
Introducció a l'anàlisi multivariant

PID_00263800

Julio Meneses

Temps mínim de dedicació recomanat: 4 hores



Julio Meneses

Professor agregat de Metodologia de la investigació dels Estudis de Psicologia i Ciències de l'Educació, investigador de l'Internet Interdisciplinary Institute (IN3) i responsable de la Unitat d'Avaluació de Projectes Institucionals de l'eLearn Center de la Universitat Oberta de Catalunya (UOC).

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Julio Meneses (2019)

Primera edició: setembre 2019

© Julio Meneses

Tots els drets reservats

© d'aquesta edició, FUOC, 2019

Av. Tibidabo, 39-43, 08035 Barcelona

Realització editorial: FUOC

Cap part d'aquesta publicació, incloent-hi el disseny general i la coberta, no pot ser copiada, reproduïda, emmagatzemada o transmesa de cap manera ni per cap mitjà, tant si és elèctric com químic, mecànic, òptic, de gravació, de fotocòpia o per altres mètodes, sense l'autorització prèvia per escrit dels titulars dels drets.

Índex

Introducció.....	5
1. El cas de la discriminació de gènere a la Universitat de Berkeley.....	7
2. Associació, confusió i causalitat.....	11
3. Disseny de la investigació i inferència estadística.....	16
4. Què és l'anàlisi multivariant i per a què serveix?.....	21
5. Una classificació de les tècniques d'anàlisi multivariant.....	26
6. Una guia per a l'elecció de les tècniques d'anàlisi multivariant.....	31
7. El procés de construcció de models multivariants.....	36
8. Bibliografia anotada.....	43
Bibliografia.....	47

Introducció

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

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

L'anàlisi multivariant pot contribuir a enriquir el debat públic sobre els fenòmens que són objecte d'interès per als professionals i els investigadors, gràcies a l'oportunitat que els ofereix per a portar a terme una anàlisi complexa de les dades obtingudes en els seus estudis. Al servei de la investigació quantitativa, i com a extensió de les tècniques d'anàlisi univariant i bivariant, l'anàlisi multivariant té com a propòsit principal modelar les múltiples relacions existents entre diverses variables de manera simultània.

La construcció de models multivariants exerceix, doncs, un paper important en el desenvolupament de les diferents disciplines basades en l'anàlisi de dades quantitatives i requereix, per tant, una atenció especial en la formació dels futurs professionals i investigadors. Conèixer-ne la lògica, les característiques específiques de les diferents tècniques disponibles, els objectius particulars que permeten assolir i les condicions en què poden ser utilitzades són alguns dels reptes importants als quals ens enfrontarem en aquest material.

Per a fer això, en aquest text ens endinsarem en els aspectes bàsics involucrats en l'anàlisi multivariant de les dades com el marc analític general que es proposa analitzar i interpretar les relacions simultànies entre diverses variables mitjançant la construcció de models estadístics complexos que permeten distingir la contribució independent de cadascuna d'elles en el sistema de relacions per, d'aquesta manera, descriure, explicar o predir els fenòmens que són objecte d'interès.

La clau d'aquest marc analític general no es troba, per tant, en el fet que els investigadors disposin de múltiples variables, sinó en la capacitat que les diferents tècniques disponibles els ofereixen per a estimar el pes específic o la importància relativa de cadascuna d'elles en els seus models. En aquest sentit, com veurem, l'anàlisi multivariant pot proporcionar les evidències necessàries que permetin establir inferències a partir de l'observació d'associacions entre les variables, de manera que sigui possible extreure conclusions no esbiaixades que, a més, siguin generalitzables més enllà dels límits dels estudis particulars sempre que sigui possible.

Aquest no és un objectiu menor i, de fet, està íntimament relacionat amb la naturalesa del disseny utilitzat en la investigació a partir de la qual s'han obtingut les dades. És per aquesta raó que, tenint en compte les paraules de Light, Singer i Willet (1990), els investigadors no han de recórrer a les potencialitats que ofereix l'anàlisi multivariant per a intentar resoldre els problemes eventu-

als que puguin sorgir en el cas que la investigació no hagi estat correctament dissenyada o desenvolupada. Al contrari, la construcció de models multivariants adquireix tot el sentit en relació amb el procediment general establert en la investigació quantitativa que, en darrer terme, és el que permet que els investigadors disposin 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 de relacions extrapolables al conjunt de la població que representa la seva mostra de participants.

Tenint en compte aquestes consideracions, començarem l'exposició prenent com a punt de partida un estudi clàssic sobre la discriminació per raó de gènere a la Universitat de Berkeley. La discussió d'aquest cas controvertit ens servirà per a introduir alguns conceptes importants, com són l'associació, la confusió i la causalitat, reconèixer explícitament la importància del disseny de la investigació per a extreure'n conclusions no esbiaixades que siguin generalitzables, i oferir una definició formal que ens permeti situar l'anàlisi multivariant com el marc analític general que permet modelar les múltiples relacions existents entre les diferents variables involucrades en una investigació de manera simultània.

Una vegada establerts aquests fonaments, desenvoluparem les implicacions de la definició descrivint-ne els objectius principals i presentarem una classificació general de les diferents tècniques disponibles que ens servirà per a oferir una panoràmica general sobre les seves característiques i les condicions en què poden ser utilitzades. D'aquesta manera, els lectors interessats disposaran d'una guia que els permetrà escollir la tècnica que millor s'ajusti a la seva investigació i, a continuació, oferirem una discussió d'alguns dels principis que, en el context de la investigació quantitativa, regeixen les diferents fases amb què és possible estructurar el procés de construcció de models multivariants.

La recapitulació d'aquestes fases ens servirà, en darrer terme, per proporcionar una perspectiva de conjunt sobre les qüestions més importants introduïdes al llarg del text. Finalment, conclourem aquesta introducció general amb una bibliografia anotada que servirà d'ajut per a complementar i ampliar la nostra aproximació als aspectes bàsics de l'anàlisi multivariant.

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

L'any 1973 va ser 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, 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 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.

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).

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 a favor de l'existència d'una discriminació per raó de gènere. De fet, si utilitzem aquesta taula de contingència per a analitzar-ne 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 ($X^2 = 111,25$, $df = 1$, $p < 0,001$). Tot i ser estadísticament significativa, però, aquesta relació no mostra una intensitat o una magnitud important (V de Cramér = 0,09).

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 preferiria els homes per davant de les dones com a estudiants dels programes de postgrau. Aquesta situació, però, resulta més complexa de la representació que ofereix l'anàlisi d'aquesta taula de contingència.

Tal com van mostrar Bickel, Hammel i O'Connell (1975), la discriminació aparent per raó de gènere es produiria únicament quan agreguem les dades per al conjunt de la Universitat. Tot i que en el seu treball no reproduïen les dades proporcionades per a cadascun dels cent un departaments que oferien aquests estudis, la seva anàlisi serveix com a 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 vuitanta-cinc departaments van resoldre les seves sol·licituds en el sentit contrari, mostrant-ne 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 aparent 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.

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

Departament	Homes		Dones	
	Sol·licituds	Percentatge d'admissió	Sol·licituds	Percentatge d'admissió
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
Total	2.691	44%	1.835	30%

Font: Freedman, Pisani i Purves (2007).

Tal com es pot observar en la taula, 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 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.

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 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 aparent contradicció.

Tenint en compte les taxes d'acceptació respectives, els departaments A i B serien els que més sol·licituds van acceptar finalment i, per tant, aquells als quals 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 tercers 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 estudiants. Però, i això és el 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 (51,47%) 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 representen gairebé la pràctica totalitat (92,75%) 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 l'elecció del departament, tal com fem en 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 que acabem d'explicar s'ha convertit en un exemple clàssic d'un fenomen que sovint es produeix en 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 el que s'ha anomenat la *paradoxa de Simpson*, expressió encunyada per Blyth (1972) a partir de l'exposició de Simpson (1951) per a 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 que va fer abans Pearson sobre l'existència de correlacions espúries entre variables quantitatives (Aldrich, 1995; David i Edwards, 2001).

Podem definir la paradoxa de Simpson 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 en cadascun dels subgrups que es conformen a partir d'una tercera variable qualitativa.

La paradoxa de Simpson no és un fenomen infreqüent en les disciplines basades en l'anàlisi de dades quantitatives, particularment en els 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 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 en cadascun dels seus departaments. En cap cas, però, és adequat interpretar aquesta aparent contradicció com el resultat d'un artefacte estadístic o com un indicatiu que la investigació hagi estat incorrectament dissenyada o desenvolupada. Les relacions observades existeixen, són reals, tant en el cas del conjunt dels candidats valorats per la Universitat de Berkeley com en el detall dels seus departaments.

El que posa de manifest la contradicció no és l'existència d'aquestes relacions en els dos nivells d'anàlisi, sinó el fet que les evidències observades d'associació entre les variables siguin emprades 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 en els 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 una **variable de confusió** és una variable estranya, no prevista o contemplada en la investigació, que pot alterar la relació entre dues variables que són 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 *variable predictora* o *explicativa*– i una **variable dependent** –també coneguda com a *variable 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 la 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, els lectors interessats poden 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 era conseqüència de la preferència dels homes i les dones pels més fàcils i més difícils d'accedir-hi, respectivament.

En qualsevol cas, 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 actua en aquest cas com un factor de confusió i, per tant, compleix la condició necessària d'estar associat tant amb el gènere dels candidats (variable independent) com amb la seva acceptació final (variable dependent). En canvi, podem il·lustrar aquest requeriment amb un exemple fictici que, 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, fins i tot, 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 haurien estat homes i 550 dones. Suposem també 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.

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

En aquest cas, tenint en compte que la taxa global d'acceptació en el conjunt dels dos departaments hauria estat del 60%, tant per als homes com per a les dones, el fet que no s'hi observi cap diferència seria una evidència en contra de l'existència d'una discriminació per raó de gènere. Si utilitzem les dades totals que es presenten en 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 hi ha 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).

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 clara a favor de l'existència d'una discriminació per raó de gènere. Tots dos departaments d'aquesta universitat preferirien, en realitat, els homes per davant de les dones com a estudiants dels seus programes.

De fet, si utilitzem les dades que es presenten en la primera i en la segona filera per a construir dues taules de contingència separades, l'anàlisi de l'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 en el cas del primer departament (V de Cramér = 0,2) que en el 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 de confusió que hauria de ser tingut en compte. Més enllà de la inspecció visual de les taxes d'acceptació de la taula 3, a continuació presentem dues taules de contingència construïdes a partir de les mateixes dades, que ens permetran determinar fins a quin punt el departament compleix la condició necessària exigida a qualsevol factor o variable de confusió i que, 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.

D'una banda, agrupant tots els candidats independentment del seu gènere, 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 estudiants 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 (33,33%) de les seves 300 sol·licituds. En comparació, el departament B seria aquell al qual hauria resultat més fàcil accedir, i hauria acceptat una mica més de dos terços (71,43%) de les 700 sol·licituds que hauria valorat.

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

D'una altra banda, agrupant ara tots els candidats independentment de la seva acceptació final als 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 (33,33%) de les sol·licituds que hauria valorat. En canvi, el departament B seria aquell al qual s'haurien presentat més dones, i hauria valorat 450 candidates que representen gairebé dos terços (64,29%) de les seves sol·licituds.

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

En aquest sentit, utilitzant aquestes dues taules de contingència per analitzar l'associació del departament amb les dues variables, podem afirmar que existeix una relació estadísticament significativa tant amb l'acceptació final dels candidats ($X^2 = 126,98$, $df = 1$, $p < 0,001$) com amb el seu gènere ($X^2 = 81,29$, $df = 1$, $p < 0,001$) que, a més a més, resulta comparativament d'una intensitat o magnitud més important en el primer cas (V de Cramér = 0,36 i 0,29, respectivament). En efecte, tal com suggeria la inspecció preliminar de les dades desagregades, el departament estaria actuant com un factor o una variable de confusió i, per tant, l'anàlisi agregada en el cas de la nostra universitat fictícia ens hauria portat a una conclusió esbiaixada.

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 les 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 incorporació d'aquests factors a l'anàlisi pot comportar l'increment, el decrement, la desaparició o, fins i tot, la inversió de les relacions observades, de manera que la mera evidència de l'existència d'una associació entre dues variables no implica, necessàriament, que aquesta relació sigui de naturalesa 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 en el context particular dels 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 emprat per obtenir-les. Si entenem l'anàlisi estadística com la culminació d'un complex procés de planificació a través del qual es porta a terme qualsevol investigació quantitativa, resulta convenient distingir dos grans tipus de dissenys: la **investigació experimental** i la **investigació observacional**.

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 poder posar a prova la seva associació mitjançant les proves estadístiques oportunes i, com dèiem, la consideració de qualsevol variable estranya que pogués actuar com a factor de confusió, és a dir, que interferís en les relacions objecte de l'anàlisi i, per tant, esdevingués 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, els lectors interessats poden 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, els investigadors tenen el control sobre els diferents nivells o les condicions d'almenys una variable independent –generalment anomenada *tractament*–, de manera que poden decidir d'acord amb la seva voluntat com seran exposats els participants. Mitjançant una assignació aleatòria, els investigadors seleccionen els individus que formen part de cadascun dels grups experimentals i, una vegada administrat el tractament, en mesuren els efectes en una o més variables dependents.

Així, quan disposen d'una mostra suficientment ampla, els investigadors igualen els diferents grups experimentals en relació amb qualsevol factor o variable de confusió, de manera que la seva influència en la variable dependent quedi neutralitzada gràcies a l'aleatorització 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 als investigadors per a atribuir, més enllà de les petites diferències entre els grups a causa de l'atzar, les variacions observades en la variable dependent com una conseqüència necessària de la manipulació de la variable independent o tractament.

D'altra banda, és possible caracteritzar la **investigació observacional** com la que es produeix quan els investigadors no tenen el control sobre els diferents nivells o les condicions d'una o més variables independents. Aquest tipus d'investigació pot adoptar moltes formes, però una de les més freqüents és basa en la utilització d'un **qüestionari** o una **enquesta**. En aquest context, els investigadors defineixen les seves variables independents i, com a conseqüència de la impossibilitat de manipular-les d'acord amb la seva voluntat, es limiten a observar-les a partir de les respostes proporcionades per una mostra generalment ampla de participants.

Una vegada administrats els seus qüestionaris, els investigadors identifiquen els individus que formen part dels diferents grups prèviament existents i en mesuren les diferències en una o més variables dependents. D'aquesta manera, amb una certa confiança, atribueixen aquestes diferències a les variacions existents en la variable independent. A diferència de la investigació experimental, però, en aquest escenari no serà possible evitar la intervenció de potencials

factors o variables de confusió en les relacions observades, de manera que els resultarà difícil excloure la possibilitat que la seva influència es converteixi en una explicació alternativa a la que proposen.

Aquests dos tipus d'investigació difereixen en la seva **validesa interna**, és a dir, en la capacitat per proporcionar les evidències necessàries que permetin determinar l'existència d'una relació de causalitat a partir de l'observació d'una associació entre les variables dependents i independents. 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 juga un paper important en el disseny de la investigació. De fet, resulta també 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 experimental o observacional, és el moment del disseny i la construcció de la mostra, quan els investigadors han de seleccionar els participants que, finalment, acabaran formant-ne 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 en 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 conformen. 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**.

De manera sintètica, considerem que una **mostra és aleatòria** quan tots i cadascun dels individus que formen part de la població tenen la mateixa probabilitat de ser seleccionats per formar part de la investigació. Partint d'una definició clara i precisa de la població que és 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 noms– i, a continuació, procedirien a escollir a l'atzar a aquells que finalment en seran els 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 disponibilitat. És per aquesta raó que, d'acord amb aquesta segona estratègia, no tots els individus que conformen la població d'interès tenen, de fet, la mateixa probabilitat de ser seleccionats.

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 seran 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 l'observacional, no només difereixen en la seva validesa interna, sinó que també poden fer-ho en la seva **validesa externa**. És a dir, en la capacitat per a proporcionar les evidències necessàries que permetin concloure amb les garanties suficients que l'existència d'una relació és generalitzable a altres situacions o a 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 en relleu la important contribució que té l'atzar en el 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
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. En primer lloc, el quadrant superior esquerre representa la investigació que, mitjançant el seu disseny, porta a terme una selecció i una assignació aleatòries dels 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 a partir de l'observació de les relacions entre les seves variables que també fos generalitzable a la població.

Al seu torn, als quadrants superior dret i inferior esquerre trobem les investigacions que únicament porten a terme una selecció o una assignació no aleatòries i que, per tant, tindrien una validesa interna o externa, respectivament, més baixa. 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 causal. 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 al conjunt de la població.

Finalment, en el pitjor dels escenaris possibles des del punt de vista tant de la validesa interna com de l'externa, el quadrant inferior dret representa la investigació que no porta a terme ni una selecció ni una assignació aleatòries dels participants. Aquest seria el cas d'una enquesta dirigida a una mostra no representativa en què, per tant, no seria possible establir ni la naturalesa causal de les relacions observades ni generalitzar-ne les conclusions obtingudes 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 juga l'**atzar** en el 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 a partir de l'observació de relacions entre les seves variables serveix com una bona il·lustració de la important contribució de l'atzar a la inferència estadística.

Si entenem la **inferència estadística** com el procés per mitjà del qual podem extreure conclusions generals a partir de l'anàlisi de les dades obtingudes d'una mostra, és necessari tenir present que aquest procés únicament és possible si la selecció dels participants o l'assignació als diferents grups han estat aleatòries. És a dir, només quan l'atzar intervé en almenys un d'aquests dos moments 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 associacions observades i decidir, amb un determinat nivell de confiança, fins a quin punt són realment el producte d'una relació causal existent al conjunt de la població. O el que és el mateix, disposar de les evidències suficients per tal de ser capaços d'excloure la possibilitat alternativa que els resultats obtinguts puguin ser, en realitat, explicats com a conseqüència d'una selecció i/o una assignació no aleatòries dels participants.

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

Tot i la importància del disseny de la investigació per tal d'extreure'n conclusions no esbiaixades que, a més, siguin generalitzables més enllà dels límits dels estudis particulars, el cert és que els investigadors no sempre poden utilitzar experiments per desenvolupar els seus treballs de camp. 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 en les disciplines basades en l'anàlisi de dades quantitatives i resulta especialment evident quan els estudis es desenvolupen, lluny de les condicions controlades dels laboratoris, en els contextos naturals en què és produeix l'activitat quotidiana de les persones.

Si, tal com plantejàvem a l'inici d'aquest text, l'objectiu és analitzar fenòmens complexos com l'eventual discriminació per raó de gènere en l'accés dels estudiants a una universitat, resulta obvi que no serà possible decidir el gènere dels candidats ni, de la mateixa manera, tampoc no es podrà escollir el departament al qual els candidats haurien de presentar-ne les 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 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 sustentar 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 determinada 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'elles en el sistema de relacions i, d'aquesta manera, descriure, explicar o predir els fenòmens que són objecte d'interès per a la investigació.

Aquest marc analític general, per tant, ofereix als investigadors l'oportunitat de portar a terme el **control estadístic** de qualsevol variable estranya que, com a eventual factor de confusió, pogués interferir en la relació entre les variables

dependents i independents que són objecte d'interès. 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 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 observacionals. Tal com ja hem explicat, l'única limitació es troba en el moment de la interpretació dels resultats i, especialment, en el risc que els investigadors estiguin disposats a assumir en el moment de 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 aquesta, 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 descriure 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 del càlcul d'una mesura de tendència central –com la mitjana, la mediana o la moda– i 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 fer una descripció de la mostra i, quan és possible, establir una inferència sobre la població a la qual representa.

Òbviament, quan els investigadors porten a terme els seus estudis mai no concentren tots els 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 variables de manera independent. 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 que, 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 del 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 en 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 a causa, en darrer terme, del 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, de manera independent. Així, 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 podrien no haver 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 entre diverses variables, però ho fa, en aquest cas, mitjançant la construcció de models complexos que permeten determinar-ne l'existència de manera simultània. Així, 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, els permet tenir en compte 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 d'interès.

Controlant estadísticament la contribució de totes aquestes variables al sistema de relacions, aquest tercer tipus d'anàlisi permet mantenir constants 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 d'aquests models permet determinar l'existència de múltiples relacions en 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 fins ara, l'estadística multivariant permetria conèixer la contribució simultània de les característiques dels estudiants i dels departaments als quals 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 pel que fa a 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 precisa sobre l'existència d'una discriminació per raó de gènere en l'accés dels estudiants a la universitat.

Resulta 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 únicament 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 l'anàlisi. D'altra banda, entenen l'estadística bivariant com l'estudi de les relacions entre parelles de variables que haurien estat obtingudes en investigacions observacionals, 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 en fan poc interessant la seva adopció. En primer lloc, 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 o observacional 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 en els seus estudis.

En segon lloc, aquest plantejament més restrictiu sobre l'anàlisi multivariant 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 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, és important tenir present que 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, i la raó per la qual resulten especialment útils per abordar problemes complexos, és la capacitat que tenen 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 fer això, en funció dels objectius de la 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 models i, d'aquesta manera, ser capaços de dur a terme una avaluació de la seva contribució específica o 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. D'una altra banda, 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 característiques i les condicions en què poden ser utilitzades i, de manera particular, servir de guia perquè els lectors interessats puguin 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 observacionals, 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 l'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 resulta 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 investigació, sinó que, 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 aquest pot ser un procés complex, especialment a les investigacions que es basen en l'avaluació d'atributs psicològics no directament observables (vegeu Meneses et al., 2014, per a una discussió més àmplia), la mesura no seria una altra cosa que l'establiment d'una correspondència entre les propietats dels fenòmens que són objecte d'interès i els nombres que les representen en una escala determinada. En aquest sentit, 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. Per a fer això, les variables qualitatives poden ser definides a partir de l'ús d'escala nominal i ordinal, 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 tots els individus que pertanyen a un determinat grup per oposició a tota la resta.

D'una 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 mitjançant tests 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 a la pràctica sovint es tracten 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 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 implicacions importants per al procés de mesura. En aquest sentit, els investigadors han d'escollir sempre les que millor reflecteixin la variabilitat observada en els fenòmens que són objecte d'interès i, per tant, aquelles que els permetin recollir adequadament la informació relativa a la presència o absència d'unes característiques determinades o, quan els seus estudis ho requereixen, el grau o la quantitat amb què aquestes característiques són presents en els participants.

Però, el que és més rellevant per a una introducció a l'anàlisi multivariant com la que ens hem proposat en aquest text és que la distinció entre variables qualitatives i quantitatives té també algunes implicacions importants per a la construcció de models complexos que permetin analitzar i interpretar 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 diferents nivells no són més que etiquetes numèriques arbitràries que serveixen per a identificar els diferents grups de participants, però en cap cas no reflecteixen el grau o la quantitat amb què una determinada característica és present en els individus. Si bé és cert que, de vegades, 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'una 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 tenint en compte, principalment, la pregunta o l'objectiu general que motiva el procés de construcció dels models. 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 pugui servir, en darrer terme, de guia per a orientar els investigadors en el moment d'escollir la tècnica que millor s'ajusti als seus objectius.

La diversitat de tècniques disponibles ens impedeix abordar-les totes, però aquesta classificació servirà per a presentar-ne algunes de les utilitzades més freqüentment. 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, així, 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 en 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 de les característiques de les dades

Objectiu general	Escenari d'aplicació	Característiques de les dades	Tècnica multivariant
Analitzar relacions d'interdependència per 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	Escalament multidimensional
Analitzar relacions de dependència per fer explicacions o prediccions	Explicació de la variabilitat dels individus	Una variable dependent quantitativa	Regressió múltiple
		Dos 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
		Dos 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 simultàniament	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

D'acord amb la classificació presentada en la taula 7, és possible establir tres grans grups de tècniques en funció de l'objectiu general al qual contribueix l'anàlisi i la interpretació del sistema de relacions mitjançant la construcció de models multivariants. A continuació, ens ocuparem de cadascun d'aquests tres grans objectius, identificarem els diferents escenaris d'aplicació i presentarem algunes de les alternatives utilitzades més freqüentment en funció de les característiques de les variables involucrades.

1) Quan no és possible distingir entre variables dependents i independents

En aquest cas, 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 existents 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 cas de l'anàlisi de components principals, 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, emergeixen com a agrupacions de les seves variables. En canvi, en el cas de l'anàlisi factorial, els investigadors parteixen d'una teoria sobre els fenòmens que són 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 tractar variables qualitatives, aquestes dues tècniques són generalment aplicades quan les variables analitzades són de naturalesa quantitativa.

En cas que les variables utilitzades siguin qualitatives, els investigadors tenen a l'abast una tècnica alternativa, l'**anàlisi de correspondències**, per tal d'assolir els mateixos objectius. Mitjançant la transformació de la informació qualita-

tiva 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 als 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 pot servir 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** o **anàlisi de clúster**.

Aquesta tècnica ofereix un conjunt de procediments que permeten 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 ells 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 investigació, la tècnica més adequada és l'**escalat multidimensional**. En aquest cas, a diferència del que succeeix amb la resta de tècniques d'anàlisi de les relacions d'interdependència que hem introduït fins ara, la cerca d'una estructura en 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 percepcions de similitud. Com succeeix en el cas de l'anàlisi de conglomerats, l'escalat multidimensional pot ser aplicat tant a variables de naturalesa quantitativa com qualitativa.

2) Quan és possible distingir entre variables dependents i independents

Els investigadors es mouen ara 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 serviren per a avaluar la contribució específica del canvi o la variació en els nivells de totes les variables independents considerades en la construcció dels models. A més, tot i que les variables independents considerades en aquests models solen ser quantitatives, totes dues tècniques són suficientment flexibles com 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 que permeten explicar. Quan els investigadors es proposen analitzar la variabilitat dels individus en una característica *i*, per tant, centren l'atenció en una única variable dependent de naturalesa quantitativa, 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 quantitatives 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 dur a terme 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 les quals els individus han estat assignats de manera aleatòria.

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 considerades en els models *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 de naturalesa quantitativa, 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 observacional, ja que permeten tenir en compte la influència de la variabilitat dels individus en altres característiques importants 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 als quals 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 involucrades, però difereixen, fonamentalment, tant en el nombre de nivells que la variable dependent qualitativa pot adoptar com 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, ho volen fer 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 hauria de ser aplicada quan segueixin una distribució normal.

En canvi, tot i que la regressió logística és aplicable únicament quan la variable dependent és dicotòmica, i per tant té només dos nivells, 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, incorporant aquestes darreres mitjançant la creació de les variables fictícies o *dummies* corresponents.

3) Quan el context no és exclusivament d'interdependència o de dependència, sinó que hi ha una combinació d'aquests dos tipus de relacions

En aquest cas, quan els investigadors estan interessats a analitzar les múltiples 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 exposades anteriorment, aquesta tècnica té com a objectiu general analitzar simultàniament 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 a partir de dos grans tipus de models.

D'una banda, d'acord amb la lògica de l'anàlisi 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. D'una altra banda, d'acord amb la lògica de l'anàlisi 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.

És important tenir present, però, que, malgrat que aquesta tècnica s'aplica generalment quan les variables considerades en la construcció d'aquests dos tipus de models són quantitatives, disposem d'alguns procediments que permeten tractar també variables de naturalesa qualitativa. 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 l'objectiu general que motiva la investigació i les característiques de les dades utilitzades per a oferir una resposta, acabarem aquesta introducció a l'anàlisi multivariant abordant el procés de construcció dels models estadístics complexos que, com hem dit, 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 text, és necessari recordar que l'anàlisi multivariant únicament adquireix tot el sentit en relació amb el procediment general establert en el context de la investigació quantitativa que, breument, podríem resumir de la manera següent:

- Formular una pregunta o un objectiu general que serveixi per a abordar un problema rellevant.
- Escollir el disseny de la investigació i especificar la mostra de participants.
- Definir adequadament totes les variables involucrades i especificar-ne les característiques.
- Desenvolupar o escollir els instruments necessaris per a portar a terme les mesures oportunes.
- Recollir les evidències necessàries que permetin respondre als objectius de la investigació.
- Resumir i tractar estadísticament les dades a fi d'avaluar les evidències obtingudes i, quan és possible, generalitzar les conclusions més enllà dels límits de l'estudi en particular.

D'acord amb aquest procediment, l'anàlisi multivariant proporciona el marc analític general que fa que els investigadors puguin descriure, explicar o predir els fenòmens que són objecte d'interès mitjançant el desenvolupament dels models estadístics més adequats que els permetin dur 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 per mitjà del qual els investigadors poden obtenir 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 cadascuna d'elles i, per tant, 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 múltiples relacions simultànies en els seus models i, seguint els procediments específics establerts en funció de la tècnica escollida, poden ser capaços de decidir fins a quin punt aquests models s'ajusten o són una bona representació de la realitat que es proposen analitzar. 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 bàsica, de manera que comparteixen uns principis generals i, com veurem a continuació, un conjunt de fases que estructurin el procés de construcció d'aquest tipus de models.

En aquest sentit, abans d'esbossar-ne les fases, resulta convenient que ens detinguem breument en els principis generals que regeixen l'anàlisi multivariant, entre els quals 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 a utilitzar-la 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 produeixin les relacions que són 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 requerirà la investigació i, en darrer terme, els permeti escollir la tècnica més adequada 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 necessària 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 seves dades. 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 a respondre als objectius. Aquesta exploració de les dades és especialment rellevant en el context

de l'anàlisi de les relacions de dependència que, com hem pogut veure, permet obtenir els indicis necessaris per a considerar el paper de qualsevol factor o variable de confusió que pugui interferir en les relacions que són objecte d'interès i, per tant, controlar-ne estadísticament la influència en els models multivariants a fi d'evitar que pugui esdevenir una explicació alternativa als resultats obtinguts.

- **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, permet comprovar fins a quin punt es compleixen els supòsits que assumeix la tècnica escollida per a modelar les múltiples relacions entre variables de manera simultània. En aquest sentit, tal com hem exposat en la classificació de les diferents tècniques disponibles, és important que els investigadors corroborin que tant la naturalesa de les relacions que es proposen analitzar com les característiques de les variables implicades s'ajusten als requeriments de la tècnica escollida. D'altra banda, més enllà dels requeriments estadístics particulars de cadascuna de les tècniques disponibles, 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 garanteix el compliment d'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 el moment d'especificar els 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 que, per tant, cal no deixar de tenir-ne en compte cap d'important–, com ho és evitar incloure qualsevol altra variable que, en realitat, no sigui rellevant per a analitzar els fenòmens que són objecte d'interès. És el que anomenem una *especificació adequada dels models*, que en cap cas es correspon amb una decisió única, sinó que forma part del procés contingut i iteratiu de construcció dels models multivariants per mitjà del qual els investigadors afegixen i treuen variables als seus models en funció dels resultats que els proporcionen. L'objectiu últim d'aquest procés és, a més, l'obtenció d'uns models que siguin parsimoniosos, és a dir, capaços de representar la màxima com-

plexitat dels fenòmens que són objecte d'interès 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 l'especificació correcta dels models multivariants i tenint en compte, per tant, que no poden ser mai utilitzats com una prova definitiva o conclouent 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 que acaben jugant 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, cal avaluar fins a quin punt la combinació de variables escollida s'ajusta raonablement bé a la variabilitat observada en 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, per tant, necessàriament incorrectes, però quan se centren en els aspectes substancials dels fenòmens es converteixen en una eina molt útil per a interpretar les múltiples relacions simultànies que són objecte d'interès per a la investigació.
- **El disseny de la investigació condiona la inferència estadística basada en l'anàlisi multivariant.** Tal com hem explicat, 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 desenvolupar 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 de relacions causals generalitzables a la població. 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 que han obtingut per a establir les seves inferències.

Finalment, 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 explicant al llarg d'aquest text i, 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 particulars amb què han de ser aplicades les diferents tècniques disponibles, aquestes deu fases fonamentals serveixen per a organitzar seqüencialment les diferents decisions que els investigadors han de prendre per a dur a terme una anàlisi complexa de les seves dades mitjançant la construcció de models multivariants. D'aquesta manera:

1) Delimitació del propòsit de l'anàlisi. La construcció de models multivariants comença sempre amb una definició precisa dels objectius particulars a partir dels 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, amb caràcter general, permeten als investigadors descriure, explicar o predir els fenòmens que són objecte del seu interès. 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 de què disposen. Aquesta és una decisió important en el procés de construcció de models multivariants, ja que la tècnica escollida condiona els procediments que els investigadors han de dur a terme al llarg de les fases següents.

3) Exploració inicial de les dades. Una vegada seleccionada la tècnica d'anàlisi multivariant més adequada, els investigadors han de familiaritzar-se amb la distribució de les variables involucrades i, a continuació, examinar-ne les relacions per parelles. Mitjançant l'aplicació de tècniques d'anàlisi univariant i bivariant, aquest primer contacte amb les dades 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 ana-

litzar el sistema de relacions que els investigadors es proposen abordar. D'una 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 donen 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'elles 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 que són objecte d'interès. Tenint en compte aquestes evidències, els investigadors hauran de portar a terme els 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 siguin parsimoniosos, de manera que els investigadors han de ser capaços de fer un balanç adequat entre l'increment del nivell de complexitat dels models i els beneficis que això comportaria en relació amb la millora substantiva en el seu ajustament a les dades.

8) Interpretació del sistema de relacions. Una vegada assolit un ajustament global acceptable, arriba el moment en què els investigadors poden utilitzar els seus models multivariants per a analitzar i interpretar 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'elles 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 en la mostra i, el que és encara més important, la significació que aquestes relacions tenen en 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 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 una publicació científica que serveix per a comunicar les principals conclusions a les quals 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 per part dels investigadors per a transmetre i fer accessibles els resultats obtinguts. 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.

8. Bibliografia anotada

Finalment, entenent que per raons d'espai no és possible abastar el detall de l'execució i la interpretació dels resultats que proporcionen les diferents tècniques disponibles en una introducció general, ni fer-ho amb tot el detall i la profunditat que poden dedicar-hi altres contribucions més avançades, els lectors interessats poden trobar diversos manuals sobre anàlisi multivariant que els poden ajudar a complementar i ampliar aquesta aproximació.

En aquest sentit, disposem d'alguns manuals publicats en el nostre entorn més proper que seria interessant tenir en compte com, per exemple, els de Martínez (1999), Díaz (2002), Peña (2002), Catena, Ramos i Trujillo (2003), Pérez (2004), Ximénez i San Martín (2004) o, més recentment, de la Garza, Morales i González (2013), López-Roldán i Fachelli (2015) o Aldas i Uriel (2017).

De manera complementària, a continuació oferim una selecció d'algunes contribucions rellevants desenvolupades en el context internacional que poden ser molt útils per a adquirir una visió més àmplia sobre la lògica general de l'anàlisi multivariant, les particularitats de les diferents tècniques, el seus fonaments matemàtics o la utilització del programari estadístic especialitzat disponible:

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

Com a introducció general, aquest llibre és una aproximació excel·lent als aspectes bàsics de l'anàlisi multivariant de les dades. Prenent com a punt de partida una primera part que, a mode de visió de conjunt, presenta les qüestions transversals que comparteixen les diferents tècniques, l'autora dedica els capítols següents a exposar-les una per una remarcant-ne les diferències i similituds, els objectius particulars que permeten assolir, les seves assumpcions bàsiques i la manera com han de ser interpretats els resultats que proporcionen. Reduint al mínim la presència de les qüestions matemàtiques relacionades amb la computació estadística dels models multivariants, aquest manual proposa una aproximació conceptual molt assequible que, a més, és il·lustrada amb diferents exemples pràctics desenvolupats mitjançant la utilització de SPSS i SAS.

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

Tot i que aquest llibre se centra principalment en el desenvolupament de les habilitats necessàries per a portar a terme una presentació efectiva dels resultats proporcionats per les diferents tècniques, la seva lectura és també molt re-

comanable per als lectors interessats a familiaritzar-se amb algunes qüestions bàsiques que van més enllà de la computació estadística dels models multivariants. En aquest sentit, adopta com a principi la necessitat de fer accessibles les interpretacions dels investigadors sobre les relacions entre les diverses variables involucrades en els seus estudis, de manera que, en darrer terme, puguin contribuir efectivament a enriquir el debat públic sobre els fenòmens que són objecte del seu interès adaptant-se a les expectatives i les necessitats dels diferents tipus d'audiències.

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

Aquesta és una referència obligada per als lectors interessats a aprofundir en el procés de construcció de models multivariants. Com que es tracta d'un manual que ha estat revisat i actualitzat regularment durant les darreres tres dècades, els autors han arribat al punt d'equilibri necessari que permet combinar, de manera senzilla i accessible per a un públic que no té una formació estadística avançada, els principals reptes vinculats a l'execució de cadascuna de les tècniques d'anàlisi multivariant i la seva aplicabilitat en la pràctica, tant en contextos acadèmics com professionals. Tot i que els exemples que utilitza per a il·lustrar les tècniques són propis de la investigació desenvolupada en l'àrea dels estudis de màrqueting, totes les orientacions, els consells i les recomanacions pràctiques resulten igualment interessants per a la resta de disciplines basades en l'anàlisi de dades quantitatives.

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

L'altra gran referència en el camp dels manuals que es proposen oferir una introducció general a les diferents tècniques d'anàlisi multivariant disponibles ha complert també tres dècades de revisió i actualització, que han fet que es converteixi en un dels llibres de capçalera per a molts dels professionals i investigadors que treballen amb dades quantitatives. Amb una orientació eminentment pràctica, aquest text centra l'atenció en una discussió profunda dels tipus de preguntes o objectius que les diferents tècniques disponibles permeten respondre, les seves limitacions i totes aquelles qüestions importants que és necessari tenir en compte en la seva aplicació a les dades obtingudes en la investigació, la manera com fer-ho mitjançant el programari estadístic especialitzat que il·lustra amb exemples per a SPSS, SAS i SYSTAT, i les estratègies més convenients per a garantir una anàlisi rigorosa i honesta per part dels investigadors.

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

La darrera actualització d'aquest manual ens ofereix una tercera aproximació que, amb caràcter general, s'ocupa de les particularitats de les diferents tècniques d'anàlisi multivariant. Tot i que dedica una part dels seus esforços a introduir els lectors a l'àlgebra matricial en què es basa la combinació lineal de variables, el cert és que l'exposició que fa en la resta de capítols sobre les tècniques, els seus supòsits i les condicions en què poden ser aplicades no obliga a aprofundir en els formalismes matemàtics que, molt probablement, podrien allunyar els lectors que no tenen una formació estadística avançada. En aquest sentit, tot i que alguns dels exercicis que proposa es basen en el càlcul manual, també ofereix una bona il·lustració dels diferents capítols amb exemples pràctics per a SPSS i SAS.

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

Amb un caràcter marcadament accessible, aquest manual té com a objectiu presentar els conceptes més importants deixant a banda en la mesura dels possibles tota la formulació matemàtica. Plantejat com un text introductori, aconsegueix de manera notable posar per davant la racionalitat de l'anàlisi multivariant amb el supòsit que, atès el nivell de sofisticació del programari especialitzat disponible per a executar l'anàlisi pròpiament dita, el que és realment important és que els lectors interessats en l'ús de les tècniques disponibles dominin la manera com s'implementen i interpreten els resultats que proporcionen. No obstant això, cada capítol proporciona un apartat específic orientat a l'exploració de les dades i una revisió dels supòsits que és necessari comprovar, així com també inclou una petita panoràmica matemàtica que permet conèixer els aspectes més tècnics vinculats a cada tècnica. Finalment, en l'annex disposa també d'una introducció a l'àlgebra matricial que sustenta tots els càlculs.

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

Aquest és un bon exemple dels manuals que remarquen la utilització del programari estadístic especialitzat que permet dur a terme l'anàlisi multivariant. En aquest sentit, el text adopta una aproximació eminentment pràctica a l'anàlisi multivariant posant en el centre de la seva atenció els procediments disponibles en el marc del programari estadístic de codi obert R. En qualsevol cas, aquest és un manual assequible per als lectors que no tenen uns coneixements avançats en l'ús de R i, a més, proporciona exemples abundants juntament amb tot el codi necessari per a portar a terme amb èxit els exercicis que, de manera pràctica, complementen totes les seves explicacions.

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

La tercera edició d'aquest llibre clàssic, publicat per primer cop a finals dels anys cinquanta com a compilació de materials docents, és probablement la referència d'elecció per als lectors interessats a aprofundir en els aspectes vinculats al tractament matemàtic de les dades en què es basen les diferents tècniques d'anàlisi multivariant. Aquest és, per tant, un text d'un nivell avançat que presenta, de manera estructurada, la teoria, els mètodes i les demostracions matemàtiques necessàries per a abordar rigorosament els diferents procediments estadístics derivats de la distribució normal multivariant i les seves característiques. Tot i que exigeix un cert domini de l'àlgebra matricial per a seguir-ne les explicacions, incorpora un annex que serveix com a introducció bàsica a les definicions i els teoremes que permeten desenvolupar els càlculs de manera autònoma.

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

En un sentit similar, aquest llibre pertany a la categoria dels manuals que prioritzen els aspectes matemàtics de l'anàlisi multivariant i, per tant, està també dirigit als lectors que tenen uns coneixements estadístics avançats. Tot i que, a diferència de l'anterior, introdueix de manera natural els aspectes bàsics relacionats amb l'àlgebra matricial, el seu objectiu és proporcionar les demostracions matemàtiques que permeten entendre el funcionament de les diferents tècniques, les seves possibilitats i limitacions, així com també els resultats que proporciona cada una. Per a fer això, els autors presenten diferents exercicis que serveixen per a il·lustrar-ne les explicacions, alguns basats en exemples ficticis simplificats per tal que sigui possible resoldre'ls a mà, i altres de més complexos a partir de dades reals que exigeixen la utilització d'algun programari estadístic especialitzat.

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

Finalment, aquest text adopta una perspectiva innovadora per a abordar les diferents tècniques d'anàlisi multivariant. Posant la representació gràfica en el centre de la discussió, els autors desenvolupen una aproximació que barreja els aspectes matemàtics essencials implicats en la computació dels models multivariants amb una exposició detallada de l'ús de SAS, STATA i SPSS per a fer-ho. Més enllà de la preceptiva introducció a l'àlgebra matricial que caracteritza aquest tipus de manuals avançats, cada capítol ve acompanyat d'un exemple real extret d'una publicació científica que permet contextualitzar els diferents tipus d'anàlisi i, en darrer terme, posar en evidència com els gràfics són aliats poderosos que permeten representar adequadament el significat de les dades també en el context de l'anàlisi multivariant.

Bibliografia

- Aldas, J., i Uriel, E. (2017). *Análisis multivariante aplicado con R* (2a ed.). Madrid: Paraninfo.
- Aldrich, A. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 10(4), 364-376.
- Anderson, T. W. (2003). *An introduction to multivariate statistic analysis* (3a ed.). Nova Jersey: Wiley-Interscience.
- Bickel, P. J., Hammel, E. A., i O'Connell, J. W. (1975). Sexbias in graduate admissions: Data from Berkeley. *Science*, 187, 398-404.
- Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338), 364-366.
- Brown, B. L., Hendrix, S. B., Hedges, D. W., i Smith, T. B. (2011). *Multivariate analysis for the biobehavioral and social sciences. A graphical approach*. Hoboken: John Wiley & Sons.
- Catena, A., Ramos, M. M., i Trujillo, H. M. (2003). *Análisis multivariado. Un manual para investigadores*. Madrid: Biblioteca Nueva.
- Díaz, V. (2002). *Técnicas de análisis multivariante para investigación social y comercial. Ejemplos prácticos utilizando SPSS versión 11*. Madrid: Ra-Ma.
- Coolican, H. (2014). *Research methods and statistics in psychology* (6a ed.). Londres: Psychology Press.
- Cozby, P. C., i Bates, S. C. (2015). *Methods in behavioral research* (12a ed.). Nova York: McGraw-Hill.
- David, H. A., i Edwards, A. W. F. (2001). *Annotated readings in the history of statistics*. Nova York: Springer.
- Everitt, B., i Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Nova York: Springer.
- Freedman, D., Pisani, R., i Purves, R. (2007). *Statistics* (4a ed.). Nova York: W. W. Norton & Company.
- de la Garza, J., Morales, B. N., i González, B. A. (2013). *Análisis estadístico multivariante*. México: McGraw-Hill.
- Hahs-Vaughn, D. L. (2017). *Applied multivariate statistical concepts*. Nova York: Routledge.
- Hair, J. F., Black, W. C., Babin, B. J., i Anderson, R. E. (2013). *Multivariate data analysis. Pearson new international edition* (7a ed.). Harlow: Pearson.
- Harlow, L. L. (2014). *The essence of multivariate thinking. Basic themes and methods* (2a ed.). Nova York: Routledge.
- Johnson, R., i Wichern, D. (2019). *Applied multivariate statistic analysis* (6a ed.). Nova Jersey: Pearson.
- Light, R. J., Singer, J. D., i Willett, J. B. (1990). *Bydesign. Planning research on higher education*. Cambridge: Harvard University Press.
- López-Roldán, P., i Fachelli, S. (2015). Metodología de la investigación social cuantitativa. Barcelona: Universitat Autònoma de Barcelona. Disponible a: <https://ddd.uab.cat/record/129382>.
- Meneses, J., Barrios, M., Bonillo, A., Cosculluela, A., Lozano, L. M., Turbany, J., i Valero, S. (2014). *Psicometría*. Barcelona: Editorial UOC.
- Martínez, R. (1999). *El análisis multivariante en la investigación científica*. Madrid: La Muralla.
- Miller, J. E. (2013). *The Chicago guide to writing about multivariate analysis* (2a ed.). Chicago: The University of Chicago Press.
- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Nova York: Cambridge University Press.

- Peña, D. (2002). *Análisis de datos multivariantes*. Madrid: McGraw Hill.
- Pérez, C. (2004). *Técnicas de análisis multivariante de datos. Aplicaciones con SPSS*. Madrid: Pearson.
- Pituch, K. A., i Stevens, J. P. (2016). *Applied multivariate statistics for the social sciences*. (6a ed.). Nova York: Routledge.
- Russo, F. (2009). *Causality and causal modelling in the social sciences*. Nova York: Springer.
- Shadish, W. R., Cook, T. D., i Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2a ed.). Boston: Houghton Mifflin.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13(2), 238-241.
- Tabachnick, B. G., i Fidell, L. S. (2019). *Using multivariate statistics* (7a ed.). Nova York: Pearson.
- Ximénez, M. C., i San Martín, R. (2004). *Fundamentos de las técnicas multivariantes*. Madrid: UNED.
- Yule, G. U. (1903). Notes on the theory of association of attributes of statistics. *Biometrika*, 2(2), 121-134.