, ni pels homes ni per les dones, en la resolució de les sol·licituds d'accés dels estudiants als seus programes. Però, si en lloc de fer una anàlisi agregada ens fixem en les dades que corresponen a cadascun dels dos departaments, la situació que ens trobem resulta molt diferent. Tenint en compte les seves respectives sol·licituds, al departament A s'hi haurien presentat 200 homes i 100 dones, dels quals haurien estat finalment acceptats un 40% i un 20%, respectivament. En un sentit similar, al departament B s'hi haurien presentat 250 homes i 450 dones, dels quals haurien estat acceptats un 76% i aproximadament un 69%, respectivament. Una diferència entre els homes i les dones de 20 punts al departament A i de 17 punts al departament B seria una evidència incontestable a favor de l'existència d'una discriminació per raó de gènere. Tots dos departaments d'aquesta universitat estarien, en realitat, preferint els homes per davant de les dones com a estudiants dels seus programes. De fet, si utilitzem les dades que es presenten a la primera i a la segona filera per a construir dues taules de contingència separades, l'anàlisi de la seva associació ens permetria afirmar que existeix una relació estadísticament significativa entre el gènere dels candidats i la seva acceptació a favor dels homes, tant al departament A (X^2 = 12, df = 1,

$p < 0,001$) com al departament B ($X^2 = 3,98$, df = 1, $p < 0,05$). La intensitat o magnitud d'aquesta relació és, però, més important al primer departament (V de Cramér = 0,2) que al segon (V de Cramér = 0,08).

En aquest sentit, l'anàlisi de les dades desagregades per a cadascun dels dos departaments de la nostra universitat fictícia suggereix l'existència d'un factor o una variable de confusió. Més enllà de la inspecció visual de les taxes d'acceptació de la taula 3, les taules 4 i 5 presenten dues taules de contingència construïdes a partir de les mateixes dades que ens permetran determinar fins a quin punt el departament compleix les condició exigides a una variable de confusió i, per tant, està efectivament relacionat tant amb l'acceptació dels candidats –és a dir, la variable dependent, resultat o explicada– com amb el seu gènere –la variable independent, predictiva o explicativa–.

Taula 4. Dades d'admissió a una universitat fictícia segons el departament escollit pels candidats

Departament	Sol·licituds	Admissions	Rebutjos	Percentatge d'admissió
A	300	100	200	33,33%
B	700	500	200	71,43%
Total	1.000	600	400	60,00%

Font: elaboració pròpia.

Taula 5. Sol·licituds d'accés a una universitat fictícia segons el gènere dels candidats

Departament	Sol·licituds	Homes	Dones	Percentatge de dones
A	300	200	100	33,33%
B	700	250	450	64,29%
Total	1.000	450	550	55,00%

Font: elaboració pròpia.

D'una banda, agrupant el gènere dels estudiants, la taula 4 presenta les dades d'admissió segons el departament escollit i mostra una important diferència en el seu comportament en relació amb l'acceptació dels candidats que s'haurien presentat. Així, el departament A seria el que més dificultats hauria posat als estudiants, de manera que hauria resolt favorablement només un terç de les seves 300 sol·licituds. En comparació, el departament B seria aquell a què hauria resultat més fàcil accedir, i hauria acceptat una mica més de dos terços de les 700 sol·licituds que hauria valorat. D'altra banda, agrupant ara els resultats de les resolucions dels dos departaments, la taula 5 presenta les sol·licituds d'accés segons el gènere dels candidats i mostra també una important diferència en el seu comportament en relació amb l'elecció del departament per a presentar les seves candidatures. Així, el departament A seria aquell que menys dones haurien escollit, de manera que les seves 100 candidates només suposen un terç de les sol·licituds que hauria valorat. En canvi, el departament B seria aquell en què més dones s'haurien presentat, valorant 450 candidates

que representen gairebé dos terços de les seves sol·licituds. En aquest sentit, utilitzant aquestes dues taules de contingència per analitzar la seva associació, podem afirmar que existeix una relació estadísticament significativa tant amb l'acceptació final dels candidats (X^2 = 126,98, df = 1, p = 0,000) com amb el seu gènere (X^2 = 81,29, df = 1, p = 0,000) que, a més a més, resulta comparativament d'una intensitat o magnitud més important (V de Cramér = 0,36 i 0,29, respectivament). En efecte, tal com suggeria la inspecció preliminar de les dades desagregades, el departament actuaria com un factor o una variable de confusió en la nostra universitat fictícia.

3. Disseny de la investigació i inferència estadística

La lliçó que podem extreure del cas de la discriminació de gènere de la Universitat de Berkeley, com a exemple clàssic de la paradoxa de Simpson, és que l'existència de potencials factors de confusió no considerats en l'anàlisi és una de las amenaces més importants per als investigadors que es plantegen fer judicis de causalitat a partir de l'observació d'associacions entre les seves variables. Tal com hem pogut veure, la seva incorporació a l'anàlisi pot comportar l'increment, el decrement, la desaparició o, fins i tot, la inversió de la seva relació, de manera que la mera evidència d'una associació entre dues variables no implica, necessàriament, l'existència d'una relació causal. De fet, la incorporació d'un factor de confusió a l'anàlisi no només pot alterar la relació observada entre dues variables, sinó que també pot fer evident una relació que, com en el cas de la nostra universitat fictícia, ni tan sols havia estat inicialment observada. Per aquesta raó, sigui quin sigui el tipus d'investigació, és obligació dels investigadors considerar l'eventual influència de qualsevol tipus de variable estranya que pogués interferir i, per tant, examinar exhaustivament les relacions entre les seves variables i els potencials factors de confusió rellevants per als seus estudis.

En aquest sentit, és important tenir present que la capacitat dels investigadors per a establir inferències causals a partir de l'anàlisi de les seves dades està molt relacionada amb la naturalesa del disseny de la investigació. Entenent l'anàlisi estadística com la culminació d'un complex procés de planificació a través del qual es porta a terme qualsevol investigació quantitativa, és possible distingir dos grans tipus de dissenys: la *investigació experimental* i la *investigació no experimental*. En tots dos casos, la investigació parteix del desenvolupament o l'adopció d'una teoria com a marc general de referència a partir de la qual sigui raonable establir una relació causal entre les variables, el plantejament d'algunes hipòtesis sobre les relacions entre les variables dependents i independents per tal de posar a prova la seva associació mitjançant les proves estadístiques oportunes, i la consideració de qualsevol variable estranya que pugui actuar com a factor de confusió interferint en aquestes relacions i, per tant, convertir-se en una explicació alternativa. La diferència substancial, com veurem a continuació, es troba en la capacitat dels investigadors per a manipular les variables independents de manera que sigui possible atribuir adequadament les diferències observades en les variables dependents a les variacions de les variables independents. Més enllà de la breu exposició que farem a continuació, el lector interessat pot trobar una discussió més profunda sobre el disseny de la investigació als treballs de Shadish, Cook i Campbell (2002), Coolican, (2014) o Cozby i Bates (2015).

D'una manera senzilla, podem caracteritzar la investigació experimental descrivint la forma més simple que pot adoptar un *experiment*. En aquest context, l'investigador té el control sobre els diferents nivells o condicions d'almenys una variable independent –generalment anomenada *tractament*–, de manera que pot decidir d'acord amb la seva voluntat la manera com seran exposats els participants de la investigació. Mitjançant una assignació aleatòria, l'investigador selecciona els individus que formen part de cadascun dels grups experimentals i, una vegada administrat el tractament, en mesura els efectes en una o més variables dependents. D'aquesta manera, quan disposa d'una mostra suficientment ampla, l'investigador iguala els diferents grups experimentals en relació amb qualsevol factor o variable de confusió, de manera que la seva influència en la variable dependent queda neutralitzada gràcies a l'assignació aleatòria dels participants. Tot i que d'acord amb aquesta lògica general un experiment pot adoptar formes molt més complexes, el seu tret característic rau en la capacitat que dóna a l'investigador per a atribuir, més enllà de les petites diferències entre els grups degudes a l'atzar, les variacions observades en la variable dependent com una conseqüència necessària de la manipulació de la seva variable independent o tractament.

D'altra banda, és possible caracteritzar la investigació no experimental com la que es produeix quan l'investigador no té el control sobre els diferents nivells o condicions d'una o més variables independents. Aquest tipus d'investigació pot adoptar moltes formes, però la més freqüent és el *qüestionari* o l'*enquesta*. En aquest context, l'investigador defineix les seves variables independents i, com que no té la capacitat de manipular-les d'acord amb la seva voluntat, es limita a observar-les a partir de les respostes proporcionades per una mostra generalment ampla de participants en una o més ocasions al llarg del temps. Una vegada administrat el qüestionari, l'investigador identifica els individus que formen part dels diferents grups prèviament existents i en mesura les diferències en una o més variables dependents. D'aquesta manera, amb una certa confiança, atribueix aquestes diferències a les variacions en la variable independent, però, a diferència de la investigació experimental, no serà possible evitar la intervenció de potencials factors o variables de confusió, de manera que resulta difícil excloure la possibilitat que la seva influència es converteixi en una explicació alternativa a la que proposa.

Aquests dos tipus d'investigació difereixen en la seva *validesa interna*, és a dir, en la seva capacitat per a proporcionar les evidències necessàries que permetin determinar l'existència d'una relació de causalitat a partir de l'observació d'una associació entre dues variables. Òbviament, els resultats d'un únic estudi no són mai suficients per a donar per provada una relació d'aquest tipus. Però el fet que els investigadors utilitzin, sempre que els resulti possible, l'assignació aleatòria dels individus als diferents grups que caracteritza la metodologia experimental, els pot permetre obtenir evidències més sòlides per a portar a terme judicis causals a partir dels seus resultats. Aquest no és, però, l'únic moment en què l'atzar fa un paper important en el disseny de la investigació. De fet, resulta determinant quan els investigadors es proposen, com

sol ser habitual, generalitzar les seves conclusions més enllà dels límits dels seus estudis particulars. Amb independència del tipus d'investigació, sigui un experiment o una enquesta, és el moment del disseny i la construcció de la mostra quan els investigadors han de triar els participants que, finalment, formaran part dels seus estudis.

Atès que, per raons pràctiques, no sempre és possible obtenir informació sobre el conjunt de la població que es proposa analitzar una investigació, sovint els investigadors porten a terme un procés de selecció amb l'objectiu d'escollir només una fracció, un subconjunt, del total d'individus que la formen. En aquest sentit, és possible identificar dos grans tipus d'estratègies per a la tria dels participants de qualsevol investigació: la *selecció aleatòria* o *probabilística* i la *selecció no aleatòria* o *intencional*. En aquest sentit, considerem que una mostra és aleatòria quan tots i cadascun dels individus que formen part de la població tenen la mateixa probabilitat de ser escollits per a participar en la investigació. Partint d'una definició clara i precisa de la població objecte d'estudi, en condicions ideals els investigadors haurien de ser capaços d'identificar-ne tots els membres –per exemple, a partir d'una llista amb els seus noms–, i, a continuació, procedirien a escollir a l'atzar els seus participants. En canvi, una mostra és no aleatòria quan els individus no han estat escollits fent servir aquesta estratègia sinó que, més aviat, són senzillament el producte accidental d'una tria intencional segons la seva conveniència o la seva disponibilitat.

Tot i que una mostra aleatòria pot adoptar formes molt més complexes, és convenient assenyalar que només quan el criteri de selecció dels participants és aleatori tindrem les garanties suficients per a considerar que les mostres són representatives. D'aquesta manera, els investigadors tindran la confiança que les relacions observades a partir de l'associació entre les seves variables són extrapolables al conjunt de la població a partir de la qual han estat extretes les mostres. És per aquesta raó que tant la investigació experimental com la no experimental no només difereixen en la seva validesa interna, sinó que també poden fer-ho en la seva *validesa externa*, és a dir, en la seva capacitat per a proporcionar les evidències necessàries que permetin determinar que l'existència d'una relació és generalitzable a altres situacions o altres individus que no han format part de l'estudi.

La taula 6 presenta esquemàticament la relació entre la selecció i l'assignació dels participants en el disseny de la investigació, que, a continuació, ens permetrà posar de relleu la important contribució de l'atzar al procés d'inferència estadística.

Taula 6. La relació entre el disseny de la investigació i la inferència estadística

	Assignació aleatòria	**Assignació no aleatòria**	
Selecció aleatòria	Relació causal generalitzable	Relació no causal generalitzable	**Alta validesa externa**

Font: elaboració pròpia.

Selecció no aleatòria	Relació causal no generalitzable	Relació no causal no generalitzable	**Baixa validesa externa**
	Alta validesa interna	**Baixa validesa interna**	

Font: elaboració pròpia.

D'acord amb aquesta taula, l'encreuament de les diferents formes amb què poden ser seleccionats i assignats els individus als diferents grups proporciona quatre tipus bàsics d'investigacions, que difereixen, fonamentalment, en la seva validesa. D'una banda, el quadrant superior esquerre representa la investigació que, a través del seu disseny, porta a terme una selecció i una assignació aleatòries dels seus participants. Seria el cas d'un experiment desenvolupat a partir d'una mostra representativa, en què la validesa interna i externa de la investigació serien òptimes i, per tant, els investigadors es trobarien en les millors condicions per a establir una relació causal que també fos generalitzable a la població. Al seu torn, als quadrants superior dret i inferior esquerre trobem les investigacions que només porten a terme una assignació o una selecció no aleatòries i que, per tant, tindrien una validesa interna o externa més baixa, respectivament. En el primer cas, es tractaria d'una enquesta administrada a una mostra representativa, que permetria establir relacions generalitzables al conjunt de la població però que, en cap cas, proporcionaria evidències suficients per a determinar-ne la naturalesa. En el segon, es tractaria del cas d'un experiment portat a terme a partir d'una mostra no representativa, que proporcionaria evidències sobre la naturalesa causal de la relació però que, en canvi, no en permetria la generalització. Finalment, en el pitjor dels escenaris des del punt de vista de la validesa, el quadrant inferior dret representa la investigació que no porta a terme ni una selecció ni una assignació aleatòries i que, per tant, com seria el cas d'una enquesta dirigida a una mostra no representativa, no permetria establir ni una relació causal ni generalitzar-ne els resultats al conjunt de la població.

Aquests quatre tipus d'investigació difereixen fonamentalment en la seva validesa i, tal com hem pogut veure, la raó per la qual això és així no és cap altra que el paper que assignen a l'aleatorització en el seu disseny. En aquest sentit, la diferent capacitat que tenen els investigadors per a determinar l'existència d'una relació causal generalitzable al conjunt de la població serveix com una bona il·lustració de la contribució de l'atzar a la *inferència estadística*. Si entenem la inferència estadística com el procés a través del qual podem extreure conclusions generals a partir de l'anàlisi de les dades d'una mostra, és necessari tenir present que aquest procés únicament és possible si la selecció o l'assignació dels participants als diferents grups han estat aleatòries. És a dir, només quan l'atzar intervé en almenys un d'aquests dos moment importants per al disseny de la investigació és possible arribar a concloure si les diferències observades en la variable dependent són conseqüència de la manipulació de la variable independent o tractament – *inferència causal*–, o si aquestes diferències són generalitzables més enllà de la mostra – *inferència a la població*–. D'aquesta manera, sempre que es compleixi aquesta condició, l'estadística inferencial

proporciona un conjunt de procediments que permet als investigadors avaluar les diferències observades i decidir, amb un determinat nivell de confiança, fins a quin punt responen a una diferència realment existent a la població o, en canvi, poden ser senzillament explicades com a resultat de l'atzar en la selecció i/o en l'assignació dels participants.

4. Què és l'anàlisi multivariant i per a què serveix?

Tot i la importància del disseny de la investigació per tal de poder extreure conclusions no esbiaixades que, a més a més, siguin generalitzables més enllà dels límits dels estudis particulars, el cert és que els investigadors no sempre poden utilitzar experiments per a desenvolupar els seu treballs. En aquest sentit, qüestions d'ordre pràctic o ètic poden desaconsellar o, fins i tot, impedir que es porti a terme una assignació aleatòria dels participants a les diferents condicions experimentals. Aquesta situació és bastant freqüent a les ciències socials i és especialment evident quan els estudis es desenvolupen, lluny de les condicions controlades dels laboratoris, als contextos naturals en què és produeix l'activitat quotidiana de les persones. Si, tal com plantejàvem a l'inici d'aquest mòdul, l'objectiu és analitzar fenòmens complexos com la discriminació per raó de gènere en l'accés dels estudiants a la universitat, resulta obvi que no serà possible decidir el gènere dels candidats ni, de la mateixa manera, tampoc no es podrà escollir el departament a què els candidats haurien de presentar les seves sol·licituds. De fet, fins i tot quan es reuneixen les condicions idònies per a fer servir experiments, els investigadors no sempre poden preveure o controlar adequadament, mitjançant el disseny de la seva investigació, tots i cadascun dels potencials factors de confusió que podrien amenaçar les seves conclusions.

És en aquest context, en què la manipulació de les variables no és una estratègia factible o suficient per a obtenir evidències sòlides que permetin portar a terme judicis de causalitat a partir de l'observació d'associacions entre variables, que l'anàlisi multivariant es presenta com el marc analític general que permet modelar les múltiples relacions existents entre les diferents variables involucrades en una investigació. En aquest sentit, podem definir l'*anàlisi multivariant* com el conjunt de tècniques estadístiques que tenen com a objectiu analitzar i interpretar les relacions entre diverses variables de manera simultània, mitjançant la construcció de models estadístics complexos que permeten distingir la contribució independent de cadascuna d'aquestes en el sistema de relacions i, d'aquesta manera, descriure, explicar o predir els fenòmens objecte d'interès per a la investigació. Aquest marc analític general, per tant, ofereix als investigadors l'oportunitat de portar a terme un control estadístic de qualsevol tercera variable que, com a eventual factor de confusió, pogués interferir en la relació entre les variables dependents i independents. És important tenir present, però, que l'elecció de les tècniques estadístiques –i l'anàlisi multivariant no n'és una excepció– no té cap relació amb el disseny que hagi estat emprat en la investigació, de manera que aquestes tècniques poden ser utilitzades per a analitzar les dades obtingudes tant en els contextos experimentals com en els no experimentals. Tal com ja hem pogut discutir àmpliament, l'única limi-

tació es troba en el moment de la interpretació dels resultats i, especialment, en el risc que els investigadors estiguin disposats a assumir per a determinar l'existència de les seves relacions a partir de les evidències de què disposen.

D'una manera senzilla, podem entendre l'anàlisi multivariant com una extensió de l'anàlisi bivariant i aquest, al seu torn, com una extensió de l'anàlisi univariant. En aquest sentit, l'*anàlisi univariant* és la forma més simple d'anàlisi estadística i es proposa la descripció de la distribució d'una única característica dels individus que formen part de la investigació. Mitjançant la construcció d'una taula de freqüències en el cas d'una variable qualitativa o bé el càlcul d'una mesura de tendència central –com la mitjana, la mediana o la moda– o de la seva dispersió –com el rang, la desviació estàndard o la variància– quan es tracta d'una variable quantitativa, la clau d'aquest tipus d'anàlisi es troba en el fet que només pren en consideració una única variable amb l'objectiu de descriure la mostra i, quan és possible, establir una inferència sobre la població que representa. Òbviament, quan els investigadors porten a terme els seus estudis mai concentren tots els seus esforços a observar únicament una variable, però, sigui quin sigui el nombre de mesures registrades en la investigació, aquest primer tipus d'anàlisi es limita a explorar cadascuna de les seves variables independentment. Així, reprenent el cas de l'estudi sobre la discriminació de gènere en l'accés a la universitat, l'estadística univariant ens permet conèixer la proporció d'estudiants de la mostra que serien homes o dones, els departaments que haurien escollit per presentar les seves sol·licituds, o la quantitat de candidats que finalment haurien estat acceptats o rebutjats per la universitat.

D'altra banda, l'*anàlisi bivariant* és una extensió de l'anàlisi univariant i, tot i mantenir la seva naturalesa exploratòria, es proposa en canvi determinar la relació existent entre dues característiques dels participants de la investigació. Mitjançant la construcció d'una taula de contingència quan es tracta de variables qualitatives o bé el càlcul d'una correlació en el cas de variables quantitatives, aquest tipus d'anàlisi té com a objectiu examinar la distribució d'una variable dependent, resultat o explicada en funció dels nivells d'una altra variable independent, predictora o explicativa. D'aquesta manera, l'observació de la seva associació permet determinar l'existència d'una relació a la mostra i, sempre que sigui possible, establir una inferència sobre la població que representa. Tal com ja hem dit, la mera evidència d'una associació entre dues variables des del punt de vista estadístic no implica, necessàriament, l'existència d'una relació causal. I això és degut, en darrer terme, al fet que aquest segon tipus d'anàlisi permet als investigadors tenir en compte les relacions entre totes i cadascuna de les possibles parelles de les seves variables, però ho fa, en cada ocasió, independentment. D'aquesta manera, no és possible descartar que qualsevol altra variable pugui interferir en aquestes relacions actuant com un potencial factor de confusió i, per tant, alterant o fins i tot fent evidents relacions entre dues variables que no haurien estat observades inicialment. Seguint amb el nostre cas, l'estadística bivariant ens permetria conèixer la relació entre

el gènere dels candidats i la seva acceptació final als programes de la universitat o, el que ha resultat més important, la relació del departament tant amb l'acceptació com amb el gènere dels candidats.

En aquest sentit, com a extensió de l'anàlisi bivariant, l'*anàlisi multivariant* es presenta com el marc analític general que es proposa analitzar i interpretar les relacions existents entre diverses variables, però ho fa, en aquest cas, mitjançant la construcció de models complexos que permeten determinar-ne l'existència de manera simultània. D'aquesta manera, més enllà de la consideració de les variables dependents i independents, aquest tipus d'anàlisi permet als investigadors incorporar als seus estudis les *variables de control* que siguin necessàries, és a dir, totes les variables estranyes que eventualment podrien actuar com a factors de confusió i que, per tant, podrien interferir en les relacions que són realment objecte del seu interès. Controlant estadísticament la contribució de totes aquestes variables al sistema de relacions, aquest tercer tipus d'anàlisi permet mantenir constantment els seus efectes i obtenir així una estimació més precisa de les relacions realment existents entre les variables dependents i les independents. Per tant, l'observació de les associacions entre les diferents variables considerades en la construcció dels models permet determinar l'existència de múltiples relacions a la mostra de participants i, quan es reuneixen les condicions necessàries, establir inferències sobre el conjunt de la població. De fet, com veurem més endavant, aquest marc analític no només permet analitzar les relacions de dependència entre les diferents variables involucrades en una investigació, sinó que també serveix per a analitzar, tenint en compte la seva interdependència, les relacions entre les variables que no poden ser considerades ni dependents ni independents des d'un punt de vista teòric. A fi d'acabar amb el cas que ens ha servit de fil conductor en aquesta introducció, l'estadística multivariant permetria conèixer la contribució simultània de les característiques dels estudiants i dels departaments a què haurien presentat les seves sol·licituds que estarien implicades en l'acceptació final dels candidats. Més enllà del paper del departament com a potencial factor de confusió, aquesta investigació podria tenir en compte també les diferències entre els homes i les dones en les seves capacitats, aptituds o habilitats controlant-ne, per exemple, l'expedient acadèmic previ o els resultats en les proves d'accés, de manera que seria possible extreure una conclusió encara més exacta sobre l'existència d'una discriminació per raó de gènere en l'accés dels estudiants a la universitat.

Seria convenient tenir present, però, que no tots els autors comparteixen aquesta manera d'entendre l'anàlisi multivariant. De fet, un corrent alternatiu considera que aquesta aproximació és poc restrictiva i, en canvi, defineix aquest tipus d'anàlisi com el que s'utilitza en investigacions que consideren múltiples variables dependents. En aquest sentit, entenen també l'anàlisi multivariant com una generalització de l'anàlisi univariant i bivariant, però ho fan prenent com a punt de partida definicions diferents d'aquests dos tipus d'anàlisi. D'una banda, defineixen l'estadística univariant com la que, en contextos experimentals, s'ocupa d'una única variable dependent i, per tant, no

exclou la possibilitat que els investigadors considerin més d'una variable independent a la seva anàlisi. D'altra banda, entenen l'estadística bivariant com l'estudi de les relacions entre parelles de variables que haurien estat obtingudes en investigacions no experimentals, de manera que, d'acord amb aquesta argumentació, no seria possible distingir entre variables dependents i independents. En aquest sentit, l'estadística multivariant no seria més que una generalització de l'anàlisi univariant, en què, sigui quin sigui el nombre de variables independents considerades, els investigadors amplien el nombre de variables dependents en la construcció dels seus models.

Aquesta aproximació alternativa planteja, però, alguns inconvenients que fan poc interessant la seva adopció. D'una banda, estableix una relació directa entre el disseny de la investigació i el tipus d'anàlisi que és possible desenvolupar. Estrictament parlant, en canvi, l'anàlisi estadística no imposa cap requeriment en relació amb la naturalesa experimental de les dades obtingudes, de manera que, com ja hem assenyalat, és responsabilitat dels investigadors valorar fins a quin punt les evidències observades d'associació entre les seves variables són suficients per a determinar l'existència de relacions de causalitat als seus estudis. D'altra banda, focalitza l'atenció únicament en les relacions de dependència entre les variables i, per tant, exclou la possibilitat que aquest marc analític general serveixi també per a analitzar les seves relacions d'interdependència. Finalment, limita el seu abast a les investigacions que consideren com a mínim dues variables dependents i, d'aquesta manera, omet altres escenaris igualment interessants en què els investigadors es proposen l'objectiu de determinar la contribució simultània de diverses variables independents en una única variable dependent.

En qualsevol cas, la clau de l'anàlisi multivariant com a marc analític general no és que els investigadors disposin de múltiples variables, perquè, com ja hem dit, els estudis no són mai dissenyats amb l'objectiu d'observar una única variable. El tret distintiu d'aquest tipus d'anàlisi és la seva capacitat de modelar les múltiples relacions existents entre les diferents variables involucrades en una investigació de manera simultània. En aquest sentit, la construcció de models complexos tant de dependència com d'interdependència comparteix una lògica comuna que es basa en la *combinació lineal de variables*. Per a ferho, en funció dels objectius de la seva investigació i, especialment, del tipus de relacions que es plantegen estudiar des d'un punt de vista teòric, els investigadors disposen de diferents procediments per a estimar, a partir de les dades obtingudes dels seus participants, el pes específic o la importància relativa de cadascuna de les variables considerades en els seus models i, d'aquesta manera, portar a terme una avaluació de la seva contribució independent al sistema de relacions. D'una banda, en el context de les relacions de dependència, la combinació lineal de variables en què es basa l'anàlisi multivariant serveix per a explicar o predir les dependents a partir de les independents i, per tant, ofereix la possibilitat de controlar l'efecte de qualsevol factor o variable de confusió que pogués interferir en les relacions que són realment d'interès per a la investigació. De l'altra, en el context de l'anàlisi de les relacions d'interdependència,

serveix per a descriure l'estructura compartida per un conjunt de variables que no poden ser identificades com a dependents ni com a independents i, per tant, ofereix la possibilitat de determinar l'existència d'una mena de supervariable o dimensió hipotètica subjacent que, malgrat no ser directament observable, podria resultar interessant interpretar.

5. Una classificació de les tècniques d'anàlisi multivariant

Definit l'anàlisi multivariant com el marc analític general que permet modelar les múltiples relacions existents entre les diferents variables involucrades en una investigació, és el moment de presentar una classificació de les diferents tècniques disponibles. Aquesta classificació general té com a objectiu oferir una panoràmica sobre les seves característiques i les condicions en què poden ser utilitzades i, de manera particular, oferir una guia que permeti al lector interessat escollir la tècnica que millor s'ajusti a la seva investigació. Tot i que, com hem dit, les tècniques d'anàlisi multivariant poden ser utilitzades per a analitzar les dades obtingudes tant en contextos experimentals com no experimentals, és important tenir present que l'elecció de la tècnica depèn de dos aspectes estretament vinculats amb el disseny de la investigació: la pregunta o objectiu general que en motiva el desenvolupament i les característiques de les dades que proporciona per a oferir una resposta. En aquest sentit, tal com hem pogut veure, l'ús de les tècniques d'anàlisi multivariant és convenient quan els investigadors es proposen respondre preguntes que tenen a veure amb l'estudi de les múltiples relacions existents, ja siguin de dependència o d'interdependència, entre les diferents variables involucrades en una investigació de manera simultània. Abans d'aprofundir en els escenaris particulars en què es pot concretar l'estudi de les relacions en aquests dos contextos, però, abordarem breument la qüestió relativa a les característiques de les dades que proporciona la investigació.

Amb independència de l'objectiu general que es plantegi, tota investigació quantitativa es basa en l'obtenció de les evidències necessàries que permetin als investigadors establir inferències a partir de l'observació d'associacions entre les seves variables. Per a fer-ho, els investigadors no només hauran de planificar com es conduirà la seva investigació sinó que, a més a més, hauran de decidir com es codificarà i registrarà la informació relativa als seus participants, de manera que pugui ser tractada mitjançant les proves estadístiques oportunes. És el moment de la mesura, el procés a través del qual els investigadors defineixen les variables d'interès i estableixen els diferents nivells que poden adoptar per tal de reflectir adequadament la variabilitat observada en els fenòmens que es proposen estudiar. Tot i que pot ser un procés complex, especialment a les investigacions que es basen en l'avaluació d'atributs psicològics no directament observables (vegeu Meneses i altres, 2013, per a una discussió més àmplia), la mesura no seria una altra cosa que l'establiment d'una correspondència entre les propietats dels fenòmens objecte d'interès i els nombres que les representen en una determinada escala. D'aquesta manera, és possible

distingir dos grans tipus de variables en funció de l'escala de mesura que hagi estat utilitzada per a definir-les: les variables qualitatives i les variables quantitatives.

D'una banda, les *variables qualitatives* o *no mètriques* són aquelles en què l'assignació dels nombres que representen els seus diferents nivells es correspon amb la presència o absència d'una determinada característica. Aquest tipus de variables no reflecteix el grau o la quantitat amb què la característica és present sinó que, en canvi, únicament permeten distingir discretament els individus que compleixen les condicions per pertànyer a un determinat nivell d'entre tots els possibles. En aquest sentit, les variables qualitatives poden ser definides a partir de l'ús d'escales nominals i ordinals, quan els seus nivells serveixen per a identificar, respectivament, individus que pertanyen a grups que són simplement diferents o que ocupen una posició relativa diferent en una sèrie ordenada. En el primer cas, utilitzen una escala nominal les variables que permeten codificar alguns atributs sociodemogràfics clàssics com són per exemple el gènere, l'ocupació o la religió i, en el context de la investigació experimental, el fet que els individus hagin estat assignats o no a una de les condicions experimentals. En el segon cas, utilitzen una escala ordinal les variables que també permeten tenir en compte l'existència d'un determinat ordre entre els seus nivells com, per exemple, l'estatus socioeconòmic o el nivell educatiu assolit però que, en cap cas, reflecteixen amb precisió la quantitat o el grau amb què la característica hi és present. Un cas particular de les variables qualitatives són les *dicotòmiques*, que únicament poden tenir dos nivells i que, en el context del desenvolupament de models multivariants, serveixen per a recodificar la informació recollida a les variables qualitatives de tres o més nivells, de manera que és possible crear una sèrie de noves variables –anomenades *fictícies* o *dummies*– que identifiquen els individus que pertanyen a cadascun dels grups en comparació amb la resta.

D'altra banda, les *variables quantitatives* o *mètriques* són aquelles en què l'assignació dels nombres que representen els seus diferents nivells es correspon exactament amb el grau o la quantitat amb què una determinada característica hi és present. Aquest tipus de variables permet distingir els individus en funció de la magnitud relativa amb què s'expressa la característica i, el que és més important, els valors que poden adoptar es corresponen amb unitats de mesura constants, de manera que qualsevol diferència entre ells reflecteix una diferència equivalent en relació amb la característica representada. En aquest sentit, les variables quantitatives poden ser definides a partir de l'ús de les escales d'interval i de raó, quan entre els seus nivells existeix un punt zero arbitrari o, en canvi, quan aquest punt zero és real i per tant representa una absència absoluta de la característica. En el primer cas, utilitzen escales d'interval les variables que recullen informació, per exemple, sobre el rendiment en un examen, els resultats d'una prova d'intel·ligència o les puntuacions obtingudes amb tests o qüestionaris dissenyats per a avaluar atributs psicològics no directament observables. Tot i que no sempre és possible demostrar l'existència d'una unitat de mesura constant en tots aquests casos, i que per tant molts

autors consideren que en realitat la seva escala hauria de ser considerada com a ordinal, el cert és que la investigació desenvolupada a les ciències socials tracta sovint aquestes variables com si realment fossin d'interval sempre que la seva distribució sigui aproximadament normal. Finalment, en el segon cas, utilitzen una escala de raó variables com l'edat, els ingressos o qualsevol tipus de recompte, en què l'existència d'un valor zero significatiu permet fer comparacions a partir de la seva magnitud i afirmar que un determinat valor és múltiple d'un altre.

Com hem dit, la distinció entre variables qualitatives i quantitatives en funció de l'escala utilitzada per a definir-ne els nivells té importants implicacions per al procés de mesura. En aquest sentit, els investigadors han d'escollir les que millor reflecteixin la variabilitat observada en els fenòmens objecte d'interès i, per tant, siguin capaces de recollir adequadament la informació relativa a la presència o absència d'unes determinades característiques o, quan els seus estudis ho requereixen, el grau o la quantitat amb què són presents en els participants. Però el que és més rellevant per a una introducció a l'anàlisi multivariant com aquesta, la distinció entre variables qualitatives i quantitatives té també algunes implicacions importants per a la construcció de models complexos que permetin analitzar i interpretat múltiples relacions de manera simultània. D'una banda, els investigadors han de conèixer, i tenir sempre ben present, l'escala de mesura de les seves variables per tal d'incorporar-les adequadament en els seus models. Això és especialment rellevant quan s'utilitzen variables qualitatives, ja que els valors que representen els seus diferents nivells no són més que etiquetes numèriques arbitràries que serveixen per a identificar els diferents grups de participants, que, en cap cas, reflecteixen el grau o la quantitat amb què una determinada característica és present en els individus. Si bé és cert que en algunes ocasions és possible tractar com a quantitatives algunes variables que, en principi, tindrien una escala ordinal, els investigadors hauran d'examinar-ne la distribució i comprovar que, almenys, és aproximadament normal. D'altra banda, com veurem a continuació, l'escala de mesura de les variables dependents i independents és un condicionant important en el moment de la tria de la tècnica d'anàlisi multivariant més adient per a assolir els objectius de la investigació.

Una vegada abordades les implicacions de les característiques de les dades que proporciona la investigació quantitativa, estem en disposició de classificar les tècniques d'anàlisi multivariant en funció de la pregunta o objectiu general que motiva la investigació. Tal com ens hem proposat, aquesta classificació ens permetrà oferir una panoràmica general sobre les seves característiques i les condicions en què poden ser utilitzades, de manera que, en darrer terme, pugui servir de guia per a orientar els investigadors en el moment d'escollir la tècnica que millor s'ajusti a la seva investigació. Malgrat que la diversitat de tècniques disponibles ens impedeix abordar-les totes i cadascuna, aquesta classificació servirà per a presentar-ne algunes de les més utilitzades i, a més a més, introduir les que seran tractades amb més detall en els mòduls posteriors d'aquest text. Per a fer això, organitzarem aquesta exposició a partir dels dos

grans contextos de dependència i interdependència en què, com hem dit, la construcció de models multivariants permet analitzar i interpretar les relacions existents entre les diferents variables involucrades en una investigació de manera simultània i, d'aquesta manera, distingir la contribució independent de cadascuna d'aquestes en el sistema de relacions. A continuació, considerarem els escenaris particulars en què aquest marc analític general pot ser utilitzat i proposarem algunes de les alternatives, presentades esquemàticament a la taula 7, de què disposen els investigadors en funció de les característiques de les seves dades.

Taula 7. Una classificació de les tècniques d'anàlisi multivariant en funció dels objectius de la investigació i les característiques de les dades

Objectiu general	Escenari d'aplicació	Característiques de les dades	Tècnica multivariant
Analitzar relacions d'interdependència per a descriure l'estructura de les dades	Identificació de grups de característiques similars	Diverses variables quantitatives	Anàlisi de components principals
			Anàlisi factorial
		Diverses variables qualitatives	Anàlisi de correspondències
	Identificació de grups d'individus similars	Diverses variables quantitatives o qualitatives	Anàlisi de conglomerats
	Identificació de grups d'objectes similars	Diverses variables quantitatives o qualitatives	Escalat multidimensional
Analitzar relacions de dependència per a fer prediccions o explicacions	Explicació de la variabilitat dels individus	Una variable dependent quantitativa	Regressió múltiple
		Dues o més variables dependents quantitatives	Correlació canònica
	Explicació de la variabilitat dels grups d'individus	Una variable dependent quantitativa	ANOVA de dos o més factors o ANCOVA
		Dues o més variables dependents quantitatives	MANOVA o MANCOVA
	Predicció de la pertinença dels individus a grups	Una variable dependent qualitativa	Anàlisi discriminant
			Regressió logística
Analitzar relacions de dependència i interdependència de manera simultània	Avaluació de l'ajustament de models concatenats	Diverses variables quantitatives	Equacions estructurals

Font: Elaboració pròpia.

6. Una guia per a l'elecció de les tècniques d'anàlisi multivariant

En primer lloc, quan no és possible distingir entre variables dependents i independents, els investigadors es mouen en el context de la interdependència i, per tant, l'objectiu general de la seva anàlisi és descriure l'estructura subjacent a les seves dades. En aquest sentit, quan la seva intenció és analitzar les relacions simultànies entre diverses variables quantitatives per tal d'identificar grups de característiques similars, les tècniques més adequades són l'*anàlisi de components principals* i l'*anàlisi factorial*. Totes dues tècniques tenen com a objectiu reduir la complexitat de les dades mitjançant l'obtenció d'un conjunt limitat de components o factors que permetria representar la variabilitat en les característiques dels individus d'una manera eficient, és a dir, conservant el màxim de la informació recollida originalment a les variables involucrades. Tant l'anàlisi de components principals com l'anàlisi factorial es basen en l'anàlisi i la interpretació de les associacions observades entre les variables, però difereixen, bàsicament, en la manera com determinen l'estructura de components o factors. En el primer cas, els investigadors no disposen d'una teoria sòlida sobre les relacions per a construir els seus models i, per tant, es limiten a determinar empíricament l'existència dels components que, de fet, agrupen les seves variables. En canvi, en el segon cas, els investigadors parteixen d'una teoria sobre els fenòmens objecte del seu interès que els informa dels diferents factors i, per tant, utilitzen aquests models per a posar a prova la contribució de les diferents variables d'acord amb les seves expectatives. És important tenir present, però, que tot i que existeixen alguns procediments per a poder tractar variables qualitatives, aquestes dues tècniques són generalment aplicades quan les variables analitzades són quantitatives. En cas que les variables utilitzades siguin qualitatives, els investigadors tenen al seu abast una tècnica alternativa, l'*anàlisi de correspondències*, per tal d'assolir els mateixos objectius. Mitjançant la transformació de la informació qualitativa per a poder tractar-la quantitativament, aquesta tècnica procedeix d'una manera comparable i, per tant, permet obtenir un conjunt de dimensions –similars als components o els factors– que reflectirien una estructura compartida per les variables considerades en la construcció dels models.

D'altra banda, l'estudi de les relacions d'interdependència amb l'objectiu de descriure l'estructura subjacent a les dades no només serveix per a identificar grups de característiques similars. Quan els investigadors estan interessats, en canvi, a identificar grups d'individus, la tècnica més adequada és l'*anàlisi de conglomerats*, també coneguda com *anàlisi de clúster*. En aquest sentit, aquesta tècnica ofereix un conjunt de procediments per tal de reduir la complexitat de les dades mitjançant l'obtenció d'un conjunt limitat de grups, exhaustius i mútuament excloents, que permetria representar la variabilitat dels individus a partir de la similitud de les seves característiques. Seleccionades les variables

que formaran part dels models, que poden ser quantitatives o qualitatives, i sempre en funció del procediment escollit pels investigadors, l'anàlisi de conglomerats es basa en l'anàlisi i la interpretació de l'associació observada entre els individus, de manera que el càlcul de la seva distància o proximitat serveix per a conformar grups homogenis en relació amb les característiques seleccionades, que, a la vegada, siguin tan heterogenis entre si com sigui possible. Finalment, quan el propòsit dels investigadors és identificar grups d'objectes similars a partir de les valoracions que proporcionen els participants de la seva investigació, la tècnica més adequada és l'*escalat multidimensional*. En aquest cas, a diferència del que succeeix amb les altres tècniques d'anàlisi de les relacions d'interdependència que hem introduït abans, la cerca d'una estructura a les dades no es basa en l'anàlisi i la interpretació de l'associació observada entre les característiques o els individus, sinó que parteix dels judicis comparatius que fan explícitament els participants sobre les parelles formades a partir d'un conjunt d'objectes, d'acord amb les seves preferències o les seves percepcions de similitud. Com en el cas de l'anàlisi de conglomerats, l'escalat multidimensional pot ser aplicat tant a variables quantitatives com qualitatives.

En segon lloc, quan és possible distingir entre variables dependents i independents, els investigadors es mouen en el context de la dependència i, per tant, l'objectiu de la seva anàlisi és explicar o predir les variables dependents a partir de les independents. En aquest sentit, quan la seva intenció és analitzar les relacions simultànies entre diverses variables quantitatives per tal d'explicar la variabilitat dels individus en una o més de les seves característiques, les tècniques més adequades són la *regressió múltiple* i la *correlació canònica*. Aquestes dues tècniques tenen com a objectiu comú determinar la intensitat o la magnitud de les relacions entre les diferents variables involucrades, de manera que servirien per a avaluar la contribució específica del canvi o la variació en els nivells de totes i cadascuna de les variables independents considerades en la construcció dels models. A més a més, tot i que les variables independents solen ser quantitatives, totes dues tècniques són suficientment flexibles per a permetre incorporar variables qualitatives mitjançant la creació de les corresponents variables fictícies o *dummies*. Tant la regressió múltiple com la correlació canònica es basen en l'anàlisi i la interpretació de les associacions observades entre les variables, però difereixen, bàsicament, en el nombre de variables dependents quantitatives que permeten explicar. Quan els investigadors es proposen analitzar la variabilitat dels individus en una característica i, per tant, centren la seva atenció en una única variable dependent, la seva tècnica d'elecció és la regressió múltiple. En canvi, podem entendre la correlació canònica com una extensió de la regressió múltiple que permet als investigadors incorporar diverses variables dependents en els seus models i, d'aquesta manera, analitzar la relació entre dos conjunts diferenciats de característiques dels individus.

D'altra banda, l'estudi de les relacions de dependència amb l'objectiu de fer explicacions o prediccions no només serveix per a analitzar la variabilitat dels individus en una o més característiques. Quan el propòsit dels investigadors

és, en canvi, analitzar les relacions simultànies entre diverses variables a fi d'explicar la variabilitat dels grups d'individus, les tècniques més adequades són l'*anàlisi de la variància* (ANOVA) de dos o més factors i l'*anàlisi multivariant de la variància* (MANOVA). En aquest sentit, les dues tècniques comparteixen l'objectiu de determinar l'existència de diferències entre els individus de manera agregada, de forma que permetrien avaluar la contribució específica de la seva pertinença a diferents grups –anomenats factors– formats a partir dels nivells d'una o més variables qualitatives. En aquest context, els factors actuarien com a variables independents en la construcció dels models i, tal com hem pogut veure en relació amb el disseny de la investigació, poden representar tant grups naturals, sobre els quals els investigadors no tindrien cap mena de control, com diferents condicions experimentals a què els individus han estat assignats d'acord amb la seva voluntat. L'ANOVA de dos o més factors i la MANOVA també es basen en l'anàlisi i la interpretació de les associacions observades entre les variables i, com en el cas de la regressió múltiple i de la correlació canònica, difereixen en el fet que permeten explicar la variabilitat dels grups en una o més variables dependents quantitatives, respectivament. D'altra banda, quan els investigadors estan interessats a considerar altres variables independents quantitatives –anomenades covariants– amb la intenció d'ajustar les diferències entre els grups en la construcció dels seus models, les tècniques més adequades són l'*anàlisi de la covariància* (ANCOVA) i l'*anàlisi multivariant de la covariància* (MANCOVA). Com a extensió de les dues anteriors, aquestes tècniques resulten especialment interessants en el context de la investigació no experimental, ja que permeten tenir en compte la influència d'altres característiques importants dels individus quan l'assignació als diferents grups no ha estat aleatòria.

Finalment, més enllà de permetre l'explicació de la variabilitat en una o més característiques dels individus, l'estudi de les relacions de dependència pot servir també per a predir-ne la pertinença a diferents grups. En aquest sentit, quan els investigadors es proposen analitzar les relacions simultànies entre diverses variables amb la intenció de classificar els individus en els diferents grups formats a partir dels nivells d'una variable qualitativa, les tècniques més adequades són l'*anàlisi discriminant* i la *regressió logística*. Totes dues tècniques tenen com a objectiu compartit determinar les característiques dels individus que serveixen per a predir amb encert els diferents grups a què pertanyen, de manera que permetrien avaluar la contribució específica de totes les variables independents considerades en la construcció dels models. Tant l'anàlisi discriminant com la regressió logística es basen també en l'anàlisi i la interpretació de les associacions observades entre les variables, però difereixen, fonamentalment, en el nombre de nivells que la variable dependent qualitativa pot adoptar i en el tipus de variables independents que permeten considerar en els models per a fer les prediccions. Quan els investigadors es proposen classificar amb encert els individus en relació amb els grups formats per una variable dependent qualitativa de dos o més nivells i, a més a més, ho fan prenent en consideració un conjunt de variables independents quantitatives, la seva tècnica d'elecció és l'anàlisi discriminant. És important tenir present,

però, que aquesta tècnica imposa una restricció en relació amb les variables independents, de manera que només pot ser aplicada quan totes i cadascuna d'aquestes segueixin una distribució normal. En canvi, tot i que la regressió logística és aplicable únicament quan la variable dependent és dicotòmica, el fet que hagi estat desenvolupada com una extensió de la regressió múltiple fa que no hagi de complir cap restricció i, per tant, permet considerar variables independents quantitatives o qualitatives.

En tercer lloc, per a tancar aquesta panoràmica general sobre les diferents tècniques d'anàlisi multivariant, en algunes ocasions els investigadors no es mouen exclusivament en el context de la interdependència o de la dependència, sinó que ho fan en la combinació d'aquests dos tipus de relacions. En aquest sentit, quan estan interessats a analitzar relacions entre les seves variables que poden ser de dependència i d'interdependència de manera simultània, la tècnica més adequada és la d'*equacions estructurals*. A diferència de totes les anteriors, aquesta tècnica té com a objectiu general analitzar de manera simultània les múltiples relacions existents entre diferents grups de variables, de manera que permetria avaluar l'ajustament de diversos models multivariants concatenats. Per a fer això, les equacions estructurals es basen en l'anàlisi i la interpretació de les associacions observades entre diverses variables, que, en termes generals, poden ser organitzades en dos grans tipus de models. D'una banda, d'acord amb la lògica de les relacions d'interdependència, un model de mesura que serveix per a identificar variables latents, similars als factors que proporciona l'anàlisi factorial, que representarien una estructura compartida entre diferents característiques dels individus. De l'altra, d'acord amb la lògica de les relacions de dependència, un model estructural que serveix per a definir un conjunt de relacions simultànies entre variables dependents i independents, que, per tant, seria equivalent al desenvolupament de diverses anàlisis de regressió múltiple o de correlació canònica de manera simultània. En aquest sentit, és important tenir present que malgrat que aquesta tècnica s'aplica generalment quan les variables considerades en la construcció d'aquests dos tipus de models són quantitatives, existeixen alguns procediments que permeten tractar també variables qualitatives. D'aquesta manera, les equacions estructurals es presenten com la tècnica d'anàlisi més eficient de què disposen els investigadors interessats a abordar fenòmens complexos i que, prenent com a punt de partida les evidències acumulades en multitud d'estudis previs, es proposen posar a prova o contrastar marcs teòrics sòlids i molt ben definits.

7. El procés de construcció de models multivariants

Tot i la diversitat de tècniques disponibles en funció de la pregunta o objectiu general que motiva la investigació i les característiques de les dades que proporciona per a oferir una resposta, acabarem aquesta introducció als aspectes bàsics de l'anàlisi multivariant abordant el procés de construcció dels models estadístics complexos que permeten analitzar i interpretar les múltiples relacions existents entre les diferents variables involucrades en una investigació de manera simultània. Abans de fer-ho, tal com hem mostrat al llarg d'aquest mòdul, és necessari recordar que l'anàlisi multivariant únicament adquireix tot el seu sentit en relació amb el procediment establert en la investigació quantitativa, que, en termes generals, podríem resumir de la manera següent:

1) Formular una pregunta o un objectiu general que serveixi per a abordar un problema rellevant.

2) Escollir el disseny de la investigació i especificar la mostra de participants.

3) Definir adequadament totes les variables involucrades i especificar-ne les característiques.

4) Desenvolupar o escollir els instruments necessaris per a portar a terme les mesures.

5) Recollir les evidències necessàries que permetin respondre als objectius de la investigació.

6) Resumir i tractar estadísticament les dades a fi d'avaluar les evidències obtingudes i, quan és possible, generalitzar les conclusions més enllà dels límits dels estudis particulars.

D'acord amb aquest procediment, l'anàlisi multivariant proporciona el marc analític general que permet als investigadors descriure, explicar o predir els fenòmens objecte d'interès mitjançant el desenvolupament dels models estadístics més adequats per a portar a terme una anàlisi complexa de les seves dades. En aquest sentit, amb independència de la tècnica escollida, és possible caracteritzar la construcció de models multivariants com el procés general a través del qual els investigadors obtenen una combinació lineal de variables que els permet estimar, a partir de les dades obtingudes dels participants, el pes específic o la importància relativa de totes i cadascuna d'aquestes i, així, avaluar-ne la contribució independent al sistema de relacions. Per a fer això, els investigadors utilitzen les associacions observades entre les seves variables com a evidència per a determinar l'existència de les múltiples relacions considerades de manera simultània en els seus models i, seguint els procedi-

ments establerts en funció de la tècnica aplicada, per a decidir fins a quin punt aquests models s'ajusten o són una bona representació de la realitat. Ja sigui en el context de l'anàlisi de les relacions d'interdependència, de dependència o en la combinació de tots dos, la diversitat de procediments que ofereixen les diferents tècniques disponibles segueixen aquesta lògica, de manera que comparteixen uns principis generals i, com veurem a continuació, un conjunt de fases que estructuren el procés de construcció d'aquest tipus de models. En aquest sentit, entre els principis generals que regeixen l'anàlisi multivariant podem assenyalar-ne alguns dels més importants:

- **La construcció de models multivariants requereix una fonamentació teòrica de les relacions**. Atesa la gran diversitat de possibilitats que, com hem pogut veure, pot oferir l'anàlisi multivariant de les dades, és important tenir present que el punt de partida de qualsevol investigació interessada en la seva utilització ha de ser, necessàriament, la formulació d'un problema rellevant que permeti identificar les relacions entre les diverses variables involucrades de manera simultània. En aquest sentit, resulta indispensable el desenvolupament o l'adopció d'una teoria com a marc general de referència a partir del qual sigui raonable esperar que es puguin produir les relacions objecte d'interès i que, per tant, serveixi de guia als investigadors per a definir els seus objectius particulars, determinar les característiques de les dades que proporcionarà la investigació i, en darrer terme, els permeti escollir la tècnica més adequada que serà convenient utilitzar per a portar a terme l'anàlisi multivariant. En un moment en què el suport dels diferents programaris estadístics especialitzats disponibles facilita enormement l'execució d'aquest tipus d'anàlisi, el repte important no és la computació estadística dels models multivariants, sinó precisament totes les decisions que els investigadors han de prendre per a poder arribar a construir-los amb èxit.

- **L'exploració de les dades és una condició prèvia per al desenvolupament de l'anàlisi multivariant**. Com a extensió de l'anàlisi univariant i bivariant, els investigadors no hauran de perdre de vista la important contribució que aquests dos tipus d'anàlisi fan en el moment de prendre contacte amb les dades obtingudes en la investigació. Si el fet de disposar d'un marc teòric sòlid és una condició indispensable per a construir models complexos que permetin analitzar i interpretar múltiples relacions de manera simultània, no és menys cert que, únicament quan els investigadors s'hagin familiaritzat amb la distribució de les variables involucrades i n'hagin examinat les relacions per parelles, estaran en disposició de considerar la conveniència de portar a terme una anàlisi multivariant per respondre als seus objectius d'investigació. Aquesta exploració de les dades és especialment rellevant en el context de l'anàlisi de les relacions de dependència en què, com hem pogut veure, permet obtenir els indicis necessaris per a considerar el paper de qualsevol factor o variable de confusió que pugui estar interferint en les relacions objecte d'interès i, per tant, contro-

lar estadísticament la seva influència en la construcció dels models multivariants a fi d'evitar que es converteixin en una explicació alternativa.

- **El compliment dels supòsits és un requeriment important per a l'aplicació de les tècniques d'anàlisi multivariant**. L'exploració inicial de les dades no només serveix per a determinar la conveniència de l'anàlisi multivariant sinó que, a més a més, permet als investigadors comprovar fins a quin punt es compleixen els supòsits sobre els quals es basa la tècnica escollida per a modelar les múltiples relacions entre variables de manera simultània. En aquest sentit, tal com hem exposat a la nostra classificació de les diferents tècniques disponibles, és important que corroborin que tant la naturalesa de les relacions que es proposen analitzar com les característiques de variables implicades s'ajusten als requeriments de la tècnica seleccionada. D'altra banda, més enllà dels requeriments estadístics particulars que té cada tècnica, és important tenir present que la inferència estadística assumeix també alguns supòsits importants en relació amb la distribució aproximadament normal de les variables, la linealitat de les relacions o l'homogeneïtat de les variàncies de les variables dependents al llarg dels diferents nivells de les independents. No és aquest el lloc per a aprofundir en aquesta qüestió, però és important tenir present que només quan es garanteixen aquests supòsits és possible generalitzar els resultats de la investigació més enllà dels límits dels estudis particulars.

- **La clau de l'èxit de l'anàlisi multivariant es troba en una especificació adequada dels models**. Un quart principi important per a la construcció de models multivariants és la selecció de les variables que finalment formaran part de l'anàlisi. Una vegada fonamentades teòricament les relacions i, per tant, d'acord amb els seus objectius particulars, els investigadors hauran de decidir quines d'entre totes les variables de què disposen seran utilitzades en l'especificació dels seus models. En aquest sentit, és convenient tenir present que és tan important seleccionar totes les variables que siguin pertinents des del punt de vista teòric, i, per tant, no deixar de tenir-ne en compte cap d'important, com evitar incloure qualsevol altra variable que, en realitat, no sigui rellevant per a analitzar els fenòmens objecte d'interès. És el que anomenem una especificació adequada dels models, que, en tot cas, no es correspon amb una decisió única, sinó que forma part del procés contingent i iteratiu de construcció dels models multivariants a través del qual els investigadors van afegint i traient variables en funció dels resultats que els proporcionen. L'objectiu d'aquest procés és, a més a més, l'obtenció d'uns models que siguin parsimoniosos, és a dir, capaços de representar la màxima complexitat dels fenòmens amb el nombre més petit possible de variables.

- **No és possible interpretar les relacions entre les variables sense una avaluació prèvia dels models**. Tot i que no és possible tenir totes les garanties sobre la correcta especificació dels models multivariants, i sabent per tant que no poden ser mai utilitzats com una prova definitiva o con-

cloent per a determinar si una teoria és o no correcta, el cert és que les decisions que prenen els investigadors durant tot aquest procés afecten els resultats de la seva anàlisi i, en conseqüència, condicionen necessàriament el paper de les diferents variables implicades en el sistema de relacions. Incloure o ometre una determinada variable pot fer que els models multivariants es comportin de manera diferent i, precisament per aquesta raó, és necessari que els investigadors avaluïn fins a quin punt la combinació de variables que proposen s'ajusta raonablement bé a la variabilitat observada a les dades i que, per tant, els seus models són una representació adequada de la realitat. En aquest sentit, és important tenir present que, com a representació simplificada de la realitat, tots els models són incomplets i, doncs, necessàriament incorrectes, però quan se centren en els aspectes substancials dels fenòmens es converteixen en una eina molt útil per a poder interpretar les múltiples relacions objecte d'interès per a la investigació.

- **El disseny de la investigació condiciona la inferència estadística basada en l'anàlisi multivariant**. Tal com hem discutit àmpliament, la capacitat dels investigadors per a extreure conclusions generals a partir de l'anàlisi de les dades d'una mostra està estretament relacionada amb el disseny utilitzat per a portar a terme la investigació. Com succeeix amb qualsevol altra tècnica estadística, tant la inferència causal com la inferència a la població que permet l'anàlisi multivariant només és possible si la selecció o l'assignació dels participants als diferents grups han estat aleatòries. És a dir, únicament quan l'atzar intervé en almenys un d'aquests dos moments importants per al disseny de la investigació, és possible disposar de les garanties suficients per a decidir si les múltiples associacions simultànies observades entre les variables són una evidència adequada per a determinar, amb una certa confiança, l'existència d'una relació causal generalitzable a la població que representa la mostra, o si, en canvi, podrien ser explicades simplement com a conseqüència de l'atzar. Tot i la complexitat dels fenòmens que permet abordar, és important tenir sempre present que la construcció de models multivariants no eximeix els investigadors de la seva responsabilitat en relació amb la valoració de l'adequació de les evidències de què disposen per a establir les seves inferències.

Finalment, per a cloure aquesta introducció, estem en disposició de recapitular les diferents fases que, de manera general, permeten estructurar el procés de construcció de models multivariants. Tenint en compte el procediment establert en la investigació quantitativa a partir del qual l'anàlisi multivariant adquireix el seu sentit i, particularment, prenent com a punt de partida els principis que acabem de presentar, aquestes fases ofereixen una perspectiva de conjunt sobre les qüestions més importants que hem anat discutint al llarg d'aquest mòdul i, a més a més, permeten posar en pràctica tots els coneixements, les habilitats i els valors vinculats amb la construcció de models multivariants. Atès el seu caràcter general i, per tant, amb independència de les especificitats dels procediments amb què han de ser aplicades les diferents tèc-

niques disponibles, aquestes deu fases fonamentals serveixen per a organitzar seqüencialment les diferents decisions que els investigadors han de prendre per a portar a terme una anàlisi complexa de les seves dades. D'aquesta manera:

1) Delimitació del propòsit de l'anàlisi. La construcció de models multivariants comença sempre amb una definició precisa dels objectius particulars per als quals els investigadors es proposen analitzar i interpretar les múltiples relacions entre diverses variables de manera simultània. Tal com hem pogut veure, les diferents tècniques d'anàlisi multivariant disponibles poden ser utilitzades amb multitud de finalitats, que, de manera general, permeten als investigadors descriure, explicar o predir els fenòmens objecte d'interès per a la seva investigació. Com a conseqüència de la formulació d'un problema rellevant amb una fonamentació teòrica adequada, un propòsit ben definit és el primer pas per a afrontar amb èxit el procés de construcció de models multivariants.

2) Elecció de la tècnica d'anàlisi. Definit el propòsit de l'anàlisi multivariant, el segon pas consisteix a escollir la tècnica més adequada. D'acord amb la classificació de les tècniques presentada anteriorment, és necessari que els investigadors decideixin si es mouen en el context de l'anàlisi de les relacions de dependència o d'interdependència, que identifiquin l'escenari particular en què es pot concretar l'estudi de les relacions en aquests dos contextos i, finalment, que identifiquin les característiques de les dades proporcionades per la seva investigació. Aquesta és una decisió important en el procés de construcció de models multivariants, ja que la tècnica finalment escollida condiciona els procediments que els investigadors han de portar a terme durant les següents fases.

3) Exploració inicial de les dades. Una vegada seleccionada la tècnica d'anàlisi multivariant, els investigadors han de familiaritzar-se amb la distribució de les variables involucrades i, a continuació, examinar les seves relacions per parelles. Mitjançant l'aplicació de tècniques d'anàlisi univariant i bivariant, aquest primer contacte amb les dades obtingudes en la investigació permet als investigadors determinar la conveniència de portar a terme una anàlisi multivariant per a respondre als seus objectius particulars. Tal com hem dit, aquesta és una fase important per a l'anàlisi de les relacions de dependència, que permet considerar l'existència de potencials factors o variables de confusió que seria convenient tenir en compte en el procés de construcció dels models multivariants.

4) Comprovació dels supòsits. L'exploració de les dades ha de servir, també, per a determinar fins a quin punt és convenient aplicar la tècnica escollida. D'una banda, des del punt de vista teòric, confirmant que serveix per a analitzar les relacions que els investigadors es proposen abordar. D'altra banda, des del punt de vista de les característiques de les seves dades, assegurant que la distribució de les variables s'ajusta als requeriments estadístics particulars de la tècnica. Finalment, des del punt de vista de la inferència, garantint que

les dades compleixen també amb els requeriments estadístics addicionals que implica, quan els investigadors tenen aquest objectiu, la generalització dels resultats a la població que representa la mostra.

5) Estimació del model. L'exploració de les dades i la comprovació dels supòsits per a poder aplicar les tècniques dóna pas, ara sí, a la computació estadística dels models multivariants. Seguint els procediments establerts per a la tècnica escollida, i sempre amb el suport del programari estadístic adequat, és el moment en què els investigadors obtenen la combinació lineal de variables que els permetrà estimar el pes específic o la importància relativa de cadascuna d'aquestes en el sistema de relacions. Com veurem a continuació, aquesta estimació no és més que un resultat inicial en el procés de construcció de models multivariants que haurà de ser avaluat i, si escau, revisat al llarg de les fases següents.

6) Avaluació de l'ajustament del model. Com hem dit, la interpretació de les relacions entre les diferents variables considerades en els models requereix una anàlisi del seu comportament global que, d'acord amb els procediments específics de cada tècnica, permeti determinar fins a quin punt s'ajusten a la variabilitat observada a les dades i, per tant, resulta raonable acceptar que són una representació adequada dels fenòmens objecte d'interès. Tenint en compte aquestes evidències, els investigadors hauran de portar a terme els seus judicis mitjançant la comparació de l'ajustament de les successives variants que, com a aproximacions complementàries o alternatives, puguin obtenir en el procés de construcció dels seus models.

7) Revisió i millora del model. L'avaluació de l'ajustament dels models condueix a la setena fase, en què els investigadors valoren la conveniència d'afegir o treure variables rellevants des del punt de vista teòric tenint en compte els seus efectes en el comportament global dels models. És important recordar que, en qualsevol cas, l'objectiu final d'aquest procés és obtenir models multivariants ben especificats, que a més a més siguin parsimoniosos, de manera que els investigadors han de ser capaços de fer un balanç adequat entre l'increment del nivell de complexitat dels seus models i els beneficis que això hauria de comportar en relació amb la millora substantiva en el seu ajustament a les dades.

8) Interpretació del sistema de relacions. Una vegada aconseguit un ajustament global acceptable, arriba el moment en què els investigadors poden utilitzar els seus models multivariants per a analitzar i interpretar la naturalesa de les relacions existents entre les diverses variables implicades. En aquest sentit, la combinació lineal de variables que proposen els serveix per a estimar els pesos o les ponderacions associades a cadascuna d'aquestes i, per tant, avaluar-ne la contribució independent al sistema de relacions. Quan l'objectiu de l'anàlisi és establir inferències causals o a la població, aquesta interpretació

permet determinar la significació estadística de les associacions observades a la mostra i, el que és més important, la significació que aquestes relacions tenen a la pràctica.

9) Validació del model final. Tot i que no sempre és possible, el procés de construcció de models multivariants hauria de considerar la conveniència de posar a prova l'eventual generalització de les conclusions més enllà dels límits dels estudis particulars. D'aquesta manera, els investigadors haurien de disposar d'una mostra de participants diferent de la que han utilitzat per a modelar les seves relacions, o almenys dividir la mostra en dues parts, per a poder oferir evidències que permetin confiar que els models no són una conseqüència de les especificitats de la mostra i que, en canvi, poden ser útils per a analitzar i interpretar les relacions en el conjunt de la població.

10) Comunicació dels resultats de l'anàlisi. El procés de construcció de models multivariants conclou amb l'elaboració d'un informe o publicació científica que serveix per a comunicar les principals conclusions a què ha arribat la investigació. Tenint en compte la complexitat dels fenòmens que permet abordar, l'anàlisi multivariant ha d'anar sempre acompanyada d'un esforç especial dels investigadors per a transmetre i fer accessibles els resultats dels seus estudis. En aquest sentit, és especialment rellevant l'ús d'un llenguatge senzill però acurat que permeti reflectir adequadament la naturalesa de les relacions observades i, quan és l'objectiu, fins a quin punt és possible utilitzar les evidències obtingudes per a establir inferències causals o a la població que representa la mostra.

Bibliografia

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

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

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

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

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

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

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

Meneses, J.; Barrios, M.; Bonillo, A.; Cosculluela, A.; Lozano, L. M.; Turbany, J.; 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*. Nova York: Springer.

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

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

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