

L'anàlisi quantitativa de dades

Antoni Coscolluela Mas
Albert Fornieles Deu
Jaume Turbany Oset

PID_00154056



Universitat Oberta
de Catalunya

www.uoc.edu

Índex

1. Mètode científic	5
1.1. El model general d'investigació	6
1.1.1. Nivell teóricoconceptual del model general d'investigació	6
1.1.2. Nivell tecnicometodològic del model general d'investigació	8
1.1.3. Nivell estadístico-analític del model general d'investigació	8
1.2. Tipus de dissenys d'investigació	8
1.2.1. Dissenys d'investigació transversals	9
1.2.2. Dissenys d'investigació longitudinals	9
1.2.3. Classificació de les metodologies d'investigació	9
1.3. Les variables	11
1.3.1. Classificació de les variables des del punt de vista metodològic	12
1.3.2. Classificació de les variables des del punt de vista estadístic	14
1.3.3. Transformació de les dades	15
1.3.4. Tipus d'escala de mesura	15
1.4. Població i mostra	18
1.4.1. Població	18
1.4.2. Mostra	19
1.4.3. El mostratge	20
1.4.4. Les tècniques de mostratge	20
1.4.5. La grandària de la mostra	24
1.5. Qualitat de les investigacions	25
1.6. Presentació d'informes científics	27
2. Anàlisi de les dades	30
2.1. Descripció de les variables	30
2.1.1. Índex de tendència central de la distribució d'una variable	31
2.1.2. Índex de dispersió o variabilitat de la distribució d'una variable	33
2.1.3. Representacions gràfiques	35
2.2. Relacions entre variables	37
2.2.1. Les proves de significació o d'hipòtesi	37
2.2.2. Les proves de relació o independència	40
2.3. Aproximació a l'anàlisi multivariable	56
Bibliografia	61

1. Mètode científic

Podem definir la **ciència** com l'activitat humana que té com a propòsit la recerca d'aproximacions als fenòmens observables per tal de descriure'ls i veure de quina manera es relacionen entre ells. En aquest sentit, estem d'acord amb Sierra Bravo (1985) quan diu que la investigació és un conjunt d'accions orientades a la **conceptualització de la realitat**. Els seus objectius són tres: **descriure, explicar i predir**. La descripció respon a allò que passa, l'explicació al perquè i la predicció tracta de fer deduccions dels fets o fenòmens per tal de poder pronosticar el seu comportament futur (què, com o quan passarà).

Com assenyala De Cabo (1996), el mètode científic intenta proporcionar una imatge imparcial del món, sense tenir en compte qui l'està observant (objectivitat). Però, com ja va remarcar Heisenberg el 1927 al seu principi d'indeterminació, el simple fet d'observar qualsevol objecte el modifica o pertorba. Com que aquest principi el va formular per a la ciència física, és fàcil deduir que els subjectes humans encara modifiquen més el seu comportament que pertorben els objectes d'estudi de la física quan se saben (o se senten) observats. Aquestes modificacions en el comportament reben el nom de *reactivitat* i constitueix un biaix en moltes de les investigacions fetes en éssers vius.

Reactivitat

La reactivitat es defineix com els canvis produïts en el comportament dels subjectes quan saben que estan essent observats.

Així, podem definir el mètode científic com el cicle complet de la investigació que ens assenyala un camí que ens porta en una direcció, cap a una finalitat. En definitiva, el mètode entès com una sèrie de procediments generals, s'especifica en unes estratègies concretes que denominem **dissenys d'investigació**, que no són altra cosa que els plans estructurats d'acció (una mena de "recepta") que ens indiquen quins passos hem de seguir per tal d'assolir una sèrie de fites.

Com proposa Arnau (1996b), el disseny inclou aspectes que van des de la formulació de problemes, la mesura de les variables i l'obtenció de les dades; fins a la seva anàlisi, la interpretació dels resultats i les conclusions que en podem extreure.

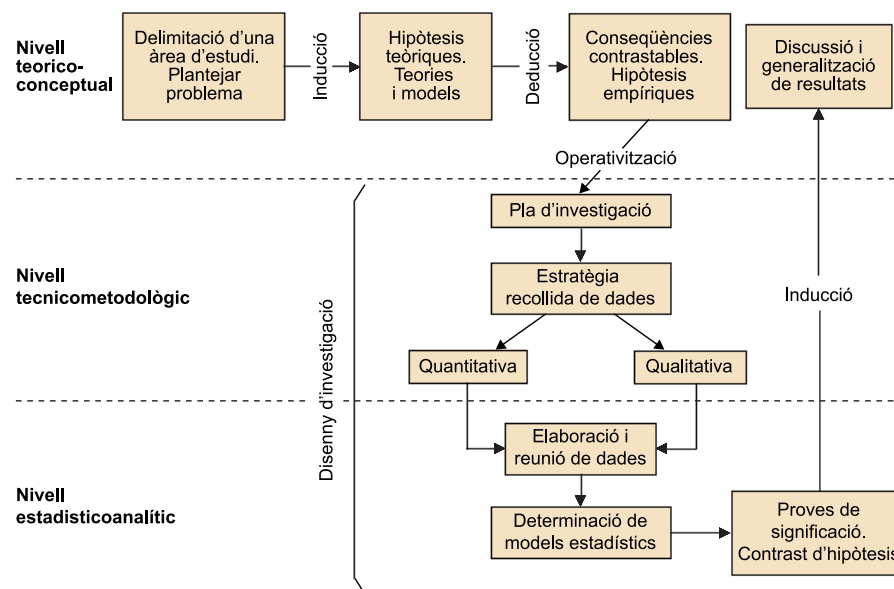
1.1. El model general d'investigació

El mètode científic es diferencia d'altres concepcions del terme *mètode* que ha de ser objectiu, sistemàtic, amb mecanismes de control sobre les expectatives de l'investigador i replicable. L'**objectivitat** implica que els resultats no poden dependre de qui fa la investigació. **Sistemàtic** significa que sempre procedeix de la mateixa manera. Els **mecanismes de control sobre les expectatives de l'investigador** són necessaris per tal de garantir que els resultats obtinguts siguin atribuïbles als factors que hem trobat en la recerca i no a la interpretació de qui la porta a terme. Finalment, perquè sigui **replicable** cal que l'informe de la investigació aporti informació suficient perquè altres investigadors la puguin repetir o replicar.

Seguint els criteris proposats per Arnau (1996a), el model general d'investigació s'articula en tres nivells d'actuació jerarquitzats, tal com mostra la figura següent:

- 1) Nivell teóricoconceptual.
- 2) Nivell tecnicometodològic.
- 3) Nivell analíticoestadístico.

Model general d'investigació en psicologia



Font: Arnau (1996). *Quantificació de les observacions: escales de mesura*.

1.1.1. Nivell teóricoconceptual del model general d'investigació

En el nivell teóricoconceptual del model general d'investigació s'elabora el marc teòric de la investigació. És a dir, és el moment de redactar el **problema**, entès com el punt de partida de la investigació. El camí per a especificar el problema és mitjançant la formulació de les **hipòtesis**; enteses com a soluci-

ons anticipades del problema. El problema científic és el punt de partida de la investigació. Es planteja generalment en forma de pregunta sobre el com, el quan o el perquè té lloc un fenomen. Els seus termes han de ser de naturalesa empírica (fenòmens observables) i han de tenir solució amb els mitjans de què disposa la ciència.

Aquest nivell constitueix el capítol d'**introducció** de l'informe científic. Durant aquesta fase és imprescindible dur a terme una tasca de **documentació** (recollida de la informació publicada prèviament sobre el tema d'estudi) que ens permeti conèixer en quin estat es troba el coneixement sobre el tema, per tal d'elaborar el marc teòric o conceptual.

Exemples de tipus de problemes d'investigació

El tipus de problema d'investigació que es vol estudiar determina en bona part la resta de fases de la recerca. Meltzoff (1998-2000) distingeix els següents tipus principals de preguntes d'investigació:

- **Preguntes d'existència.** Es qüestiona l'existència d'alguna cosa. Per exemple, fa anys que es qüestiona l'existència de l'inconscient, o de la comunicació simbòlica en ximpanzés. Actualment, es debat sobre la presència en els humans de períodes sensibles per a aprendre determinades habilitats, com ara els idiomes.
- **Preguntes de descripció i classificació.** Un cop establerta la seva existència, les següents preguntes que es plantegen són sobre la seva descripció i classificació. Continuant amb l'exemple de la comunicació simbòlica entre ximpanzés, ens plantejarem com és aquesta comunicació, quines són les seves característiques, si varia o és constant, si és un tipus de comunicació única o pertany a alguna classe coneguda, etc.
- **Preguntes de relació.** Una de les preguntes més freqüents en qualsevol disciplina és si existeix relació entre entitats. Es poden formular equivalentment com a preguntes de relació, per exemple: la memòria es relaciona amb l'edat?; o com a preguntes comparatives, per exemple: els joves tenen més memòria que la gent gran?
- **Preguntes de causalitat.** Plantegen la influència o l'efecte d'una variable independent (causa) sobre una variable dependent (efecte). Per exemple: el tractament conductual provoca una remissió de l'anorèxia? Aquestes preguntes requereixen investigació experimental, en la qual l'investigador fa servir l'aleatorització i manipula la variable independent, o investigació quasi experimental, que fa servir altres estratègies de control de les variables estranyes. Quan la pregunta és de tipus causal, tractarem de redactar la hipòtesis en termes implicatius. És a dir, "Si es manipula la variable independent de determinada manera, **aleshores** cal esperar que es produeixin els següents canvis en la variable dependent".

Exemples d'hipòtesis

Alguns exemples d'hipòtesis són els següents:

1) Hipòtesis descriptives

- Quin és el percentatge de llars a Catalunya que disposen de connexió a Internet?
- Què opinen els catalans sobre la Llei de normalització lingüística?
- Els estudiants estan satisfets amb els materials d'estudi proporcionats?
- Quina és la mitjana salarial d'un determinat nivell professional?

2) Hipòtesis associatives

- Hi ha diferències en la intenció de vot entre els votants que viuen en hàbitats rurals, semiurbans i urbans?
- Hi ha relació entre el sexe, l'edat i el consum d'alcohol els caps de setmana?
- La taxa d'atur varia segons el nivell de formació?
- Els homes utilitzen Internet amb més freqüència que les dones?
- Si augmenta el consum d'alcohol, **aleshores** s'incrementarà l'accidentalitat?

Vegeu també

El subapartat 1.6 d'aquest mòdul tracta sobre l'informe científic.

Vegeu també

Podeu consultar la unitat de vídeo 16 per a complementar la informació d'aquest subapartat.

3) Hipòtesis predictives

a) Si els alumnes aprenen amb ajuda de l'ordinador, aleshores milloren el seu rendiment acadèmic mesurat per les notes dels exàmens?

b) A partir de les dades recollides en una enquesta salarial, quin és el salari de mercat d'un director de recursos humans amb unes característiques determinades?

1.1.2. Nivell tecnicometodològic del model general d'investigació

El nivell tecnicometodològic del model general d'investigació és la fase de planificació de la recerca; correspon a l'estratègia o **disseny d'investigació**.

Entenem com a *disseny d'investigació* el pla que permet donar solució al problema, verificant les hipòtesis plantejades en connectar la teoria i la realitat (món empíric).

Vegeu també

En el subapartat 1.2 hi ha diferents classificacions dels dissenys d'investigació.

En l'informe científic, el nivell tecnicometodològic es materialitza en l'apartat "**Mètode**" i el subapartat "**Materials, participants i procediment**". En el cas de dissenys d'investigació molt complexos, es pot incloure un subapartat "Disseny i variables".

Per tant, en aquesta fase es trien les variables que volem enregistrar i **s'operativitzen**. Es posen a punt els instruments de mesura i recollida de dades (materials), es decideix quins participants formaran part de la investigació (participants), i per quines fases passaran aquests subjectes (procediment).

1.1.3. Nivell estadístico-analític del model general d'investigació

El nivell estadístico-analític del model general d'investigació és la fase d'obtenció de dades per a la seva posterior **anàlisi estadística** per tal de verificar les hipòtesis. La interpretació dels resultats es concreta en els apartats "**Resultats**" i "**Discussió de l'informe**". D'altra banda, cal assenyalar que la discussió connecta amb el primer nivell "(teóricoconceptual) del model general d'investigació", ja que es verifica el grau de compliment dels supòsits formulats en les hipòtesis, tancant d'aquesta manera el cicle de la investigació.

1.2. Tipus de dissenys d'investigació

En aquest subapartat ens centrarem en uns dels aspectes més importants a l'hora de classificar els dissenys: la seva **temporalitat** –per exemple: Arnau (1996b) i León i Montero (2003, 2009). Així doncs, seguint aquest criteri, podem dir que els dissenys d'investigació poden ser transversals o longitudinals.

1.2.1. Dissenys d'investigació transversals

Els dissenys transversals són aquells que es fan amb la intenció de descriure la població en un moment determinat.

Els dissenys transversals també permeten trobar diferències entre grups i relacions entre variables. Aquesta estratègia temporal implica que únicament es portarà a terme una mesura per unitat d'anàlisi (per exemple: s'enquestarà el subjecte un sol cop). Com veurem a l'apartat 2, dedicat a l'anàlisi de dades, els índex que més s'utilitzen són els estadístics descriptius (mitjana, percentatge...) per a descriure les dades; els quocients de correlació; el khi quadrat (χ^2) per tal d'establir relacions entre variables; i les proves de diferències de mitjanes entre grups.

Vegeu també

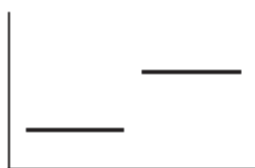
En l'apartat 2 d'aquest mòdul hi ha una breu explicació dels índex estadístics més habituals.

1.2.2. Dissenys d'investigació longitudinals

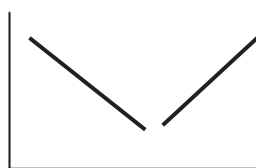
Al llarg del temps, en els dissenys longitudinals es prenen diverses mesures, ja sigui als mateixos subjectes, o a subjectes diferents.

La intenció d'aquests estudis és estudiar el canvi o l'evolució, avaluant els nivells o les tendències, per la qual cosa es fan comparacions entre períodes diferents. Per tant, el més rellevant d'aquests tipus de dissenys és la incorporació de la dimensió temporal.

Exemples de dissenys longitudinals



Exemple: canvi de nivell



Exemple: canvi de tendència

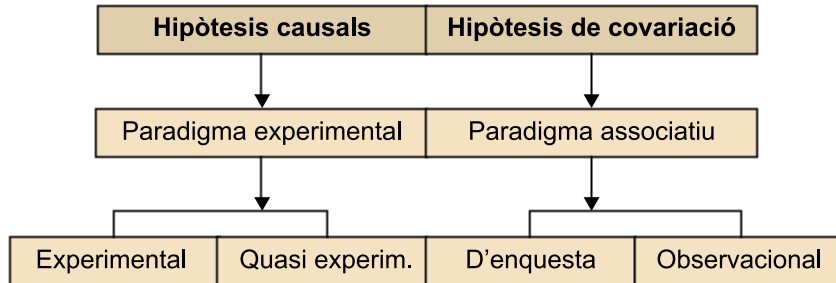
Els dissenys longitudinals estudien canvis de nivell o de tendència al llarg del temps.

1.2.3. Classificació de les metodologies d'investigació

El **mètode científic**, malgrat ser únic i general per a totes les branques de la ciència, es pot manifestar en diferents estratègies que, en darre- ra instància, es poden classificar en **metodologies quantitatives** (manipulativa experimental) i **qualitatives** (correlacional, associativa dife- rencial), definides en funció del grau de control de la investigació (Ar- nau, 1996b).

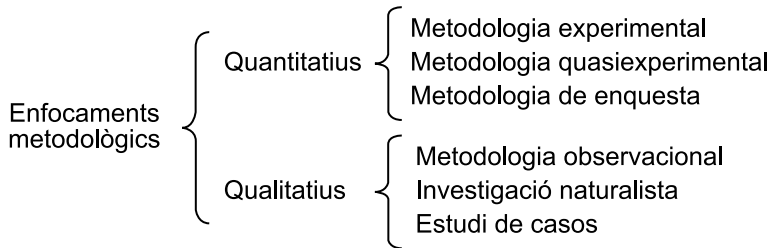
A continuació, podem veure esquemàticament alguns aspectes de les metodologies d'investigació en funció dels tipus d'hipòtesis, la naturalesa de la informació, el grau de control (validesa interna), el tipus d'intervenció, etc.

Segons el tipus d'hipòtesis



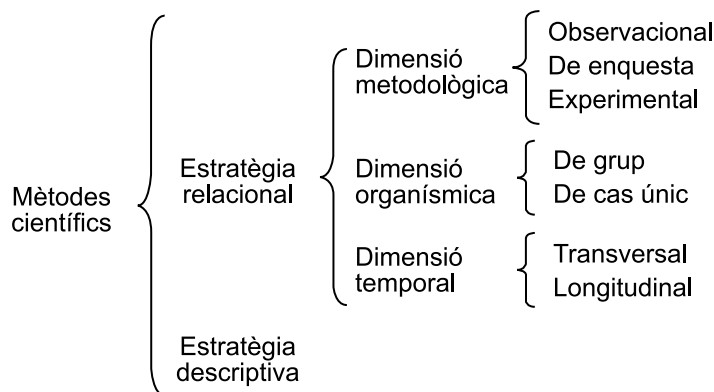
Font: J. Arnau (1996). *Quantificació de les observacions: escales de mesura.*

Segons la naturalesa de la informació



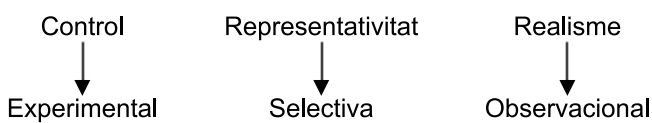
Font: J. Arnau (1996). *Quantificació de les observacions: escales de mesura.*

Diversos criteris de classificació



Font: Ato (1991)

Segons l'objectiu principal



Font: Kish (1987)

Segons el grau de validesa interna



Font: Anguera (1990)

Segons el tipus d'intervenció

		Intervenció en context	
		Sí	No
Intervenció en variable independent	Sí	Experiment de laboratori	Experiment de camp
	No	Selectiva	Observacional

Font: Moreno, Martínez i Chacón (2000)

Classificació adoptada			
Experimental	Quasi experimental	Selectiva	Observacional
<ul style="list-style-type: none"> Hipòtesis causals. Aleatorització. Manipulació de variables independents. Control experimental. Màxima validesa interna. Mínima validesa externa. 	<ul style="list-style-type: none"> Hipòtesis causals. Grups naturals. Manipulació de variables independents. Control experimental i estadístic. Alta validesa interna. Baixa validesa externa. 	<ul style="list-style-type: none"> Hipòtesis correlacionals i estudis exploratoris. Importància mostra. Grups representatius. Importància instruments de mesura. Control estadístic. Baixa validesa interna. Alta validesa externa. 	<ul style="list-style-type: none"> Hipòtesis correlacionals i estudis exploratoris. Registre en situacions naturals. Importància entrenament observadors. Mínima validesa interna. Màxima validesa externa.

L'elecció de la metodologia d'investigació no està determinada només pel tipus de problema d'investigació sinó també per les limitacions ètiques, econòmiques, humanes, etc.

1.3. Les variables

A l'hora de dur a terme una investigació cal determinar els aspectes de la realitat que es volen estudiar.

Podem definir una **variable** com qualsevol dimensió de canvi susceptible de prendre més d'un valor (contràriament a la constant) que expressa o representa un tret o propietat mesurable d'una realitat (per exemple: pes, alçada, quocient intel·lectual, grau de satisfacció, sexe...).

Per tal d'utilitzar les variables en una investigació, cal **operativitzar**, és a dir, que estigui perfectament especificat què i com s'ha de mesurar la variable (instruments, escala de mesura...).

Les variables s'acostumen a classificar segons dos punts de vista:

- 1) el metodològic, que indica el rol que juga la variable dins de la investigació, i
- 2) l'estadístic, que depèn del tipus de metrització amb què la variable ha estat mesurada.

1.3.1. Classificació de les variables des del punt de vista metodològic

Des d'un punt de vista metodològic, les variables s'acostumen a classificar segons el paper que exerceixen en el context de la investigació en variables independents o predictores, variables dependents o criteri, i variables estranyes.

La **variable independent** o **predictor** també denomina *variable de tractament*, *factor* o *predictora*, segons si es tracta d'una investigació experimental o no experimental. Parlarem de variables independents sempre que la investigació tingui com a objectiu l'obtenció d'una relació causal; com en el cas dels dissenys experimentals, on aquestes variables són les que l'investigador manipula amb l'objectiu de comprovar el seu efecte sobre alguns aspectes de la conducta (variables dependents). Una investigació pot tenir una sola variable independent (simple o unifactorial) o més d'una (disseny factorial).

Es pot distingir entre variables independents de manipulació directa o de selecció de valors. Les primeres són les purament experimentals. En aquest cas, l'experimentador les manipula triant els valors que pensa que li interessin per a la seva investigació (anomenats *tractaments*). Al contrari, en les de selecció de valors, l'investigador es limita a triar els subjectes que, per les seves característiques, compleixen les condicions requerides per a formar part d'un determinat grup experimental (per exemple: el sexe, el nivell d'intel·ligència, el nivell d'ansietat, l'extraversió, tenir un determinat rendiment acadèmic, l'edat...).

Variables independents de manipulació directa

Alguns exemples de variables independents de manipulació directa els tenim quan apliquem un mètode d'ensenyament interactiu o un de tradicional, la dosi d'un fàrmac, en fer una teràpia o una altra, en rebre uns estímuls o uns altres...

Variables independents de selecció de valors

Si volem trobar diferències en l'opinió sobre el masclisme d'homes i de dones, triarem les respostes donades al qüestionari pels enquestats per tal de comparar-les.

La **variable dependent** o **criteri** –o també **resposta**– tracta de l'aspecte sobre el qual s'espera observar els canvis produïts per la manipulació de la variable independent (per exemple, com un determinat mètode d'ensenyament influeix sobre el rendiment o sobre el fracàs escolar) o en un context no experimental sobre el qual es pretén obtenir informació (les opinions, les actituds,

Vegeu també

Podeu consultar les unitats de vídeo 15, 16 i 18 per a complementar la informació d'aquest subapartat.

la intenció de vot...) en funció de les variables predictores. Per tant, s'espera que sigui sensible a les variacions de les variables predictores. En definitiva, és l'aspecte que es mesura en la investigació.

Les **variables estranyes** o **de confusió**, igual que les variables independents o predictores, també tenen un efecte potencial sobre les variables dependents o criteri. Per tant, les variables estranyes s'han de controlar per tal d'eliminar o neutralitzar el seu possible efecte contaminant sobre les variables criteri. El grau de control és màxim en la metodologia experimental, escàs en la d'enquestes i molt limitat en l'observació natural. La manca de control de les variables estranyes en les metodologies no experimentals impossibilita establir relacions causals.

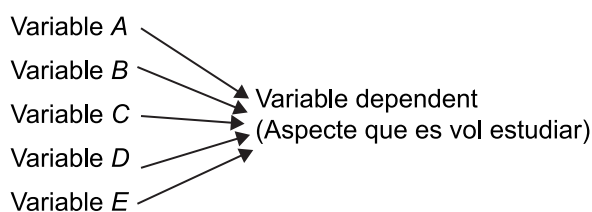
Podem definir la **relació causal** en funció de tres criteris (Kenny, 1979, citat per Domènech, 1995). Per a considerar una relació entre dues variables com a causal cal que es compleixin els tres requisits següents:

- 1) **Associació.** Les dues variables han de covariar (els canvis en els valors d'una s'han de reflectir en els valors de l'altra).
- 2) **Temporalitat.** La causa (variable independent) ha de ser prèvia a l'efecte (variable dependent).
- 3) **Absència d'esperitat.** La variable independent (causa) ha de ser l'única explicació dels canvis apreciats en la variable dependent (efecte).

Exemple del rol que exerceix les variables dins la investigació

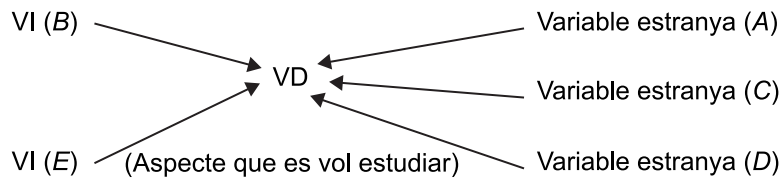
Si fem una enquesta a les empreses per tal de determinar quin ha de ser el salari dels treballadors d'un determinat nivell professional (salari de mercat), el sou serà la variable criteri. D'altra banda, les variables predictores podran ser determinats aspectes de les empreses (grandària de l'empresa, sector d'activitat, zona geogràfica en què es troba...) i de l'individu (formació, experiència, responsabilitat...). Ara bé, si algunes d'aquestes variables predictores no es tenen en compte com a tal, cal controlar-les (per exemple, mantenint el seu valor constant) perquè si no poden esdevenir variables de confusió. Un exemple d'això és el fet de comparar el salari de diferents zones geogràfiques sense tenir en compte la grandària de l'empresa.

Esquema del rol de les variables en la investigació



Normalment hi ha una seguit de variables que influeixen sobre una variable dependent. Per tant, l'investigador triarà –en funció dels seus interessos– una o unes quantes d'aquestes variables com a objectiu del seu treball. Aquestes variables passaran a ser les variables independents i caldrà un control de la resta de variables ja que constitueixen possibles fonts de variació que produïrien error.

Si en aquest exemple l'investigador decideix estudiar l'efecte de les variables *B* i *E* sobre la variable dependent, l'esquema quedaria de la següent manera.



Manipulació **Mesura** **Control**
(Operacions que es fan sobre les variables)

1.3.2. Classificació de les variables des del punt de vista estadístic

Des del punt de vista estadístic, les variables estan relacionades amb l'escala de mesura en què han estat recollides. Per tant, ajuntarem el tipus de variable des del punt de vista estadístic i el seu corresponent per cada **escala de mesura**.

Identificar correctament el tipus d'escala amb què s'ha mesurat una variable és fonamental, ja que determina quin tipus d'anàlisi de dades s'hi pot fer.

Com que l'objectiu del mòdul no és, ni de bon tros, entrar a fons en la **teoria de la mesura**, en el subapartat 1.3.4 esmentarem únicament la classificació feta per Stevens el 1951. Aquesta classificació s'articula en funció de la distinció entre quatre tipus d'escales: nominal, ordinal, d'interval i de raó.

Els conceptes mètrics poden assignar nombres reals o vectors a les diferents característiques. En el cas de les magnituds escalars, l'escala de mesura assigna un nombre real a l'objecte. Aquest és el cas de la majoria de variables, per exemple l'edat, l'alçada, el nombre de respostes correctes, etc.

Parlarem de magnituds vectorials en variables en les quals és important, no solament el valor numèric, sinó també la direcció, com per exemple en variables físiques com la velocitat o la força... En el cas de les ciències socials, aquest concepte es pot aplicar a les actituds, ja que tenen magnituds tant de força com de direcció (positiva o negativa) cap a l'objecte de l'actitud.

En el nostre cas, ens limitem a les anomenades *magnituds escalars*, encara que es pot generalitzar perfectament a tot tipus de conceptes mètrics.

Lectures recomanades

L. Jáñez (1989). *Fundamentos de psicología matemática*. Madrid: Pirámide.

F. Salvador (1996). *Quantificació de les observacions: escales de mesura*. Barcelona: UOC.

A. Cosculluela; A. Fornieles; J. Turbany (2008). *Tècniques d'anàlisi de dades quantitatives*. Barcelona: UOC.

Lectura recomanada

Per a una exposició més ampla de la teoria de la mesura consulteu, per exemple, l'obra de L. Jáñez (1989). *Fundamentos de psicología matemática*. Madrid: Pirámide.

1.3.3. Transformació de les dades

Formalment, en l'àmbit científic podem definir la transformació d'una variable com el resultat de fer alguna operació matemàtica idèntica sobre tots els seus valors, de manera que cada un d'ells guardi la mateixa correspondència amb les dades de la variable original. En aquest mateix sentit, en el *Diccionari de la llengua catalana* (Institut d'Estudis Catalans), es defineix *transformació* en l'accepció matemàtica com a 'aplicació bijectiva entre dos subconjunts d'un espai'.

Transformar al DRAE

Transformar, segons el diccionari de la Reial Acadèmia de la Llengua Espanyola, és 'hacer cambiar de forma a alguien o algo. Transmutar algo en otra cosa'.

Com assenyala Salvador (1996), les transformacions consisteixen a obtenir valors numèrics diferents dels originals, mantenint la capacitat de representació de les relacions empíriques.

Exemples de transformacions de variables

En la nostra societat hi ha nombrosos exemples de transformacions de variables, ja sigui per a facilitar els càlculs o per motius culturals. Així, és fàcil trobar regles amb escales en centímetres i en polzades, balances amb grams i lliures ($1 \text{ kg} = 0,45 \text{ lliures}$), termòmetres amb graduació en graus centígrads i Fahrenheit ($C = 5/9 F - 160/9$), etc. Com veurem més endavant, la major part d'aquestes transformacions són de tipus lineal i solen consistir en una simple combinació d'operacions aritmètiques.

Un bon indicador de la utilitat de les transformacions el podem trobar en el fet que pràcticament tots els programes informàtics de tractament de dades tenen implementades multitud de funcions i utilitats destinades a facilitar-ne la realització. Així, per exemple, l'SPSS té diverses instruccions exclusivament dedicades a les transformacions.

L'objectiu de les transformacions de les dades és aconseguir algun avantatge en les anàlisis perservant, alhora, la informació rellevant i no dificultant –més ben dit, en general, facilitant– la interpretació dels resultats. Algunes transformacions permeten, a més de simplificar els càlculs, augmentar el nivell de potència, millorar la simetria de la distribució, comparar valors de distribucions diferents, etc., cosa que facilita l'aplicabilitat de les proves estadístiques amb què usualment s'analitzen les dades. A més, moltes vegades, en cas que ens interressi, permeten apropar la forma d'una relació no lineal a una recta, en modificar l'escalat de la variable.

1.3.4. Tipus d'escales de mesura

Mesurar és el procés d'atribuir valors numèrics (conceptes mètrics) a les variables de la mostra o població, amb la intenció de representar els atributs mesurats. Aquest procés es fa per mitjà d'una escala de mesura, amb unes regles

Lectura recomanada

F. Salvador (1996). *Quantificació de les observacions: escales de mesura*. Barcelona: UOC.

establertes de les quals s'ocupa la teoria de la mida. En funció de les característiques del que s'ha mesurat i de com s'ha fet la mesura, es poden establir diferents tipus d'escala de mesura:

1) **Nominal**. Es tracta del nivell de llenguatge qualitatiu. La classificació o taxonomia solament ens permet verificar empíricament la relació d'**igualtat** o **desigualtat**. Una classificació d'un grup determinat d'objectes o successos (domini) en subgrups representa una partició del conjunt. En termes estadístics s'acostumen a anomenar *variables qualitatives* o *categòriques*.

Les categories han de estar ben definides operacionalment i cal que siguin **exhaustives** (han de cobrir completament tot el domini o possibles valors de la variable) i **mútuament excloents** (no hi pot haver intersecció entre les diferents categories). És a dir, cada objecte o succés per classificar ha de poder ser assignat únicament a una categoria.

a) Transformacions possibles. Qualsevol, els valors numèrics són meres etiquetes de les categories, per tant, poden ser substituïdes per qualsevol altra, sempre que els nous valors segueixin essent diferents entre ells.

b) Estadístics. Els propis de **variables qualitatives** o **categòriques**, com per exemple freqüències, percentatges, proporcions, moda.

c) Exemples. Sexe, nacionalitat, color del cabell...

2) **Ordinal**. Nivell de llenguatge comparatiu que manté les característiques de les anteriors i, a més de verificar la igualtat/desigualtat, incorpora l'**ordre**.

En la classificació estadística seran **qualitatives ordenades** o simplement **variables ordinals**.

a) Transformacions. Isotònica, els diferents valors numèrics ens indiquen l'ordenació de les categories, per tant, la transformació ha de mantenir l'ordenació original.

b) Estadístics. Els de les escales nominals més la mediana, els centils i de més índex de posició i la correlació ordinal.

c) Exemples. L'ordre d'arribada en una cursa, les qualificacions (suspès, aprovat, notable, excel·lent, matrícula d'honor), una gradació del grau d'acord (totalment d'acord, d'acord, indiferent, en desacord, totalment en desacord), el nivell d'estudis (primaris, secundaris, mitjans, superiors...), etc.

3) **Interval**. A més de verificar la igualtat/desigualtat i l'ordre, ens indica la **magnitud de les diferències**. El **valor nul** de l'escala és **designat arbitràriament** (per convenció). És a dir, el valor 0 no indica la manca total de l'atribut i en molts casos pot tenir sentit parlar de valors negatius. En la temperatura

mesurada en graus centígrads té sentit parlar de $-5\text{ }^{\circ}\text{C}$, ja que $0\text{ }^{\circ}\text{C}$ no és la mínima temperatura possible. En termes estadístics es tracta de **variables quantitatives**.

a) Nivell de llenguatge quantitatiu: magnituds intensives; parlem de magnituds intensives quan la combinació d'objectes no és additiva.

b) Transformació: lineal.

c) Estadístics: els de les escales anteriors més la mitjana, la variància i la desviació tipus, la correlació lineal, etc.

Exemple de magnitud intensiva

Si ajuntem dos líquids de diferents temperatures, la combinació resultant no tindrà una temperatura que sigui la suma de les dues anteriors.

d) Exemples: temperatura en $^{\circ}\text{C}$, quocient intel·lectual, etc.

4) **Raó**. També anomenades *de proporció*. Ens permeten verificar totes les relacions de les anteriors afegint, respecte a la d'interval, el fet que en les escales de raó hi ha una unitat empírica de mesura i, per tant, un valor nul (**0 absolut**) que indica absència de la característica o variable que s'ha de mesurar.

a) El nivell de llenguatge també és quantitatiu: magnituds extensives; es parla de magnituds extensives o additives quan la combinació d'objectes és igual al seu sumatori (per exemple: si ajuntem dos líquids de diferents volums, la combinació resultant tindrà un volum igual a la suma dels dos anteriors). En estadística són també **variables quantitatives**.

b) Transformacions: el subconjunt de les funcions lineals en què la transformació manté el valor 0 absolut.

c) Estadístics: els mateixos que en les escales d'interval.

d) Exemples: longitud, pes, temps de reacció, nombre de respostes, etc.

Finalment, cal assenyalar que les variables quantitatives se subdivideixen en variables quantitatives discretes i contínues. Les primeres són aquelles en les quals no és possible inserir cap valor entre un valor determinat i el següent, com ara el nombre de fills, el nombre de cadires que hi ha en una habitació... És a dir, d'alguna manera, no té sentit parlar de decimals.

Per contra, en les variables quantitatives contínues és possible intercalar infinits valors entre dos valors qualssevol. El nombre de decimals que incorporarem dependrà de la precisió que necessitem i que tingui el nostre instrument

de mesura, però no de les característiques de la variable. Així, per exemple, entre 168 i 168 cm podem posar infinits valors (1.681, 1.682, 1.683, 1.684... mm) i així successivament.

Escales	Transformació verbal	Transformació formal	Invariància	Exemples
Nominal	(Pertinença) Tr. biunívoca		Manté les categories	Materials de construcció, sexe...
Ordinal	(Precedència o preferència) Tr. isotònica	$X_i < X_j \implies X'_i < X'_j$	Manté l'ordre dels valors	Nivell de formació, preferència d'assignatures...
Interval	(Diferència, magnitud) Tr. funcions lineals	$X' = aX + b$	Manté la magnitud de les diferències	Temperatura en °C, quocient intel·lectual...
Raó	(Magnituds extensives i 0 absolut) Tr. funcions lineals de semblança	$X' = aX$	Manté la raó entre els valors i el 0 absolut	Pes, longitud...

1.4. Població i mostra

Algunes vegades, quan la grandària de la població no és molt gran, és possible fer investigacions amb la població sencera i estudiar totes les unitats d'anàlisi de la població. Però, en general, resultaria molt costós i, de vegades, fins i tot impossible. Per això l'habitual en les investigacions és treballar amb subconjunts de la població als quals denominarem *mostres*, que són triades seguint unes determinades normes o tècniques de mostratge.

Vegeu també

Podeu consultar la unitat de vídeo 17 per a complementar la informació d'aquest subapartat.

1.4.1. Població

Entenem per *població* tot el col·lectiu d'unitats d'anàlisi que té algun tret comú i del qual pretenem obtenir alguna informació (León i Montero, 2003; 2009).

Quan parlem d'unitats d'anàlisi, ens referim als subjectes, empreses, països, comunitats... dels quals volem obtenir la informació. En general, la unitat d'anàlisi coincideix amb el concepte de subjectes, però això, com veurem en els exemples, no sempre és així.

1.4.2. Mostra

La mostra és un subconjunt de la població. A partir de la mostra prenem decisions sobre el conjunt de la població. Per tant, la idea fonamental és suposar que si la **mostra** ha estat ben triada, serà **representativa** de la població de referència, és a dir, els resultats obtinguts en la mostra podran ser extrapolats a la població.

Per exemple, per conèixer la intenció de vot per a les pròximes eleccions, les empreses d'estudis estadístics trien una mostra d'uns quants milers de subjectes i extrapolen els resultats que han obtingut d'aquests subjectes.

Que una mostra sigui representativa implica que les característiques del conjunt dels subjectes que la componen siguin semblants al conjunt de subjectes que formen la població. És a dir, per tal de predir quins seran els resultats de les properes eleccions, caldrà que la mostra estigui formada per persones de totes les característiques dels votants. Per tant, serà necessari que hi hagi homes i dones; joves, adults i vells; estudiants, professionals, pagesos... gent de ciutats grans, petites, de pobles, etc.

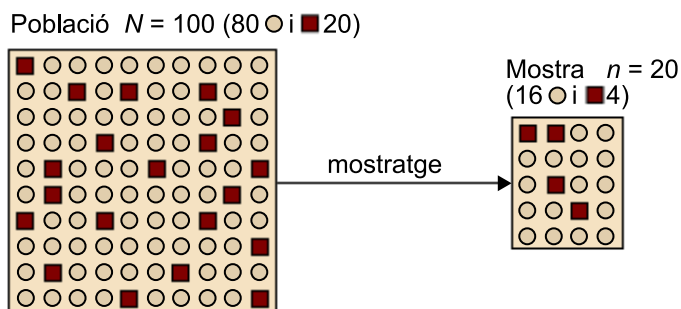
Exemples de poblacions

Els següents són alguns exemples de poblacions:

- 1) Poblacions en què les unitats d'anàlisi són els subjectes:
 - a) Persones amb dret a vot en les properes eleccions autonòmiques.
 - b) Estudiants universitaris a Catalunya.
 - c) Dones histerotomitzades.
 - d) Turistes alemanys a la Costa Brava.
- 2) Poblacions en què les unitats d'anàlisi no són els subjectes:
 - a) Empreses a la província de Barcelona
 - b) Països africans.
 - c) Organitzacions no governamentals.

Exemple

En la següent figura podem observar com es mantenen les proporcions de quadres negres i cercles blancs (afixació proporcional). La mostra és representativa. Si la mostra, per exemple, tingués 15 quadres negres i 5 cercles blancs, direm que està esbiaixada.



No actuar d'aquesta manera i fer una mostra, per exemple, amb la intenció de vot dels estudiants de la Universitat de Barcelona en les properes eleccions autonòmiques impli-

Vegeu també

Les enquestes formen part de les metodologies quantitatives, en concret d'una categoria anomenada *metodologies selectives*. La justificació d'aquest qualificatiu la veurem posteriorment. Podeu consultar la unitat de vídeo 18 per a complementar la informació d'aquest subapartat.

caria, molt probablement, cometre un greu error a l'hora de fer la predicció dels resultats de les eleccions basant-nos en les dades obtingudes amb la nostra mostra.

Una de les dificultats més grans que ens trobem a l'hora de treballar amb mostres és que molts cops no són representatives, ja que encara que hàgim fet el mostratge correctament, es dona el problema de l'autoselecció. Aquest problema radica en el fet que molts cops no participen en la nostra investigació tots els subjectes escollits per a formar part de la mostra, sinó que únicament ho fan aquells que ho volen, i no tenim cap garantia de si aquests subjectes tenen o no alguna característica diferencial respecte als subjectes que no ho fan i, per tant, si són o no representatius de tot el col·lectiu (població).

1.4.3. El mostratge

El mostratge és el procés pel qual algunes de les unitats de la població passen a formar part de la mostra.

El mostratge és un procés que té una importància fonamental en alguns dissenys, com el cas dels dissenys d'enquesta, ja que en no poder exercir altres tipus de controls, la manca de rigor en el mostratge pot provocar **estimacions esbiaixades** de les característiques de la població (Domènech i altres, 1998).

Aquest procés s'ha de portar a terme seguint una sèrie de principis que s'enmarquen en la denominació de *tècniques de mostratge*.

Per tal d'iniciar el procés, abans de començar a mostrejar, cal que fem un cens de les unitats de la població. El cens consisteix a recomptar-les i identificar-les. Hem de tenir en compte que en poblacions grans o disperses, l'elaboració del cens pot resultar força complicada, i un cens esbiaixat pot representar un greu problema a l'hora de trobar una mostra representativa.

Problemes en l'elaboració d'un cens

Si volem saber què pensen els adolescents sobre el problema del consum de drogues, pot ser que ens costi més censar alguns dels joves que precisament tenen aquest problema perquè es troben en ambients marginals; per tant, els resultats que obtindrem de la mostra que hàgim pogut extreure d'aquest cens no donaran una visió realista del problema.

1.4.4. Les tècniques de mostratge

Per a fer el mostratge s'han de seguir unes tècniques que ens indiquen la manera de fer-lo. En general, podem distingir dos grans tipus de tècniques: les tècniques de mostratge probabilístiques i les no probabilístiques.

Mostratge probabilístic

Les tècniques de mostratge probabilístiques són aquelles en les quals la **selecció de les unitats** (subjectes, empreses...) que integraran la mostra es fa **aleatòriament** (per sorteig) i, per tant, totes les unitats de la població tenen les mateixes probabilitats de passar a formar part de la mostra.

Aquesta és la manera més segura de garantir la representativitat de la mostra (de fet, pràcticament és l'única manera que la representativitat no quedi compromesa), ja que s'espera que l'atzar distribueixi totes les característiques de la població entre les unitats que componen la mostra.

Les principals tècniques probabilístiques són el mostratge aleatori simple, l'aleatori estratificat i el mostratge per conglomerats.

En el **mostratge aleatori simple** cada element de la població té la mateixa probabilitat de ser inclòs en la mostra. La tècnica consisteix senzillament en un **sorteig** entre tots els components de la població.

El primer pas per a fer el sorteig és numerar totes les unitats de la població per a, seguidament, triar a l'atzar quines formaran la mostra. Per a fer el sorteig es poden utilitzar molts procediments, com ara posar els números en un bombo, utilitzar un programa informàtic o una calculadora perquè ens faci una llista de nombres aleatoris, utilitzar una taula de nombres aleatoris...

Taula de nombres aleatoris

Una taula de nombres aleatoris és una llista de nombres que s'ha creat a l'atzar i en la qual, per tant, no segueixen cap patró sistemàtic.

Aquesta tècnica és efectiva quan el nombre de població no és gaire gran. Si el nombre de població és molt gran, com ara, per exemple, les persones amb dret a vot en les properes eleccions autonòmiques, és difícil numerar totes les unitats per tal de fer el sorteig i és preferible recórrer al mostratge aleatori estratificat.

El **mostratge aleatori estratificat** consisteix en la creació de diferents **estrats uniformes** (parts) de la població.

Exemple de mostratge aleatori estratificat

Per exemple, si agafem el cas de la població de votants potencials per a les properes eleccions, podrem definir tres estrats a partir del tipus d'hàbitat dels subjectes i tindrem un estrat urbà, un semiurbà i un rural. Aleshores, farem un sorteig aleatori dins de cadascun dels estrats i podrem estar prou segurs que en la mostra hi haurà subjectes de tots els hàbitats (estrats).

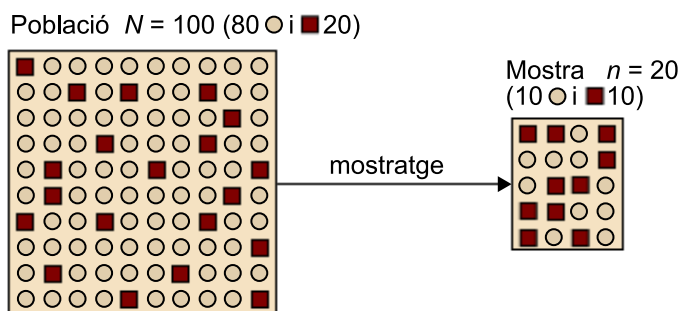
Ara bé, a l'hora de decidir quants subjectes tenim de cada hàbitat en la mostra podem fer dues coses:

- 1) agafar el mateix nombre de subjectes de cada estrat, que s'anomena **afixació simple**, o bé
- 2) fer una **afixació proporcional** i mantenir els percentatges que cada hàbitat representa en la població a la mostra.

Així, si la població està repartida, per exemple, en 50% urbà, 30% semiurbà i 20% de població rural; si decidim triar una mostra de 1.200 subjectes, en el cas de l'afixació simple, agafarem 400 de cadascun dels hàbitats, mentre que si fem servir una afixació proporcional, agafarem 600 de l'hàbitat urbà (50%), 360 del semiurbà (30%) i 240 del rural (20%).

Exemple d'afixació simple

La figura següent mostra un exemple d'afixació simple (50% de quadres negres i cercles blancs). En la figura anterior del subapartat 1.4.2 es pot observar com es mantenen les proporcions de quadres negres i cercles blancs (afixació proporcional).



En el **mostratge per conglomerats** es varia l'estratègia, ja que el sorteig no el farem amb les unitats directament, sinó que agafarem **grups sencers d'unitats**, que per alguna causa constitueixen grups naturals i farem el sorteig de grups, per tant, les unitats que formin part dels grups triats seran les que compondran la mostra.

En l'exemple anterior, en comptes de sortejar els subjectes individualment, ho farem per blocs de pisos, enquestant tots els subjectes que visquin als blocs triats.

Aquest tipus de mostratge s'acostuma a fer quan les poblacions són molt grans o quan la població està distribuïda en grups naturals que impossibiliten o dificulten treballar amb les unitats per separat. Per exemple, si volem fer una enquesta als escolars que cursen 3r. d'ESO, pot ser més senzill seleccionar per conglomerats unes quantes aules i enquestar-los a la mateixa classe, que fer un sorteig entre tot l'alumnat i passar-los l'enquesta fora de l'horari escolar.

En aquest cas correm el risc que els subjectes de conglomerats diferents tinguin també característiques diferents.

Per acabar, cal no confondre entre conglomerats i estrats, ja que els primers poden ser molt heterogenis (de vegades més que la mateixa població), mentre que quan estratifiquem la població busquem trobar molta homogeneïtat en cada estrat (Domènech i altres, 1998).

Mostratge no probabilístic

En les tècniques de mostratge no probabilístiques **no es fa cap sorteig** i, per tant, no totes les unitats de la població tenen les mateixes probabilitats de passar a formar part de la mostra.

Malgrat que es fan servir molt, aquestes maneres de fer les mostres comporten un **greu risc per a la representativitat** de la mostra. Com que són seleccionades per alguna característica concreta o bé pel simple fet que en molts casos són participants voluntaris, no hi ha cap garantia que les unitats triades siguin representatives de tota la població a què pertanyen.

Hi ha diverses tècniques no probabilístiques. Algunes són les següents:

1) Les **mostres accidentals** són aquelles en què els subjectes participen per la seva **pròpia voluntat** en l'enquesta en resposta a una demanda de qui l'organitza. Generalment, aquesta demanda es vehicula per algun mitjà de comunicació com ara un diari o revista, la ràdio o la televisió. En aquestes ocasions, sovint responen molts subjectes que tenen opinions extremes sobre el tema tractat.

Exemple de mostratge accidental a la televisió

En molts programes de televisió es fa una pregunta sobre què opinen els oients o televidents i es demana que truquin a un telèfon determinat o escriguin dient què opinen sobre el tema.

Un altre tipus de mostratge accidental és quan es treballa amb les unitats **disponibles per a l'investigador**.

Exemple de mostratge accidental amb les unitats disponibles per a l'investigador

Un investigador és professor d'una universitat i fa una enquesta als seus alumnes per saber quina és l'opinió dels joves sobre el consum de tabac.

Aquestes mostres són esbiaixades i els resultats obtinguts únicament poden ser útils quan les característiques particulars dels subjectes no difereixen gaire de les generals de la població.

2) El **mostratge fet a propòsit** o **intencional** es produeix quan es decideix per raons històriques o teòriques utilitzar un **determinat col·lectiu de subjectes** com a representació d'una població. Per tal que sigui útil, ha de tenir el suport de resultats anteriors que permetin pensar que les dades mostrals d'aquell col·lectiu són un bon indicador (representatives) de les de la població.

Exemple de mostratge intencional

Per exemple, un investigador d'opinió ha constatat que els resultats generals de les eleccions autonòmiques sempre coincideixen molt aproximadament amb les votacions fetes en un determinat municipi. Es pot plantejar la possibilitat d'enquestar tots els votants del municipi, en comptes de fer un mostratge estratificat, per tal de predir el resultat de tota la població.

3) **Mostratge per quotes**. Cal tenir un bon coneixement de les proporcions que representen els estrats en la població. És semblant al mostratge estratificat aleatori, però sense el caràcter probabilístic. Un cop coneguts els percentatges, fixarem les quotes. És a dir, el nombre de subjectes que compleixen les condicions requerides.

Exemple de mostratge per quotes

Tenim 30 subjectes masculins, de 20 a 30 anys, etc. Un cop determinada la quota, es trien els primers participants que compleixen les condicions fins a completar totes les quotes.

4) **Bola de neu**. Es trien uns quants subjectes i, un cop han finalitzat la seva participació, se'ls demana que convoquin coneguts seus, i així successivament fins a tenir el nombre de participants necessari. Aquest sistema s'utilitza molt en poblacions marginals.

1.4.5. La grandària de la mostra

Haver fet correctament el mostratge no implica que les dades obtingudes siguin extrapolables a la població. Encara que la mostra sigui representativa, cal que contingui prou elements per a poder extreure conclusions generalitzables. No s'ha d'oblidar que l'**error mostral** està relacionat amb la grandària de la mostra.

Per tant, la grandària de la mostra està directament relacionada amb la **precisió de l'estimació** que farem, i també amb el grau de seguretat que tindrem. En definitiva, com més gran és la mostra, més precisa serà la nostra predicció i més segurs estarem dels resultats. Però cal tenir en compte que les mostres grans són més "cares" d'aconseguir, per tant, cal trobar un punt d'equilibri entre la precisió i el cost econòmic i personal. A més, cal considerar que malgrat el que acabem de dir, una mostra gran no és, per si sola, una garantia de la seva representativitat. La grandària mostral és una condició necessària però no suficient de representativitat.

Per a calcular la grandària de la mostra requerida per a cada cas concret, hi ha una sèrie de procediments estadístics que dependran, a més de la precisió i la seguretat, d'aspectes com l'estadístic emprat.

En general, podem dir que la grandària de la mostra que necessitem depèn de diferents aspectes, entre els quals destaquen: la grandària de la població, el tractament estadístic que volem donar a les dades, la variabilitat de la mesura a la població, els marges admissibles de les nostres prediccions i el nivell de seguretat que volem tenir.

1.5. Qualitat de les investigacions

En aquest subapartat ens referirem a un parell de qüestions fonamentals en tota investigació. La **validesa** fa referència a les possibilitats reals que tenim de generalitzar les nostres troballes i en quin nivell de seguretat ho podem fer.

La **validesa interna** d'una investigació fa referència al grau de seguretat que tenim que la relació que hem establert entre la variable independent i la variable dependent és **causal**. És a dir, el grau de confiança amb què podem establir que els canvis enregistrats en la variable dependent s'han produït com a efecte dels canvis que hem introduït en la variable independent.

Si recordem la definició de *relació causal*, en la validesa interna estarà fortament implicat el concepte de *control de les variables estranyes*, ja que solament amb aquests controls podem garantir que les variables independents són l'única causa dels canvis enregistrats a la variable dependent.

La validesa interna es pot veure afectada per una sèrie d'amenaques que tenen com a problema principal el fet que les variables estranyes actuïn de manera diferenciada sobre els grups experimentals. Seguint el criteri clàssic (Campbell i Stanley, 1966), podem dir que aquestes amenaces són:

1) **La història**. Es refereix a tot allò que succeeix a cada subjecte experimental durant la seva participació a l'experiment. Per tant, es tracta de com afecten el subjecte experimental esdeveniments que són externs al subjecte mateix.

2) **La maduració**. És semblant a la història, la diferència principal radica en el fet que la maduració es refereix a esdeveniments interns del subjecte mateix. La maduració pot ser tant biològica com psicològica.

Lectura recomanada

Podeu trobar gairebé en tots els manuals d'estadística prou informació per a poder fer els càlculs necessaris a fi de trobar la grandària de la mostra.

3) **Els pretests.** Les mesures prèvies a l'aplicació dels tractaments, especialment la mesura de la variable dependent, poden afectar per diferents causes (familiarització, creació d'expectatives...) el rendiment dels subjectes en la tasca experimental.

4) **La instrumentació.** Es refereix a la precisió i constància de mesura dels instruments.

5) **La regressió a la mitjana.** Aquesta amenaça pot afectar les investigacions que facin servir com a variable independent valors extrems d'alguns tret dels subjectes –per exemple, quocient intel·lectual alt enfront de baix, extravertits enfront d'introvertits... L'error es pot produir a l'hora de fer la mesura per classificar els subjectes com extrems, quan alguns d'ells potser no ho són i han obtingut una puntuació més extrema del que és normal en ells.

6) **La selecció diferencial.** Fa referència a un error en l'assignació dels subjectes als grups, que motiva que aquests ja no siguin comparables inicialment. Un cas de selecció diferencial és que un dels grups estigui format per subjectes voluntaris, mentre que l'altre ho estigui per subjectes no voluntaris.

7) **La mortalitat experimental.** S'entén com a *mortalitat experimental* el fet que al llarg de la investigació alguns subjectes deixen de participar-hi. Per tant, això pot afectar especialment els estudis longitudinals, en què de vegades es poden perdre més subjectes d'uns grups que d'altres. Això constitueix una mortalitat experimental diferencial que pot afectar la validesa interna.

A més, cal dir que aquestes amenaces poden interactuar entre elles, augmentant d'aquesta manera el seu efecte sobre la validesa interna.

La **validesa externa** indica el nivell de representativitat i, per tant, fa referència al grau en què podem generalitzar els resultats obtinguts en una investigació (les relacions trobades), o bé en altres situacions (també anomenada *ecològica*), en altres subjectes o bé en altres variables.

Diferència entre validesa interna i externa

La validesa interna té a veure amb el control de les variables estranyes. La validesa externa amb la representativitat i la generalització de subjectes, situacions o variables.

En quines altres situacions, subjectes o poblacions podem generalitzar els efectes obtinguts en la nostra investigació? Com podem deduir fàcilment, la validesa interna és prèvia i més important que l'externa, ja que si no podem estar segurs que la relació entre variable independent i variable dependent és la desitjada, no servirà de gaire poder-la generalitzar.

Lectures recomanades

Per a aprofundir més sobre la validesa interna i l'externa podeu llegir les següents obres:

O. León; I. Montero (2003). *Métodos de investigación* (3a. ed.) Madrid: McGraw-Hill.

O. León; I. Montero (2009). *Métodes d'investigació quantitativa*. Barcelona: UOC.

1.6. Presentació d'informes científics

L'informe científic es compon d'una sèrie d'apartats fixos que a continuació assenyalarem breument:

1) **Títol.** Ha de donar una idea precisa del contingut del treball.

2) **Autors.** Nom i cognoms dels autors. Cal posar una adreça, en general institucional, de contacte (universitat, empresa, institució...), per si algú que llegeix el treball vol obtenir més informació.

3) **Resum.** El treball ha de contenir un resum d'un màxim de 15 línies en què se sintetitzen els aspectes més rellevants del marc teòric que ha promogut el treball, els objectius, els resultats més importants i les conclusions que se'n poden extreure. El resum és molt important, ja que és una informació que s'integra en les bases de dades de consultes (CD-ROM, publicacions d'abstractes...) i possibilita que les persones potencialment interessades puguin decidir si volen aprofundir en el treball o no.

4) **Paraules clau (keywords).** Les paraules clau són quatre o cinc paraules que recullen els temes principals de la investigació. Igual que els apartats anteriors, aquesta informació és un camp de les bases de dades de consulta. Mitjançant aquestes paraules clau es poden establir filtres per a obtenir únicament els treballs que corresponen als temes triats.

Exemple de paraules clau

Si publiquem els resultats d'una enquesta salarial feta per a estudiar les remuneracions de mercat, les paraules clau poden ser: enquestes, salaris, remuneració de mercat...

5) **Introducció.** En la introducció s'estableix el **marc teòric** de la recerca. És a dir, es planteja el **problema** i les **hipòtesis** a partir dels resultats obtinguts en investigacions anteriors i dels objectius dels autors. Per aquest motiu, en aquesta fase és fonamental la **documentació** per tal d'establir l'estat de la recerca en el tema d'interès. La documentació és el procés de recollida d'informació provinent d'investigacions anteriors sobre el tema. Tanmateix, es concreten els **objectius** de la recerca.

6) **Mètode.** En aquest apartat s'especifiquen totes les dades metodològiques que permetin replicar el treball. Se subdivideix en tres subapartats:

a) **Materials.** En aquest subapartat s'especifiquen tots els materials: qüestionaris, instruments, ordinadors i programes emprats per al tractament estadístic de les dades, etc. En el cas dels qüestionaris i les entrevistes, és convenient adjuntar un exemplar a l'annex dels diferents tipus de qüestionaris o del protocol de l'entrevista, ja que aquests acostumen a ser eines creades *ad hoc* i, per tant, no estandarditzades.

b) Participants. En el subapartat de participants s'especifiquen les característiques rellevants dels subjectes que han contestat a les preguntes de la investigació (dades demogràfiques, trets distintius...). Tanmateix, s'han d'especificar les tècniques de mostreig que s'han emprat, el nombre total de subjectes i, en general, qualsevol altra qüestió relacionada amb els subjectes.

c) Procediment. En el procediment s'especifiquen tots els **passos de la investigació** (com s'ha fet), especialment les que fan referència a la funció de les unitats d'anàlisi (generalment subjectes) des de l'inici de la seva participació en la investigació fins al final. En aquest subapartat també farem referència a l'operativització i mesura de les variables i a l'estratègia utilitzada per a obtenir la informació (per correu, entrevistes...).

7) Resultats. En aquest apartat es presenten i comenten tots els resultats obtinguts en la investigació. Cal que recullin, com a mínim, **informació descriptiva** de cadascuna de les variables estudiades. Aquesta informació es pot presentar mitjançant **taules i gràfics**. Aquest últim procediment és força recomanable ja que permet una visió ràpida dels trets generals més importants.

En la major part de recerques també és interessant trobar **relacions entre variables**. Aquestes es poden presentar de diferents formes, o bé amb una mera descripció segmentada ("tallada") d'una variable en funció d'una altra (per exemple, veure la distribució salarial per homes i per dones per separat), o bé amb els índex estadístics corresponents (χ^2 , correlacions i regressions, diferències entre mitjanes o percentatges...).

Si es tracta d'un disseny d'enquestes, en aquest apartat s'acostuma a presentar la fitxa tècnica de l'enquesta, en què s'especifiquen la grandària de la mostra, la precisió obtinguda, els marges d'error i altres aspectes tècnics de la recerca.

8) Discussió. En aquest apartat s'analitzen els aspectes més rellevants trobats en els resultats i s'extreuen les conclusions més importants. D'altra banda, es comparen les troballes amb els objectius i, en cas d'haver-ne, amb les hipòtesis, per tal de veure en quina mesura s'han complert ambdós aspectes.

Per acabar, s'exposen les possibles **repercussions** que es puguin produir a partir de la informació obtinguda i les possibles investigacions futures que completin o complementin el treball i s'indiquen les possibles mancances o **limitacions** que hagi pogut tenir el treball.

9) Referències. En aquest apartat s'han d'incloure totes les referències bibliogràfiques que s'hagin emprat al llarg del treball. Per a fer-ho se seguirà l'ordre alfabètic pel cognom del primer autor que signa el treball. Per a citar més d'un treball del mateix autor, se segueix el criteri cronològic.

Lectures recomanades

Per a aprofundir sobre la presentació dels informes científics podeu consultar les següents obres:

O. León; I. Montero (2003). *Métodos de investigación* (3a. ed.). Madrid: McGraw-Hill.

O. León; I. Montero (2009). *Métodos d'investigació quantitativa*. Barcelona: UOC.

Per tal de fer correctament les referències bibliogràfiques, hi ha una sèrie d'estàndards per a les citacions. Aquests estàndards depenen de la disciplina científica amb la qual s'hagi dut a terme la investigació però, en general, tots inclouen la següent informació:

Nom dels autors i institució que fa el treball, data de publicació, títol del treball, nom de la publicació (revista, llibre...), número i pàgines on es troba el treball (en cas de ser una revista o una col·lecció), ciutat on s'ha editat i nom de l'editorial.

D'altra banda, cal tenir en compte que cada tipus de treball se cita de manera diferent (tesis, llibres, articles...).

Exemples de referències bibliogràfiques

A continuació, posem l'exemple de diferents publicacions en format APA (American Psychological Association): un llibre, un article publicat en una revista, el capítol d'una compilació de treballs i una tesi doctoral:

A. **Andrés Pueyo** (1993). *La inteligencia como fenómeno natural*. València: Promolibro.

J. J. **Aparicio**; J. L. **Zaccagnini** (1980). Memoria y adquisición del conocimiento. *Estudios de Psicología*, 2, 78-92.

G. H. **Bower** (1975). Cognitive Psychology: an introduction. En W.K. Estes (ed.), *Handbook of learning and cognitive processes*. Nova York: Wiley.

A. **Coscolluela** (1990). *Rendimiento escolar, inteligencia y velocidad de procesamiento de la información*. Tesis doctoral no publicada. Universidad de Barcelona.

10) Annex. En l'annex es posa tota la informació complementària que els autors pensen que pot ser d'utilitat per als lectors: un exemplar de cada qüestionari, llistats de resultats o de llocs d'interès, taules...

2. Anàlisi de les dades

Com ja hem comentat en el subapartat 1.3 dedicat a les variables, el tipus d'anàlisi que podrem fer amb les dades depèn, a banda dels nostres interessos, bàsicament de l'escala en què hagin estat mesurades. Malgrat això, el procés sempre començarà per la tabulació de les dades, seguirà amb la seva descripció i representació gràfica i, en el cas d'aquest tipus de dissenys, sovint afegirem la recerca de relacions entre variables.

2.1. Descripció de les variables

L'objectiu de l'estadística és **reduir la informació** a una sèrie d'índexs que puguin ser interpretats. Per tant, el primer pas, un cop tabulades les dades (introduïdes en una matriu de dades), consistirà a **descriure** les variables que hem estudiat (què és?, com són?). El següent pas, propi de l'estadística **inferencial (bàsicament paramètrica)** serà explicar com són les relacions entre variables o per què els fenòmens es produeixen d'una determinada manera (per què?). Finalment, algunes vegades també ens interessarà poder fer prediccions de com o quan es produirà un fenomen determinat a partir de la informació que tenim.

Com que la nostra pretensió es troba lluny de fer un manual d'estadística, la intenció d'aquest subapartat és presentar, amb l'ajuda d'un exemple, alguns dels índexs estadístics necessaris per a l'anàlisi de les dades.

Un investigador vol estudiar la relació entre la valoració que fan els alumnes d'una aula de 3r. d'ESO de l'assignatura de Matemàtiques mitjançant una escala d'opinió i les notes que obtenen en aquesta assignatura. Per tal de dur a terme la investigació, elabora un qüestionari on es reflecteixen les valoracions de diferents aspectes de les Matemàtiques que finalment queden recollides en una puntuació global que es troba en un interval entre 50 i 200.

La taula que presentem a continuació ens informa, per cada un d'aquests 15 alumnes, del sexe, del valor obtingut en aquesta escala i de la nota final del curs en l'assignatura de Matemàtiques, recollida quantitativament i qualitativament.

Alumne	Sexe	Valoració (x)	Nota (y)	Nota qualitativa
1	Noia	62	5,09	Aprovat
2	Noia	90	5,16	Aprovat
3	Noi	103	5,22	Aprovat
4	Noia	120	5,14	Aprovat
5	Noi	126	5,75	Aprovat
6	Noi	194	6,86	Notable
7	Noia	193	7,34	Notable
8	Noia	188	7,34	Notable
9	Noi	130	6,47	Aprovat
10	Noia	174	6,47	Aprovat
11	Noi	155	6,22	Aprovat
12	Noi	136	6,23	Aprovat
13	Noia	156	6,26	Aprovat
14	Noi	122	5,16	Aprovat
15	Noia	113	5,16	Aprovat
Total		2.062		89,87

2.1.1. Índex de tendència central de la distribució d'una variable

Els estadístics de nivell o tendència central intenten descriure quin és el valor representatiu del centre d'un conjunt de dades d'una variable. Els més emprats són la **moda** (valor més freqüent), la **mediana** (valor que divideix la distribució de les dades en dues parts iguals, un cop hem ordenat tots els valors) i la **mitjana** (valor mitjà de totes les dades presents en la distribució).

Vegeu també

Podeu consultar la unitat de vídeo 4 per a complementar la informació d'aquest subapartat.

Mitjana d'una variable

$$\bar{X} = \frac{\sum X_i}{n}$$

La mitjana del nivell de l'escala de valoració $x = 2.062/15 = 137,47$.

La mitjana de les notes de Matemàtiques $y = 89,87/15 = 5,99$.

Mediana de la variable escala de valoració

Primer ordenem els valors

Alumne	Valoració (x)
1	62
2	90
3	103
15	113
4	120
14	122
5	126
9	130
12	136
11	155
13	156
10	174
8	188
7	193
6	194

En aquests casos tenim un nombre senar d'observacions (15), per tant, la mediana és el valor que ocupa la posició 8. Si tenim un nombre parell d'observacions, la mediana serà el terme mitjà de les dues observacions centrals.

La mediana de l'escala de valoració és 130.

La mediana de les notes de Matemàtiques és 6,22.

Moda d'una variable

La moda té el problema de la seva ambigüitat. Podem tenir més d'un valor repetit (moltes modes) o no tenir-ne cap. En el nostre exemple no hi ha moda en la variable *escala de valoració*, mentre que la variable *nota de matemàtiques* té un valor igual a 5,16.

2.1.2. Índex de dispersió o variabilitat de la distribució d'una variable

Els índexs descriptius de variabilitat ens indiquen la dispersió de les dades de la distribució al voltant d'un valor central (normalment la mitjana). Els més utilitzats són la **variància** (terme mitjà de les distàncies quadràtiques de les puntuacions respecte a la mitjana –vegeu quadre de dispersió o variabilitat–), i la **desviació típica** o desviació estàndard (índex basat en l'anterior).

Vegeu també

Podeu consultar la unitat de vídeo 6 per a complementar la informació d'aquest subapartat.

Variància d'una variable

$$S_x^2 = \frac{\sum (X_i - \bar{X}_i)^2}{n} = \frac{\sum X_i^2}{n} - \bar{X}_i^2$$

Desviació típica

$$S_x = \sqrt{S_x^2}$$

x	y	(x - \bar{x})	(y - \bar{y})	(x - \bar{x}) ²	(y - \bar{y}) ²
62	5,09	-75,47	-0,9	5.695,72	0,81
90	5,16	-47,47	-0,83	2.253,4	0,69
103	5,22	-34,47	-0,77	1.188,18	0,59
120	5,14	-17,47	-0,85	305,2	0,72
126	5,75	-11,47	-0,24	131,56	0,06
194	6,86	56,53	0,87	3.195,64	0,76
193	7,34	55,53	1,35	3.083,58	1,82
188	7,34	50,53	1,35	2.553,28	1,87
130	6,47	-7,47	0,48	55,8	0,23
174	6,47	36,53	0,48	1.334,44	0,23
155	6,22	17,53	0,23	307,3	0,05
136	6,23	-1,47	0,24	2,16	0,06
156	6,26	18,53	0,27	343,36	0,07
122	5,16	-15,47	-0,83	239,32	0,69
113	5,16	-24,47	-0,83	598,78	0,69
Total				21.287,7	9,34

La variància de l'escala $s_x^2 = 21.287,72/14 = 1.520,55$

La desviació típica $s_x = 38,99$

La variància de y : $s_y^2 = 9,34/14 = 0,67$

La desviació típica de Y : $s_y = 0,82$

La variància és un valor difícil d'interpretar atès que la seva unitat de mesura és l'original de la variable però elevada al quadrat. La desviació típica conserva les unitats de mesura originals.

Taules de freqüències

Les taules de freqüències es poden utilitzar tant per a descriure variables de tipus qualitatiu com quantitatiu. En aquest últim cas, caldrà "ajuntar" les dades formant intervals.

Exemple de taula de freqüències de la variable *sexe*

Valor	Freqüència	Percentatge	Percentatge acumulat
Noi	7	53,33	53,33
Noia	8	46,67	100
Total	15	100	

Exemple de taula de freqüències de la variable *escala de valoració de les matemàtiques*

Interval	Valor central	Freqüència	Percentatge	Percentatge acumulat
50-69	60	1	6,7	6,7
70-89	80	0	0	6,7
90-109	100	2	13,3	20
110-129	120	4	26,7	46,7
130-149	140	2	13,3	60
150-169	160	2	13,3	73,3
170-189	180	2	13,3	86,7
190-209	200	2	13,3	100
Total		15	100	

Dispersió o variabilitat

La dispersió o variabilitat es refereix a la molta o poca concentració de les dades respecte del valor central. Les dues sèries següents tenen la mateixa mitjana = 10, però la primera és més dispersa que la segona.

1) 2, 4, 7, 10, 13, 16, 18

2) 7, 8, 9, 10, 11, 12, 13

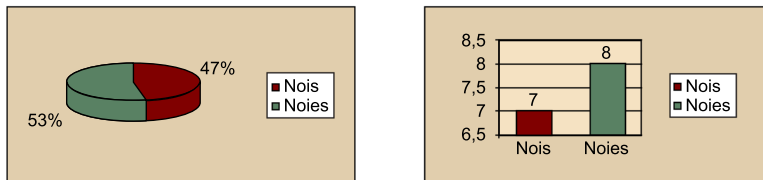
2.1.3. Representacions gràfiques

La representació gràfica de les variables depèn del tipus de variables que volem representar. Les variables qualitatives s'acostumen a representar en gràfics de sectors (pastissos o formatges), o bé en gràfics de barres. Els valors poden estar en freqüència absoluta (el valor del recompte) o en percentatges.

Vegeu també

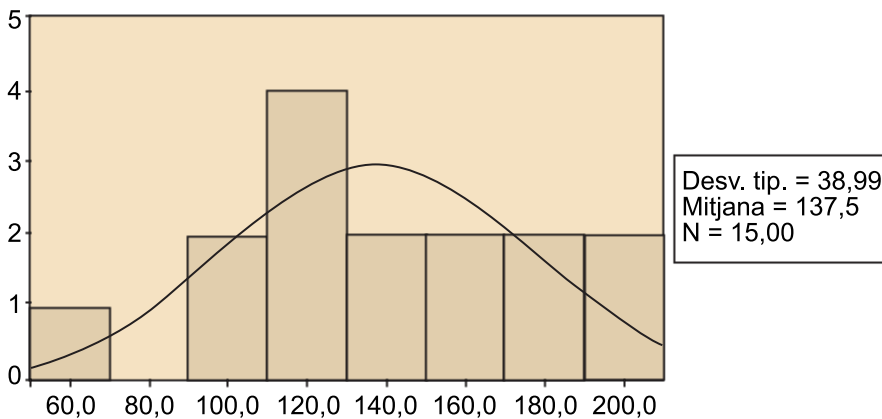
Podeu consultar les unitats de vídeo 3, 5 i 7 per a complementar la informació d'aquest subapartat.

Exemple de representació gràfica de la variable *sexe*

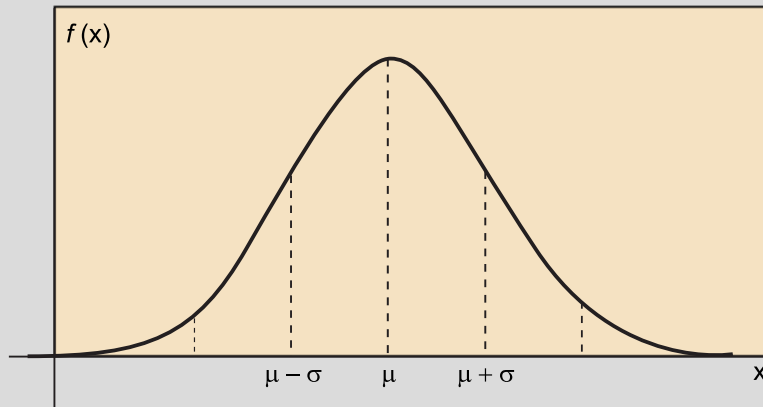


Les variables quantitatives poden ser representades de moltes maneres. Seguidament en mostrarem dues de les més freqüents.

Exemple d'histograma de freqüències (amb la distribució de la corba normal) de la variable *escala de valoració de les matemàtiques*



La **distribució normal** és un model de probabilitat que segueixen determinades variables quantitatives contínues.

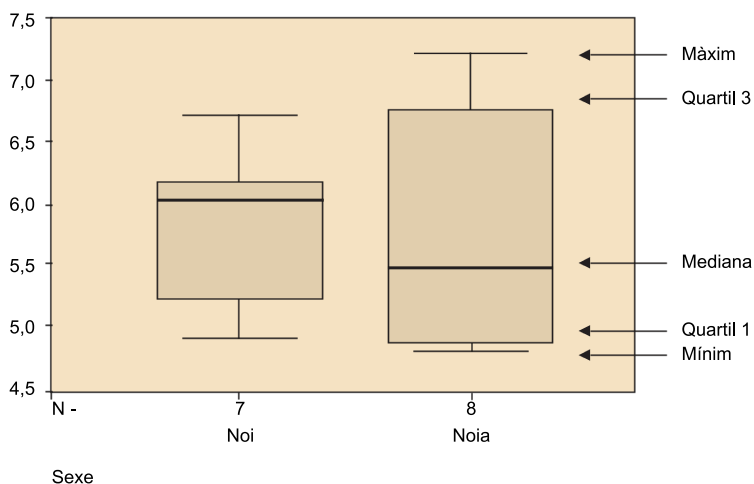


Com veiem en la figura, aquesta distribució és simètrica respecte al valor central. Aquest valor central és representat per la mitjana (μ) de la distribució.

La distribució de valors teòrics és asimptòtica respecte a l'eix horitzontal, vol dir que fluctua entre $-\infty$ i $+\infty$.

La mitjana (μ) i la variància (σ^2) es coneixen com els *paràmetres de la distribució*.

Exemple de diagrama de caixa (boxplot) de la variable notes de matemàtiques. És útil per a veure la distribució de la variables i permet comparar submostres.



En el diagrama de caixa representem cinc valors de càlcul molt senzill, els valors extrems (el més petit i el més gran de la sèrie de dades), la mediana (comentat anteriorment) i els quartils 1 i 3. Aquests valors es calculen de manera simple: un cop calculada la mediana, aquesta ens divideix la sèrie de dades en dues parts iguals. La mediana de la sèrie inferior de les dades representa el quartil 1, la mediana de la sèrie superior serà el quartil 3.

En el nostre exemple, la comparació dels dos diagrames de caixa de *la nota de matemàtiques* en funció del sexe ens permet descobrir diversos aspectes. En efecte, veiem que la nota mediana dels nois és superior a la de les noies, encara que aquestes presenten més dispersió, de fet, la millor nota correspon a una noia. També podem veure que les dades no presenten simetria, ja que en cap dels dos casos la mediana es troba centrada respecte als quartils i als extrems.

2.2. Relacions entre variables

2.2.1. Les proves de significació o d'hipòtesi

Quan intentem trobar relacions entre variables realitzem el que anomenem una *prova de significació o d'hipòtesi*. A continuació, farem algunes consideracions generals al respecte d'aquestes proves.

Aquestes proves consisteixen en una sèrie de procediments encaminats a la presa de decisions estadístiques. Fonamentalment, les decisions seran sobre qüestions que impliquen trobar igualtat o diferència, dependència o independència, ajustament o desajustament. Les preguntes que ens farem seran de l'estil següent:

- Hi ha relació entre la hiperactivitat dels infants segons el sexe?
- Hi ha diferència en la quantitat de hores que veuen la televisió entre els infants catalans i els alemanys?
- Hi ha diferències en el coneixement de l'idioma anglès segons la zona geogràfica de l'Estat espanyol?
- Estudien més hores les noies que els nois durant el Batxillerat?
- El consum d'alcohol és igual en homes i dones?

Per tal de portar a terme una prova d'hipòtesi, cal seguir una sèrie de passos que detallarem a continuació:

1) **Formular la hipòtesi nul·la.** En primer lloc, en aquestes proves estadístiques partim de la formulació d'una mena de "supòsit d'innocència" anomenada *hipòtesi nul·la* (H_0).

La H_0 ens diu que no hi ha diferències entre els grups o relació entre les variables.

Hipòtesi nul·la

En termes generals plantejarem: $H_0: A = B$, $H_1: A \neq B$, $H_1: A < B$, $H_1: A > B$

Vegeu també

Podeu consultar les unitats de vídeo 20 i 21 per a complementar la informació d'aquest subapartat.

a) Atès que la H_0 acostuma a contradir la hipòtesi de treball de la investigació (per exemple, en igualtat de condicions laborals, els homes cobren més que les dones), els anàlisi intentaràn demostrar que, amb un cert risc de error, podem rebutjar la H_0 i, per tant, confirmar la hipòtesi de treball.

Aquesta hipòtesi és complementària de la H_0 i s'anomena *hipòtesi alternativa* (H_1) i indica diferència, desajustament o dependència entre grups, tractaments, variables...

Exemples d'hipòtesi nul·la

H_0 : No hi ha relació entre hiperactivitat i el sexe.

H_0 : No hi ha diferències en les hores d'estudi entre nois i noies durant el batxillerat.

H_0 : No hi ha diferències en el consum d'alcohol entre els homes i les dones.

H_0 : No hi ha diferències en el nivell intel·lectual entre homes i dones.

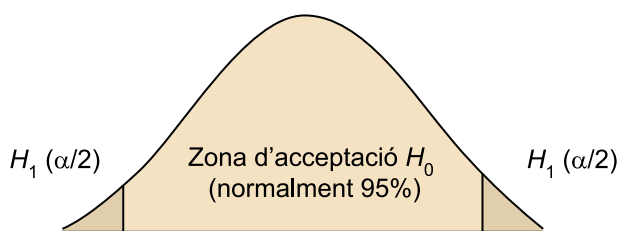
H_1 : Hi ha diferència en el consum de tabac entre els dos sexes.

H_1 : Les noies estudien més hores que els nois durant el batxillerat.

H_1 : Hi ha més nens amb trastorn d'hiperactivitat que nenes.

H_1 : L'alumnat prefereix les Ciències socials a les Matemàtiques.

2) **Nivell de risc.** Com tota presa de decisions, les decisions estadístiques comporten una sèrie de riscos que caldrà preveure. Per tant, el primer pas serà veure quin nivell de risc estem disposats a assumir. D'aquesta manera, fixarem un risc α . Això consisteix a fixar *a priori* el nivell de significació, establint la regió crítica. Aquest risc és el que acceptem d'equivocar-nos rebutjant una H_0 (per exemple, dient que hi ha diferència entre els grups) quan és vertadera (en realitat no hi ha diferències). En la següent figura es pot veure com es concreta aquest pas.



Les proves d'hipòtesi poden ser unilaterals o bilaterals, segons la hipòtesi que vulguem estudiar. Utilitzarem una prova bilateral per a demostrar diferències sense tenir una hipòtesi del sentit d'aquesta diferència ($\mu_a \neq \mu_b$), mentre que farem servir una prova unilateral si únicament ens interessa saber si, per exemple, A és més gran que B ($\mu_A > \mu_B$). És més fàcil demostrar diferències en proves unilaterals que en bilaterals.

3) **Presa de la decisió.** El test d'hipòtesi és el procediment pel qual decidim entre la hipòtesi de nul·litat i l'alternativa. Aquesta decisió es pren en funció de si el resultat de la prova estadística cau en la zona d'acceptació de la H_0 o fora (regió crítica).

En tota presa de decisions estadístiques correm el risc de cometre errors. Els errors que poden fer-nos equivocar la decisió estadística són l'error de tipus I i el de tipus II. Cometem un **error de tipus I** (o de primera espècie) quan rebutgem una H_0 que en realitat és vertadera. És a dir, quan, per exemple, afirmem que hi ha diferències entre dos grups i en realitat no n'hi ha.

Podem conèixer el seu valor ja que la probabilitat de cometre aquest error és el risc α o nivell de significació que fixem en la prova estadística (habitualment $\alpha < 0,05$). Actualment, els paquets estadístics proporcionen la probabilitat exacta de cometre un error de tipus I (afirmar que hi ha diferències o relació quan en realitat no n'hi ha) anomenada *nivell de significació* o simplement P . En efecte, el nivell de significació (*p-value*, o valor P) ens informa de la probabilitat real que sigui certa la H_0 : si aquesta probabilitat és alta l'acceptarem; per contra, si és baixa la podrem rebutjar. En ciències socials els valors d'alfa solen ser 0,01; 0,05; 0,10; essent el més utilitzat el del 0,05. Per tant, sempre que:

- $P \geq \alpha \rightarrow$ decidirem acceptar H_0
- $P < \alpha \rightarrow$ decidirem rebutjar H_0

Cometem un **error de tipus II** (o de segona espècie) quan acceptem una H_0 que en realitat és falsa (direm que són iguals quan en realitat són diferents). El risc de cometre aquest error se denomina β i acostuma a ser desconegut, però està relacionat amb la grandària de la mostra i també amb el nivell de significació, de manera que podem dir que en disminuir un risc, n'augmentarem l'altre, però la manera correcta de disminuir aquest risc és augmentar la grandària de la mostra.

Taula resum d'errors tipus I i II

	H_0 vertadera	H_0 falsa
Rebutgem H_0	Error tipus I	Correcte
Acceptem H_0	Correcte	Error tipus II

Consideracions finals

Quan agafem una mostra molt gran, pràcticament podrem rebutjar qualsevol H_0 , malgrat que la diferència real (grandària de l'efecte) sigui molt petita. Això es deu al fet que la significació estadística està relacionada amb la grandària de l'efecte i amb la grandària de la mostra.

2.2.2. Les proves de relació o independència

La prova de relació que utilitzarem estarà en funció del tipus de variables amb què estem treballant. A la taula següent podem veure un resum de les principals proves de relació entre parelles de variables.

Taula resum dels tipus de proves de relació més emprades en els diferents casos

Tipus de variables	Prova estadística
Dues variables qualitatives	Prova de χ^2
Dues variables: una qualitativa i una quantitativa	Dos grups: <i>t</i> de Student Més de 2 grups: ANOVA
Dues variables quantitatives	Covariància i correlació de Pearson Model de la regressió lineal

Relació entre variables qualitatives

La prova de χ^2 permet estudiar la relació o dependència entre dues variables qualitatives a partir de les diferències trobades entre les seves freqüències empíriques i les teòriques (les esperades en el cas d'independència o hipòtesi de nul·litat).

Per fer la prova hem de construir la taula de contingència, la qual constitueix una taula de doble entrada en què expressarem les freqüències que hem registrat en la mostra.

Imaginem un exemple en què tenim una mostra de 150 subjectes, en què hem registrat la variable *sexe* i la nota que han obtingut en una prova (aprobat o notable). Volem veure si hi ha relació entre aquestes dues variables, o sigui, si el sexe dels subjectes pot influir o no en la nota obtinguda.

Taula de contingència *nota* × *sexe*

	Noi	Noia	Total
Recompte d'aprovat	60	60	120
Recompte de notable	10	20	30
Recompte total	70	80	150

Primer de tot, hem de calcular la freqüència (freqüència esperada o teòrica) que hauríem d'haver trobat, en el cas que no hi hagués relació entre les variables. Per aconseguir-ho, per cada casella multiplicarem els totals de la freqüència observada en la filera i la columna i ho dividim pel total de la mostra. Per exemple, per a la combinació *noi* i *aprobat*, multipliquem 120 (nombre d'aprovat) per 70 (nombre de nois) i ho dividim per 150 (total de la mostra):

Lectura recomanada

Cal tenir en consideració que per a emprar totes aquestes proves s'han de complir una sèrie de condicions d'aplicació (per exemple, normalitat de les distribucions, nombre mínim de casos, homogeneïtat de variàncies...). Si esteu interessats a aprofundir en aquest tema en podeu trobar informació en pràcticament qualsevol manual d'estadística.

Lectures recomanades

Per a aprofundir sobre la relació entre variables qualitatives podeu consultar les següents obres.

J. Guàrdia; M. Freixa; M. Peró; J. Turbany (2007). *Análisis de datos en psicología*. Madrid: Delta.

D. S. Moore (2004). *Estadística aplicada básica*. Barcelona: Antoni Bosch.

$$F_e(\text{noi/aprovat}) = \frac{120 \cdot 70}{150} = 56$$

En la taula següent tenim tant les freqüències observades com les freqüències esperades, i en negreta hi ha els valors esperats en cada casella en el cas que no hi hagi relació.

Taula de contingència *nota × sexe*

		Noi	Noia	Total
Aprovats	Recompte	60	60	120
	Freqüència esperada	56	64	120
Notables	Recompte	10	20	30
	Freqüència esperada	14	16	30
Recompte total	Recompte	70	80	150
	Freqüència esperada	70	80	150

La lògica de la prova serà la següent: si les freqüències observades són similars a les esperades amb el supòsit de la no-relació (freqüències esperades), llavors acceptarem la H_0 (no hi ha relació). Si les dues freqüències són diferents, llavors decidirem rebutjar la H_0 i, per tant, conclourem que sí que hi ha relació entre les dues variables.

Formalment, expressem les hipòtesis de la següent manera:

- $H_0: f_{\text{observades}} \approx f_{\text{esperades}}$
- $H_1: f_{\text{observades}} \neq f_{\text{esperades}}$

L'estadístic de contrast que utilitzarem es denomina χ^2 , ja que es distribueix seguint aquesta distribució teòrica coneguda. Aquest fet ens serveix per a calcular un valor amb l'expressió de càlcul i saber el valor del grau de significació (*p-value*), que ens proporciona la probabilitat que sigui certa la H_0 .

La fórmula de càlcul de l'estadístic de contrast de χ^2 és la següent:

$$\chi^2 = \sum \sum \frac{(F_o - F_e)^2}{F_e}$$

Aplicant la fórmula a les nostres dades:

$$\chi^2 = \frac{(60-56)^2}{56} + \frac{(60-64)^2}{64} + \frac{(10-14)^2}{14} + \frac{(20-16)^2}{16} = 2,6786$$

El valor de χ^2 és de 2,6786. Per a saber si es tracta d'un valor alt o baix necessitem conèixer la probabilitat que acompanya aquest valor. La distribució de χ^2 no és única, fluctua en funció dels anomenats *graus de llibertat*. Per a aquesta prova els calculem amb $(k - 1) \cdot (l - 1)$, en què k i l són les categories de les dues variables. En el nostre cas tenim dues categories en cada variable categòrica, per tant, els graus de llibertat són $(2 - 1) \cdot (2 - 1) = 1$.

Amb l'ajuda de l'Excel o de qualsevol programa estadístic podem veure que la probabilitat (*p-value*) associada al valor 2,6786 en una distribució de χ^2 amb un grau de llibertat és 0,1017.

Aquest valor representa la probabilitat que sigui certa la H_0 o, dit d'una altra manera, la probabilitat que tenim d'equivocar-nos en el cas de rebutjar la H_0 . Si creiem que aquesta probabilitat és alta acceptarem la H_0 ; si considerem que és baixa la rebutjarem i direm que sembla que hi ha relació entre ambdues variables.

En el nostre cas, una probabilitat 0,1017 (grau de significació) és prou alta, per tant acceptarem la hipòtesi de nul·litat i direm que no hi ha relació entre les dues variables. Cal tenir en compte que el valor 0,1017 representa la probabilitat de cometre un error del tipus I. Clàssicament el valor màxim (α) de cometre un error del tipus I més utilitzat és 0,05; com que 0,1017 és més gran que 0,05, considerem que és prou alta la probabilitat que sigui certa la H_0 , i decidirem acceptar-la (vegeu el subapartat 2.2.1).

Relació entre dues variables: una qualitativa i una quantitativa. Diferència entre mitjanes

Les proves de diferència de mitjanes permeten establir a partir de quina diferència es pot considerar que les mitjanes observades són significativament diferents. També permeten comparar una mitjana observada amb la mitjana de la població.

Les proves es poden fer per a dues mitjanes: les mitjanes obtingudes pels mateixos subjectes (mesures repetides) o per subjectes diferents (dades independents).

Finalment, cal assenyalar que quan comparem dues mitjanes utilitzem la prova de comparació de mitjanes t de Student, mentre que si són més de dues mitjanes, analitzarem les dades amb l'anàlisi de la variància (ANOVA).

Prova t de Student per a grups independents

Aplicarem la prova t de Student per a grups independents quan tinguem dos grups diferents de subjectes (o sigui, si hi ha una variable categòrica amb dues categories), sobre els quals hem registrat una variable quantitativa. Per veure si hi ha o no relació entre les dues variables, haurem d'observar si hi ha diferències estadísticament significatives entre les mitjanes dels dos grups formats.

H_0 : mitjana₁ \approx mitjana₂

H_1 : mitjana₁ \neq mitjana₂

Estandardització

Les unitats de mesura sempre ens afecten en el sentit que no podem saber *a priori* si una diferència és molt gran o poc. La utilització d'un estadístic que segueix una distribució teòrica coneguda (sigui la normal, la t de Student, la khi quadrat, etc.) ens estandarditza aquestes diferències i podem prendre les decisions de manera independent de les unitats originals de les variables.

Suposem que tenim els següents resultats de la prova de comparació de mitjanes entre nois i noies respecte a les puntuacions en l'escala de valoració.

Estadístics del grup

	Sexe	N	Mitjana	Desviació	Error típic de la mitjana
Escala	Noi	7	138,0000	29,2062	11,0389
	Noia	8	137,0000	48,0565	16,9905

L'estadístic de contrast que utilitzarem per a estandarditzar la diferència entre les mitjanes es denomina t de *Student*, ja que es distribueix seguint aquesta distribució teòrica coneguda. La distribució t és una distribució teòrica molt similar a la distribució normal, però no és única sinó que fluctua en funció dels diferents graus de llibertat. En aquesta prova els graus de llibertat són $(n_1 - 1) + (n_2 - 1)$, o sigui, es perd un grau de llibertat en cada un dels dos grups, o el que és el mateix $n - 2$.

És important tenir en compte que a l'hora d'aplicar aquesta prova s'han de comprovar dos supòsits, sobretot si la mostra és petita, inferior a $n = 30$. Suposarem que la variable quantitativa –en el nostre cas la variable escala de valoració– es distribueix normalment a la població d'origen i les variàncies dels dos grups han de ser similars (homocedasticitat). Si no es compleix una de les dues condicions, o cap de les dues, podem aplicar una prova alternativa que s'anomena *prova no paramètrica* de la U de Mann-Whitney.

La fórmula de càlcul de l'estadístic de contrast t és la següent:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

En què s_p^2 és un càlcul de variància agrupada o ponderada en què intervenen les variàncies dels dos grups, segon la següent fórmula:

$$s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}$$

Utilitzant els valors de l'exemple:

$$s_p^2 = \frac{(7-1) \cdot 29,2^2 + (8-1) \cdot 48,06^2}{7+8-2} = 1.637,245$$

$$t = \frac{138-137}{\sqrt{\frac{1637,245}{7} + \frac{1637,245}{8}}} = 0,04$$

Prova de mostres independents

Prova T per a la igualtat de mitjanes					
	<i>t</i>	<i>gl</i>	Sig. (bilateral)	Diferència de mitjanes	Error típic de la diferència
Notes	-0,018	13	0,986	-7,86 E-03	0,4376

S'observa una petita diferència amb el resultat obtingut amb el programa SPSS, de totes maneres és molt petita i es deu als diferents decimals emprats.

També veiem que el valor de t que fem servir és de signe positiu, mentre que el valor de t de la llista ens el dona en negatiu. Això només és degut a quina de les dues mitjanes posem primer, de totes maneres, no té cap incidència ja que la distribució t de Student, igual que la normal, és simètrica i centrada en zero. Per tant, és exactament el mateix treballar en un cantó de la distribució (el dret si és positiu) o en l'altre (l'esquerre si el valor és negatiu).

Amb l'ajuda de l'Excel o de qualsevol programa estadístic podem veure que la probabilitat (valor P) associada al valor $t = 0,04$ en una distribució amb tretze graus de llibertat és 0,484. Aquí també veiem una diferència amb la llista, que en realitat està donant el doble (bilateral) de la probabilitat que representa el grau de significació.

Aquest valor representa la probabilitat que sigui certa la H_0 ; en aquest cas veiem que és molt alta, per tant acceptarem la H_0 , cosa que vol dir que sembla que no hi ha relació entre les dues variables. Podem dir que les dues mitjanes són similars, el sexe dels subjectes sembla no influir en les puntuacions de l'escala de valoració.

Prova t de Student per a dues mesures repetides

Aplicarem la prova t de Student per a dues mesures repetides en el cas que tinguem la mesura d'una variable quantitativa, feta en dos moments temporals diferents, en una mateixa mostra de subjectes.

Els dissenys de mesures repetides tenen l'avantatge sobre les dades independents, en la comparació del mateix grup de subjectes amb ells mateixos, a més a més, si no dividim la mostra en dos grups maximitzem el nombre de subjectes. Per contra, el fet d'intervenir el temps pot aparèixer el que s'anomena *efecte serial*, o sigui, contaminació de la resposta per la mesura anterior.

Suposem que tenim els següents resultats d'una mostra de 15 nois i noies que han contestat una enquesta de valoració al principi i al final d'un determinat curs.

Alumne	Valoració inicial	Valoració final	Y
1	62	98	-36
2	90	116	-26
3	103	198	-95
4	120	140	-20
5	126	75	51
6	194	186	8
7	193	194	-1
8	188	194	-6
9	130	147	-17
10	174	174	0
11	155	192	-37
12	136	183	-47
13	156	166	-10
14	122	116	6
15	113	116	-3

L'última columna de la taula representa la diferència de les dues mesures per a cada subjecte. Per veure si hi ha o no relació entre la variable quantitativa i el moment de mesura (variable qualitativa) haurem d'observar si la mitjana de les diferències és significativament diferent de zero.

$$H_0: \text{mitjana}_Y \approx 0$$

$$H_1: \text{mitjana}_Y \neq 0$$

L'estadístic de contrast que utilitzarem per a estandarditzar la mitjana de les diferències és la t de Student, ja que es distribueix seguint aquesta distribució teòrica coneguda. En aquest cas, com que es tracta d'un sol grup de subjectes només es perd un grau de llibertat, és a dir, els graus de llibertat són $n - 1$.

Suposarem que la variable quantitativa Y (diferència de les puntuacions) es distribueix normalment en la població d'origen. Si tenim evidències que això no és així, haurem d'aplicar la prova no paramètrica de la T de Wilcoxon.

La fórmula de càlcul de l'estadístic de contrast t és la següent:

$$t = \frac{\bar{y}}{\frac{s_y}{\sqrt{n}}}$$

En el nostre exemple la mitjana de la variable Y és $-15,53$ i la desviació típica és $32,17$. Fent servir els valors:

$$t = \frac{-15,53}{\frac{32,17}{\sqrt{15}}} = -1,86$$

Prova t per a mitjanes de dues mostres aparellades

	Valoració inicial	Valoració final
Mitjana	137,4666667	153
Variància	1.520,552381	1.637,714286
Observacions	15	15
Quocient de correlació de Pearson	0,672712441	
Diferència hipotètica de les mitjanes	0	
Graus de llibertat	14	
Estadístic t	-1,869881748	
$P(T \leq t)$ una cua	0,0412786	
Valor crític de t (una cua)	1,761310115	

	Valoració inicial	Valoració final
$P(T \leq t)$ dues cues	0,082557199	
Valor crític de t (dues cues)	2,144786681	

El valor negatiu $(-1,87)$ ens indica que la valoració en conjunt és més alta en la segona mesura, al final del curs, que a l'inici. De totes maneres, el símbol no és important, ja que podríem haver fet la diferència restant a la segona columna la primera. Com que la distribució t de Student és simètrica, és igual el signe que presenti l'estadístic de contrast; en tot cas aquest ens pot ajudar a interpretar la possible relació.

En aquest cas estem al límit, si fem una prova unilateral concentrant alfa en un cantó, i si aquest és del 0,05, haurem de rebutjar la hipòtesi nul·la i, per tant, direm que sembla que si hi ha relació, la valoració és superior al final que a l'inici. En canvi, si escollim fer una prova bilateral, en què repartim alfa a cada cantó, haurem d'acceptar la hipòtesi nul·la: no hi ha prou diferència per a determinar que la valoració inicial i final són diferents.

Prova de l'anàlisi de la variància per a grups independents

Aplicarem la prova de l'anàlisi de la variància (AVAR o ANOVA) per a grups independents quan tinguem dos o més grups diferents de subjectes (o sigui, si hi ha una variable qualitativa amb dues o més categories), sobre els quals hem registrat una variable quantitativa. Per veure si hi ha o no relació entre les dues variables haurem d'observar si hi ha diferències estadísticament significatives entre les mitjanes dels diferents grups formats.

$$H_0: \text{mitjana}_1 \approx \text{mitjana}_2 \approx \text{mitjana}_3 \approx \dots \approx \text{mitjana}_k$$

$$H_1: \text{mitjana}_1 \neq \text{mitjana}_2 \neq \text{mitjana}_3 \neq \dots \neq \text{mitjana}_k$$

Suposem que tenim els següents resultats d'una mostra de 15 nois i noies, dels quals tenim la nota final obtinguda en una prova (amb tres resultats: suspens, aprovat i notable) i les puntuacions en una escala de valoració del curs.

Alumne	Nota final	Valoració final
1	Suspens	98
2	Aprovat	116
3	Notable	198
4	Suspens	140
5	Suspens	75

Alumne	Nota final	Valoració final
6	Aprovat	186
7	Notable	194
8	Aprovat	194
9	Aprovat	147
10	Aprovat	174
11	Notable	192
12	Aprovat	183
13	Suspens	166
14	Suspens	116
15	Notable	116
Suma		2.295

La variació total que presenten les dades es pot descompondre en dues fonts de variabilitat: d'una banda, la produïda pel fet de pertànyer a diferents grups (els suspensos, els aprovats i els notables) s'anomena *suma de quadrats entregrups*, de l'altra, la produïda dintre de cada grup (suma de quadrats intragrups). A partir d'aquestes variabilitats (sumes de quadrats) es fan dues estimacions independents de la variància de les dades. La raó d'aquestes dues variàncies (anomenades *quadrats mitjans*) segueix la distribució *F* de Snedecor.

Els quadrats mitjans (variàncies) es calculen fent el quocient de les sumes de quadrats entregrups i intragrups amb els seus respectius graus de llibertat.

Els graus de llibertat, igual que les sumes de quadrats, presenten una relació additiva:

$$\text{graus de llibertat}_{\text{total}} (n - 1) = \text{graus de llibertat}_{\text{entre}} (k - 1) + \text{graus de llibertat}_{\text{intra}} (n - k)$$

Utilitzarem les dades de l'exemple per a presentar les fórmules més ràpides de càlcul de les sumes de quadrats:

Font de variació entregrups (o explicada):

$$SQ_{\text{entre}} = \sum_{i=1}^{j=k} \frac{\left(\sum_{i=1}^{i=n_k} X_{ik} \right)^2}{n_k} - \frac{\left(\sum_{i=1}^{i=n} X_{ik} \right)^2}{n}$$

$$SQ_{\text{entre}} = \frac{595^2}{5} + \frac{1.000^2}{6} + \frac{700^2}{4} - \frac{2.295^2}{15} = 8.836,67$$

Font de variació intragrups (residual o no explicada):

$$SQ_{\text{intra}} = \sum_{i=1}^{i=n} x_{ik}^2 - \sum_{i=1}^{i=k} \frac{\left(\sum_{i=1}^{n_k} x_{ik} \right)^2}{n_k}$$

$$SQ_{\text{intra}} = (98^2 + 116^2 + \dots + 116^2) - \left(\frac{595^2}{5} + \frac{1.000^2}{6} + \frac{700^2}{4} \right) = 14.091,33$$

Font de variació total:

$$SQ_{\text{total}} = \sum_{i=1}^{i=n} x_{ik}^2 - \frac{\left(\sum_{i=1}^{i=n} x_{ik} \right)^2}{n}$$

$$SQ_{\text{total}} = (98^2 + 116^2 + \dots + 116^2) - \frac{2.295^2}{15} = 22.928$$

Els resultats s'acostumen a presentar en una taula resum de l'anàlisi de la variància. Un exemple aplicat a aquestes mateixes dades és la taula obtinguda mitjançant el programa Excel.

Resum de l'anàlisi de variància d'un factor

Grups	Compte	Suma	Mitjana	Variància
Suspens	5	595	119	1.259
Aprovat	6	1.000	166,6	879,07
Notable	4	700	175	1.553,3

Anàlisi de variància

Origen de les variacions	Suma de quadrats	Graus de llibertat	Mitjana dels quadrats	F	Probabilitat	Valor crític per a F
Entre grups	8.836,67	2	4.418,3	3,76	0,053	3,88
Dins dels grups	1.4091,33	12	1.174,2			
Total	22.928	14				

Es calculen els quadrats mitjans (variàncies) *entre* i *intra* fent la raó entre les respectives sumes de quadrats i els graus de llibertat corresponents:

$$QM_{\text{entre}} = \frac{SQ_{\text{entre}}}{g.ll._{\text{entre}}} = \frac{8.836,67}{3-1} = 4.418,33$$

$$QM_{\text{intra}} = \frac{SQ_{\text{intra}}}{g.ll._{\text{intra}}} = \frac{14.091,33}{15-3} = 1.174,27$$

Finalment, per obtenir l'estadístic de contrast farem el quocient entre les dues estimacions de les variàncies:

$$F_{EC} = \frac{QM_{\text{entre}}}{QM_{\text{intra}}} = \frac{4.418,33}{1174,27} = 3,76$$

La distribució teòrica coneguda que segueix l'estadístic de contrast és la F de Snedecor: aquesta distribució fluctua en funció de dos tipus de graus de llibertat, els anomenats *graus de llibertat del numerador* (que corresponen a la font de variació entresubjectes) i els del *denominador* (corresponen a la font intrasubjectes).

És important destacar que per a suposar la distribució teòrica de l'estadístic de contrast i, per tant, poder aplicar correctament la prova és necessari que es compleixin determinats supòsits: que la distribució de la variable sigui quantitativa (en el nostre cas la puntuació de valoració), les diferents observacions siguin independents i que les variàncies dels diferents grups siguin homogènies (similars). Si no es compleix alguna de les condicions d'aplicació haurem d'utilitzar alguna prova alternativa del tipus de les anomenades *proves no paramètriques*.

Amb l'ajuda de taules, o bé d'una llista d'algun programa estadístic podrem prendre la decisió. En aquest cas, veiem que el grau de significació és del 0,053, si utilitzem com a criteri l'habitual 0,05 veiem que en aquest cas és superior per molt poc i, per tant, hem de concloure que la més probable és la H_0 . Per tant, en el nostre exemple, voldrà dir que la nota obtinguda no afecta a la valoració final expressada pels subjectes

En cas d'aplicar la prova i arribar a la conclusió de rebutjar la H_0 , interpretarem que hi ha diferències significatives entre les mitjanes dels diferents grups. En aquest cas, s'haurà de completar amb alguna prova de contrast per a veure quines són les mitjanes diferents entre elles.

Mesura de la covariació o variabilitat conjunta de dues variables quantitatives

Els índexs descriptius de relació entre dues variables ens indiquen que la variabilitat conjunta es produeix entre els diferents valors de les variables (com la modificació dels valors d'una condueix a la modificació dels valors de l'altra).

Com es pot deduir de les fórmules, la variància d'una variable és una covariància d'una variable amb si mateixa ($s_x^2 = \text{COV}_{xx}$).

Vegeu també

Podeu consultar les unitats de vídeo 9, 11 i 13 per a complementar la informació d'aquest subapartat.

Exemple de càlcul de la covariància

(x)	(y)	(x - \bar{x})	(y - \bar{y})	(x - \bar{x}) · (y - \bar{y})
62	5,09	-75,47	-0,9	67,92
90	5,16	-47,47	-0,83	39,40
103	5,22	-34,47	-0,77	26,54
120	5,14	-17,47	-0,85	14,85
126	5,75	-11,47	-0,24	2,75
194	6,86	56,53	0,87	49,18
193	7,34	55,53	1,35	74,96
188	7,34	50,53	1,35	68,21
130	6,47	-7,47	0,48	-3,59
174	6,47	36,53	0,48	17,53
155	6,22	17,53	0,23	4,03
136	6,23	-1,47	0,24	-0,35
156	6,26	18,53	0,27	5,00
122	5,16	-15,47	-0,83	12,84
113	5,16	-24,47	-0,83	20,31
Total				399,58

La covariància entre X i Y es $COV_{xy} = 399,58/14 = 28,54$

Mesura de la correlació entre dues variables

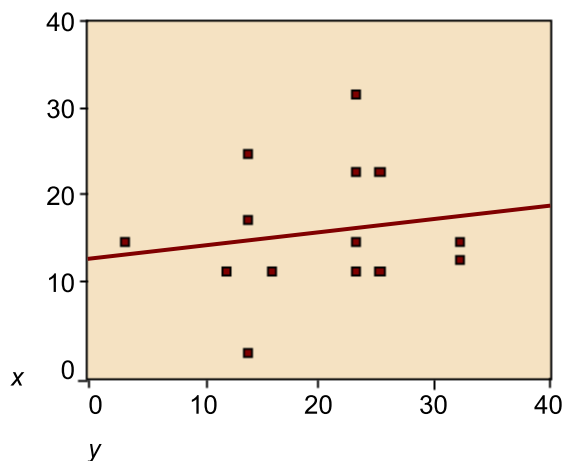
El problema de la covariància és que es troba afectada per la unitat de mesura, i no permet comparar les diferents covariacions entre diferents parelles de variables. Una manera de posar-hi remei és el càlcul del quocient de correlació moment-producte de Pearson. Aquest valor constitueix, de fet, una covariància estandarditzada.

$$\text{Correlació Pearson: } r_{xy} = \frac{\sum z_x z_y}{n} = \frac{S_{xy}}{S_x S_y} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$$

Els valors del quocient de correlació de Pearson fluctuen entre -1 i 1, i s'interpreten de la següent manera:

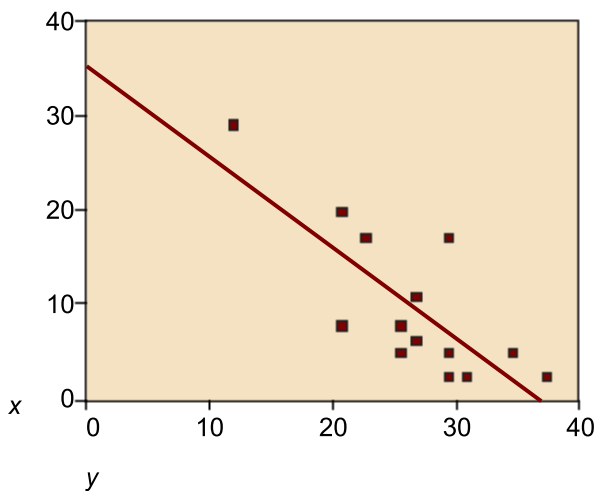
1) Un valor de correlació 0 o pròxim a 0 indica absència de relació lineal entre les dues variables.

Exemple de no-correlació



2) Un valor de correlació igual o pròxim a -1 ens indica una forta relació negativa o inversa (a mesura que augmenten els valors d'una de les variables disminueixen els de l'altra). El valor -1 és la correlació inversa perfecta.

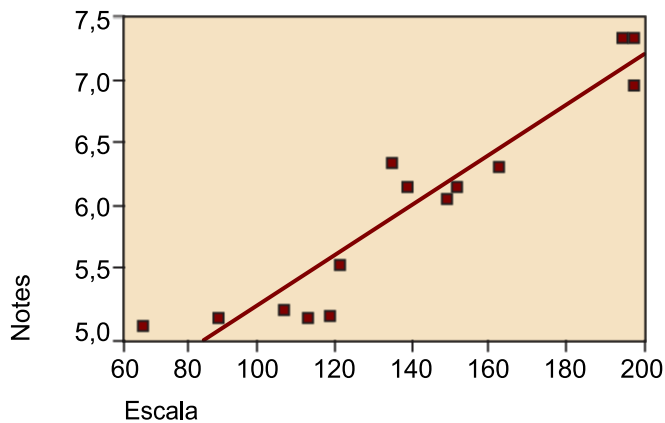
Exemple de correlació negativa



3) Un valor de correlació pròxim a 1 ens indica una forta relació positiva o directa (a mesura que augmenten els valors d'una de les variables també augmenten els de l'altra). El valor 1 és la correlació perfecta.

En l'exemple proposat, el valor del quocient de correlació és igual a 0,89. Aquest valor és indicador d'una forta relació positiva entre les dues variables. A mesura que augmenten la valoració de la assignatura de Matemàtiques, es detecta un augment de la nota obtinguda per l'alumnat en aquesta assignatura.

Exemple de correlació positiva (notes i valoració)



Representació gràfica de la relació entre la variable escala de valoració de les Matemàtiques i les notes en Matemàtiques mitjançant un gràfic de dispersió amb la recta de regressió.

Regressió lineal

La regressió es fonamenta en la relació que s'estableix entre les variables per a, mitjançant una equació, poder predir el valor que esperem que prengui una variable criteri (Y) en funció del valor de l'altra variable predictor (X).

Els paràmetres de l'equació de la recta són el seu pendent (b) i la intercepció o punt de tall de l'eix d'abscisses (a). D'altra banda, el grau de seguretat que tenim que les nostres prediccions són correctes és determinat pel quadrat de l'índex de correlació, anomenat *quotient de determinació* (correspon al percentatge de variància compartida).

Vegeu també

Podeu consultar la unitat de vídeo 12 per a complementar la informació d'aquest subapartat.

La recta de regressió de Y sobre X es

$$Y' = b_{yx} X_i + a_{yx}$$

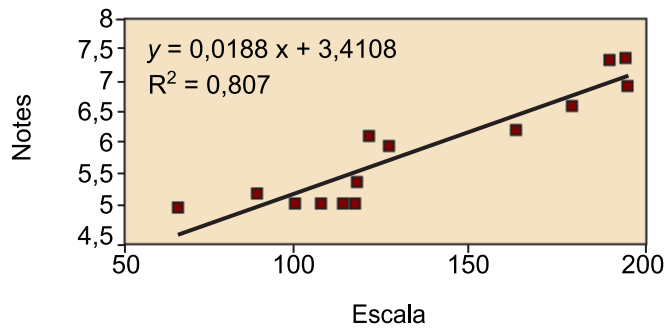
on el pendent de la recta és

$$b_{yx} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

i l'ordenada en l'origen

$$a_{yx} = \bar{Y} - b_{yx} \bar{X}$$

En l'exemple, els resultats de la regressió són:



Per tant, l'equació de regressió serà: **Nota = 3,41 + (0,0188 escala)**

Per exemple, un alumne que hagi puntuat 100 en l'escala, és d'esperar que tingui una nota de $3,41 + 1,88 = 5,29$.

La regressió representa un pas endavant en l'estudi de les possibles relacions entre variables. En efecte, en aplicar la tècnica de regressió estem ajustant les nostres dades a un model, en aquest cas a l'anomenat *model lineal general*. Dintre d'aquest model lineal s'emmarca tant la tècnica de regressió simple que hem comentat, com el model de l'anàlisi de la variància. D'altra banda, aquest model lineal general també ens obre la porta de tota l'estadística multivariant. No ens limitarem a buscar relacions entre dues variables, sinó que buscarem models d'ajust a conjunts de variables.

Lectures recomanades

Per a aprofundir sobre la regressió lineal consulteu la següent bibliografia.

J. Guàrdia; M. Freixa; M. Però; J. Turbany (2007). *Análisis de datos en psicología*. Madrid: Delta.

D. Peña (2002). *Regresión y el diseño de experimentos*. Madrid: Alianza.

Regressió múltiple

La regressió múltiple és una tècnica multivariant que ens servirà per a predir els valors d'una variable dependent a partir de dues o més variables independents. En el cas més simple d'una VD i dues VI el model serà ajustat per l'equació d'un pla:

$$Y' = b_1X_{1i} + b_2X_{2i} + a$$

Suposem que tenim una mostra de 15 subjectes i que volem predir el valor de la nota de Matemàtiques (VD), a partir de les puntuacions obtingudes en una escala determinada (VI) i el quocient d'intel·ligència (VI).

Subjecte	Escala	QI	Matemàtiques
1	62	91	5,09
2	90	93	5,16
3	70	100	5,22
4	65	98	5,14
5	68	103	5,75

Subjecte	Escala	QI	Matemàtiques
6	122	102	6,86
7	126	110	7,34
8	130	115	7,34
9	122	103	6,47
10	120	105	6,47
11	111	110	6,22
12	100	112	6,23
13	103	108	6,26
14	88	95	5,16
15	80	96	5,16

La llista que produeix el programa SPSS:

Variables Entered/Removed (b)

Model	Variables Entered	Variables Removed	Method
1	QI, Escala (a)	.	Enter

a) All requested variables entered.

b) Dependent Variable: Mates

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,945(a)	,892	,874	,28871

a) Predictors: (Constant), QI, Escala

ANOVA(b)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	8,296	2	4,148	49,760	,000(a)
	Residual	1,000	12	,083		
	Total	9,296	14			

a) Predictors: (Constant), QI, Escala

b) Dependent Variable: Mates

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta	B	Std. Error
1	(Constant)>	-,663	1,247		-,532	,604

a) Dependent Variable: Mates

	Escala	,021	,004	,622	4,764	,000
	QI	,045	,015	,402	3,080	,010

a) *Dependent Variable: Mates*

El valor de R representa l'indicador de la correlació múltiple entre la variable dependent i les altres dues variables independents. Recordem que en el cas de la regressió simple (1 VD i 1 VI) aquest valor coincideix amb el valor de correlació de Pearson.

El valor de R de Square (R^2) representa el percentatge de variabilitat explicada de la variable dependent per les dues independents. En aquest cas, un 89,2% de la variància de la nota de Matemàtiques està explicada per la variabilitat produïda entre les altres dues variables. Un valor igual a 1 o el 100% indica una relació perfecta entre les variables. La predicció no produeix cap error. Aquest valor, juntament amb la taula de l'ANOVA, ens indica si el model té un bon ajust o no. En aquest cas, veiem que és un ajust molt bo, ja que les dues variables independents ens expliquen molta part de la variabilitat presentada per la variable dependent. Si ens fixem en la taula de l'ANOVA veiem que el grau de significació és pròxim a zero, cosa que vol dir que podem rebutjar la H_0 , que ens diu que no hi ha relació entre les variables.

La taula de quocients ens ofereix l'estimació dels quocients de l'expressió del pla de regressió. En aquest cas, a més a més de donar un indicador general de la significació de la regressió (com el que hem vist amb l'ANOVA), també ens ofereix la significació dels diferents quocients. En aquest cas, veiem que els dos quocients associats a les dues variables independents són significatius (pròxims a 0); en canvi, el terme constant (a) no és significatiu (el grau de significació associat és superior a 0,05). Si un quocient que afecta una variable independent no és significatiu, llavors traurem la variable del model de regressió; en canvi, el terme constant sempre l'inclourem en el model sigui o no significatiu.

En aquest cas, l'expressió que ens servirà per a predir la nota de Matemàtiques, a partir de la nota de l'Escala i el quocient d'intel·ligència, serà la següent:

$$\text{Pred_Matemàtiques} = 0,021 \cdot \text{Escala} + 0,045 \cdot \text{QI} - 0,663$$

2.3. Aproximació a l'anàlisi multivariable

La regressió múltiple ens ha obert la porta de totes les tècniques d'anàlisi multivariable. La complexitat dels fenòmens que normalment volem estudiar, on intervenen moltes variables, junt amb el gran desenvolupament que s'ha produït en l'aplicació dels diferents programes de paquets estadístics, ha

permès un gran ús de les diferents tècniques multivariables. En aquest subpartat farem una petita aproximació als objectius d'algunes de les tècniques d'ús més comú.

Es poden fer diferents classificacions segons diversos criteris; una primera classificació pot ser en funció de la finalitat que persegueixen. En el quadre següent veiem una classificació de les diferents tècniques que presentem.

Descriptives	Anàlisi en components principals. Anàlisi de correspondències. Anàlisi de conglomerats.
Explicatives	Anàlisi de regressió múltiple. Anàlisi multivariàble de la variància. Anàlisi discriminant. Anàlisi d'equacions estructurals.

Les tècniques que anomenem **descriptives** s'utilitzen normalment en contextos no experimentals i la seva finalitat és intentar reduir la quantitat de variables inicial, en un nombre més petit de dimensions rellevants.

Les tècniques **explicatives** són més pròpies del context experimental; pretenen veure l'impacte d'un conjunt de variables independents (predictores) sobre una o diverses variables dependents.

En general, les diferents tècniques intenten buscar un model de relació de totes les variables disponibles, reduint les que no siguin significatives i, per tant, obtenint un model tan parsimoniós (simple) com sigui possible.

Hi ha diferents models i tècniques. Bàsicament són les següents:

1) Model lineal general

a) Regressió múltiple. Ens serveix per a predir una variable dependent o criteri, a partir d'una o diverses variables predictores o independents.

b) Anàlisi multivariàble de la variància (MANOVA). Representa una extensió de l'anàlisi de la variància (ANOVA); en aquest cas disposarem de dues o més variables independents o factors. Si ens fixem en l'exemple que hem vist en l'aplicació de l'ANOVA veurem la valoració dels subjectes en funció de la nota obtinguda; si incorporem el gènere dels subjectes, el MANOVA ens permetrà veure si hi ha diferències en la valoració en funció de la nota, en funció del sexe, i en funció de la interacció de les dues variables.

En la següent taula veiem el tipus de variables sobre les quals apliquem aquestes tècniques:

Tècnica	Variable dependent	Variable/s independent/s
Regressió múltiple	Quantitativa	Quantitativa/es
Anàlisi de la variància	Quantitativa	Qualitativa/es

2) Tècniques factorials

a) Anàlisi en components principals. És una tècnica estadística de reducció de la dimensió de les dades que s'utilitza per a explicar la variabilitat entre les variables observades en funció d'un nombre de variables latents més petit (no observades) anomenades *components* o *factors*.

b) Anàlisi de correspondències. És una tècnica similar a l'anàlisi en components principals però aplicada a variables categòriques.

c) Anàlisi discriminant. S'aplica aquesta tècnica quan disposem d'una variable dependent categòrica (dos o més grups) que volem predir a partir d'un conjunt de variables quantitatives. En aquest cas, les variables quantitatives s'agrupen en combinació lineal en el que s'anomena la *funció discriminant*, que ens ajudarà a predir la pertinença dels subjectes a cada un dels grups.

En la següent taula veiem el tipus de variables sobre les quals apliquem aquestes tècniques i la reducció que n'obtenim.

Tècnica	Variables	Formació de grups
Anàlisi en components principals	Quantitatives	Variables
Anàlisi de correspondències	Qualitatives	Categories de variables
Anàlisi discriminant	Quantitatives	Categories de variable (VD)

3) Tècniques de classificació

a) Anàlisi de conglomerats (clúster). Classifica una mostra (es pot utilitzar tant des de la perspectiva dels subjectes agrupant-los, com de les variables) en un nombre més petit de grups de manera que, o bé agruparem els subjectes en funció de les seves similituds (similar al discriminant però sense saber *a priori* quins grups tenim), o bé agruparem les diferents variables reduint la dimensió de les variables originals.

La taula ens mostra el tipus de variables i els grups formats en la tècnica de l'anàlisi clúster.

Tècnica	Variables	Formació de grups
Anàlisi de conglomerats	Quantitatives i/o qualitatives	Variables i/o subjectes

4) Tècniques estructurals

a) **Anàlisi d'equacions estructurals.** Analitza les relacions existents entre un grup de variables, representades per sistemes d'equacions simultànies en les quals suposem que algunes (anomenades *constructes*) es mesuren (amb error) a partir de variables observables anomenades *indicadors*. Els models s'especifiquen en dues parts: un model estructural que especifica les relacions de dependència entre els constructes latents i un model de mesura que especifica com els indicadors es relacionen amb els seus corresponents constructes.

La següent taula presenta el tipus de variables sobre les quals apliquem l'anàlisi de les equacions estructurals.

Tècnica	Variable dependent	Variables independents
Anàlisi d'equacions estructurals	Quantitativa	Quantitatives i/o qualitatives

En aquest apartat hem intentat fer una primera aproximació a les principals tècniques, no totes, d'anàlisi multivariable. Qualsevol manual especialitzat ens permetrà aprofundir en les tècniques que podem necessitar. També hem de tenir en compte que totes aquestes tècniques tenen uns requisits (supòsits) d'aplicació, i per això hem de ser curiosos a l'hora de comprovar que es compleixen realment. Això últim és especialment adient perquè, atès que els programes estadístics són fàcils d'aplicar, molts cops s'apliquen aquestes tècniques sense comprovar-ne els supòsits o condicions d'aplicació.

Lectura recomanada

Per a aprofundir sobre les tècniques d'anàlisi multivariable podeu consultar la següent obra:

J. P. Lévy; J. Varela (ed.) (2003). *Anàlisi multivariable para las ciencias sociales*. Madrid: McGraw-Hill.

Bibliografia

- Arnau, J.** (1996a). *Model general d'investigació psicològica*. Barcelona: UOC.
- Arnau, J.** (1996b). *Quantificació de les observacions: escales de mesura*. Barcelona: UOC.
- Cabo, I. de** (1996). *Metodologia de les ciències socials i humanes*. Barcelona: UOC.
- Cook, T. D.; Campbell, D. T.** (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.
- Coscolluela, A.; Fornieles, A.; Turbany, J.** (2008). *Tècniques d'anàlisi de dades quantitatives*. Barcelona: UOC.
- Doménech, J. M.** (1995). "El modelo de investigación científica". *Métodos estadísticos en ciencias de la salud*, 4. Barcelona: Signo.
- Gómez, J.** (1996). *Construcció d'instruments de mesura*. Barcelona: UOC.
- Guàrdia, J.; Freixa, M.; Però, M.; Turbany, J.** (2007). *Análisis de datos en psicología*. Madrid: Delta.
- Jáñez, L.** (1989). *Fundamentos de psicología matemática*. Madrid: Pirámide.
- Kenny, D. A.** (1979). *Correlation and Causality*. Nova York: Wiley.
- León, O.; Montero, I.** (2003). *Métodos de investigación* (3a. ed.). Madrid: McGraw Hill.
- León, O.; Montero, I.** (2009). *Mètodes d'investigació qualitativa*. Barcelona: UOC.
- Lévy, J. P.; Varela, J.** (ed.) (2003). *Análisis multivariable para las ciencias sociales*. Madrid: McGrawHill.
- Morales, P.** (1988). *Medición de actitudes en psicología y educación*. San Sebastián: Tarttalo.
- Peña, D.** (2002). *Regresión y diseño de experimentos*. Madrid: Alianza.
- Salvador, F.** (1996). *Model general d'investigació psicològica*. Barcelona: UOC.
- Sierra Bravo, R.** (1985). *Técnicas de investigación social: Teoría y ejercicios*. Madrid: Paraninfo.
- Shuman, H.; Presser, S.** (1981). *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wordin and Context*. Nova York: Academic Press.
- Summers, G.** (1982). *Medición de actitudes*. Mèxic, DF: Trillas.
- Viladrich, M. C.; Doval, E. Prat, R.; Vall-Ilovera, M.** (1997). *Psicometria* (2a. ed.). Terrassa: Cardellach.

