Tema 1 – El procés d'investigació científica i l'estadística

1. L'estadística

• Té el seu origen en l'interés dels *Estats* per conèixer els recursos amb què comptaven: nombre d'habitants, edat, tipus de treball que realitzaven, condicions de vida, propietats, etc. Ja a l'Egipte antic i durant l'Imperi Romà es poden trobar manifestacions d'aquest interès per procediments que permeteren obtenir dades estadístiques (*d'estat*). En qualsevol cas, serà en el període de domini napoleònic quan es produïsca el salt més substancial en aquest afany per disposar d'informació estadística. Posteriorment, aquest interès ha crescut i s'ha estès, de manera que ha anat abastant altres nivells d'anàlisi més moleculars (regions, ciutats, barris, col·legis, grups concrets de persones...) o també, a vegades, més molars (grups de nacions, continents, el món...). D'altra banda, l'interès de l'anàlisi estadística s'ha ampliat a tota mena de variables més enllà de les que típicament cobreix el cens, l'hereu directe d'aquell interès històric en què es pot trobar l'origen de l'estadística.

• Encara que l'estadística aparega originalment associada a l'interès per conèixer més sobre els habitants de nacions, regions o altres tipus d'agrupacions, prompte va transcendir la seua aplicació a altres unitats d'observació que no necessàriament eren persones, però en les quals era també habitual la recollida de volums amplis de dades dels quals es volia extraure algun tipus d'informació d'interès –aquest és el cas, per exemple, de la investigació astronòmica (unitat d'observació: estreles, planetes, constel·lacions…) o, més recent en el temps, la investigació sobre productes de consum (unitat d'observació: aliments, ordinadors, cotxes…).

• L'interès per l'estadística ha donat lloc al desenvolupament d'una sèrie de coneixements i procediments orientats a satisfer dos grans nivells de competències: (1) resumir la informació recollida, habitualment quantiosa, d'una manera que resulte més comprensible i permeta prendre decisions útils; i (2) inferir sobre una població nombrosa en la seua grandària, a partir d'un subconjunt reduït de membres d'aquesta població. Totes dues necessitats han donat lloc a les dues grans branques que tradicionalment se solen diferenciar dins del camp de l'estadística: l'estadística descriptiva i l'estadística inferencial. L'aplicació d'ambdues no és excloent sinó, amb freqüència, complementària.

• Una altra diferenciació també tradicional és la d'estadística teòrica *versus* estadística aplicada, la primera més dirigida al desenvolupament i estudi de mètodes formalment vàlids per a fer estadística, mentre que la segona estaria orientada a l'aplicació d'aquests mètodes a camps d'estudi concrets. L'anàlisi estadística o anàlisi de dades són termes que amb freqüència s'usen amb el mateix sentit que el d'estadística aplicada.

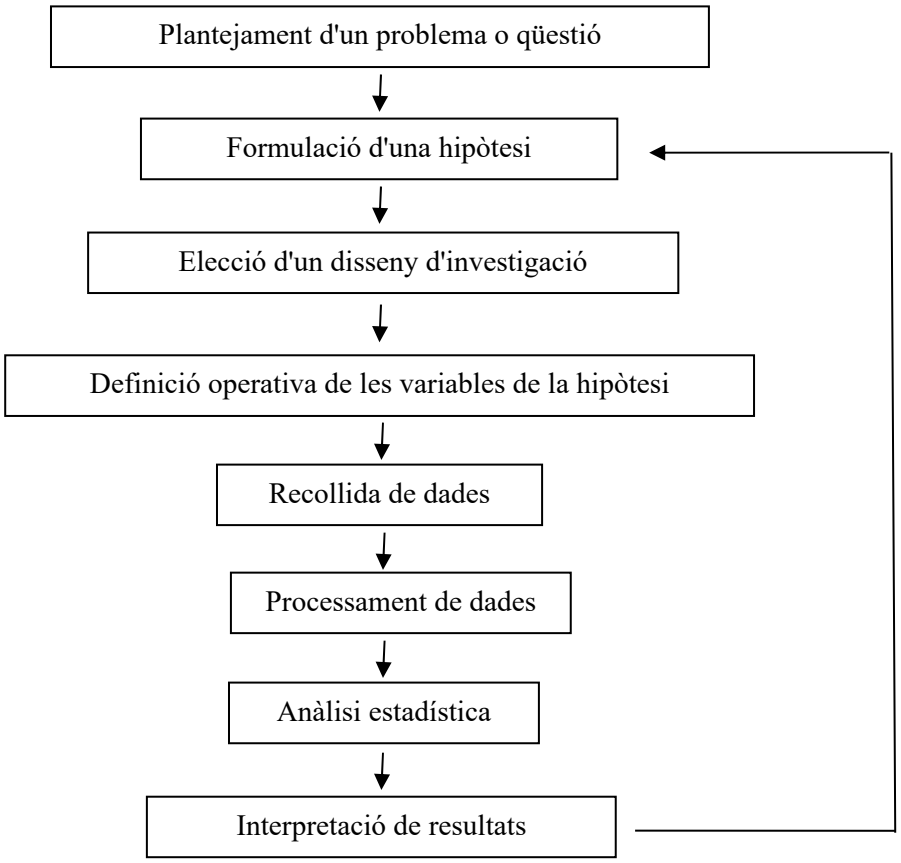
2. El mètode científic i l'anàlisi estadística

• L'estadística fa un paper molt important dins del conegut com a mètode científic, el qual utilitza la psicologia, igual com altres àrees de coneixement, en el seu desenvolupament com a ciència.

• El mètode científic representa una estratègia ordenada i sistemàtica d'actuació en la realització d'un estudi o investigació. És precisament l'aplicació d'aquest el que permet dotar a una investigació el qualificatiu de *científica*. Les següents són característiques bàsiques inherents al mètode científic:

(1) s'aplica a qüestions empíricament contrastables;
(2) cerca resultats lliures de biaix i d'aplicació tan general com siga possible;
(3) dona peu a la replicació dels estudis basats en la seua aplicació;
(4) busca explicacions parsimonioses.

• Passos que es poden diferenciar en l'aplicació del mètode científic (Navas, 2001):



Exemple il·lustratiu de l'aplicació d'algunes de les etapes del mètode científic en la cerca d'una resposta a la qüestió: "Per quina raó les persones difereixen en el seu nivell de tolerància al dolor?"

– Formulació d'una hipòtesi:

Exemples d'aquesta hipòtesi en funció del marc teòric de referència considerat:

- Sociologia/antropologia: "Les diferències depenen de factors culturals."
- Psicologia social: "Les diferències depenen del context grupal al qual pertany el subjecte."
- Psicologia de la personalitat: "Les diferències van en funció d'un tret de personalitat, la introversió-extraversió, de manera que les persones introvertides tenen més tolerància al dolor."

(Assumim com a nostra l'última d'aquestes hipòtesis a fi de continuar.)

– Elecció d'un disseny d'investigació:

- De qui es recollirà informació? (→ Disseny mostral)
- Quin procés lògic se seguirà per a recollir la informació empírica que ens permeta contrastar la nostra hipòtesi? (→ Disseny de recollida de les dades)
- Quines estratègies s'aplicaran per a controlar possibles variables estranyes (aquelles que poden afectar la variable objecte d'interés, però l'efecte del qual en realitat no ens interessa)?

– Definició operativa de les variables implicades en la hipòtesi:

- De qui, en concret, es recollirà informació empírica?
- Quin tipus d'instrument o procediment utilitzarem per a obtenir les dades? (com mesurar si una persona és extravertida o introvertida?; com mesurar la tolerància al dolor?)
- És fiable i vàlid el procediment que utilitzarem en la mesura de les variables?

– Recollida de les dades

– Processament i anàlisi estadística de les dades:

Suposem que, en el nostre exemple, obtenim el següent resum de les dades recollides (→ *estadística descriptiva*):

Mitjana aritmètica de tolerància al dolor de les persones introvertides = 8,1

Mitjana aritmètica de tolerància al dolor de les persones extravertides = 6,7

(La tolerància al dolor oscil·la entre 1 [baixa tolerància] i 10 [alta tolerància])

Podem considerar que hi ha diferències estadísticament significatives entre aquestes mitjanes a nivell poblacional? (→ *estadística inferencial*)

– Interpretació dels resultats:

- Aquests resultats donen suport a la hipòtesi inicial?
- Quines conseqüències es deriven dels resultats obtinguts?

• En síntesi, l'anàlisi estadística ens permetrà satisfer l'objectiu de resumir i transmetre d'una manera comprensible la informació procedent de dades empíriques (estadística descriptiva) així com, quan siga oportú, generalitzar a partir de la informació recollida d'un conjunt reduït de subjectes a una població més àmplia a la qual aquests representen (estadística inferencial).

3. Per què l'estadística en les ciències de la salut?

- Perquè ens proporcionarà un tipus de coneixements i competències que afavoreixen el pensament analític i crític.
- Perquè ens capacitarà per a realitzar estudis (investigacions) en els quals poder buscar la resposta a preguntes que ens sorgisquen o bé posar a prova idees que tinguem (hipòtesis).
- Perquè ens permetrà poder llegir, entendre i valorar informació especialitzada sobre temes psicològics (articles en revistes de divulgació científica, informes d'investigació, llibres, etc.), per a la nostra formació i, en el futur, la nostra especialització i actualització com a professionals.

Per posar-ne un exemple, encara que no es tracte més que d'un fragment dels resultats d'un article d'investigació, hem de poder interpretar la informació que s'hi conté.

"Per comprovar si existeixen diferències estadísticament significatives en l'apoderament i en el rendiment acadèmic en funció de tenir o no discapacitat i en funció del sexe es va utilitzar la t de Student per a dues mostres independents. Per a les comparacions de més de dos grups (edat) es va utilitzar l'ANOVA. Prèviament a això, es va comprovar que es complien els supòsits d'homocedasticitat, normalitat de les distribucions i independència de variables. Això es va comprovar a través de la prova de Levene, de Kolmogorov-Smirnov i khi-quadrat, respectivament. Així mateix, es va calcular la grandària de l'efecte (diferència *d*; Cohen, 1988), que indica si la magnitud de les diferències trobades és xicoteta, moderada o gran." (Suriá i Villegas, 2020)

4. Alguns conceptes bàsics d'estadística

• Un cas o unitat d'observació és cadascun dels elements sobre els quals es desitja recollir informació en un determinat estudi. De manera sinònima són utilitzats també amb freqüència els termes *entitat*, *participant* i *subjecte*. Aquesta última denominació és apropiada quan les unitats d'observació són persones individuals, la qual cosa, encara que freqüent, no és sempre el cas: d'una banda, les unitats d'observació poden ser parelles (p. ex., mare-fill) o grups (p. ex., famílies, associacions, col·legis, empreses); d'altra banda, poden ser animals (com és comú en la investigació psicofisiològica) o objectes (p. ex., joguets, anuncis radiofònics...).

• Una variable és el conjunt de valors resultants de mesurar una característica (propietat, atribut) en les diferents unitats d'observació objecte d'estudi.

• Una dada (valor observat, observació) és un valor que proporciona informació d'un cas per a una variable concreta. Utilitzarem les claus { }, per enumerar les dades observades en una determinada variable. Per exemple, per a una hipotètica variable *Z,* de la qual es van obtenir dades de 9 casos, tindríem:

$$Z\text{: } \{1; 3; 2; 2; 5; 0; 3; 2; 0\}$$

Exercici 1: En un estudi amb un grup de 45 persones que havien sigut tractades psicològicament per problemes d'ansietat, s'hi va preguntar quantes vegades havien patit un atac d'ansietat des que va acabar el tractament rebut. També se'ls va preguntar si consideraven que havia millorat la seua

qualitat de vida arran del tractament psicològic que van rebre. Qüestions: (a) Quants casos i quantes variables apareixen implicats en aquest estudi?; (b) Quantes dades s'hauran obtingudes en la recollida de dades?

• En la literatura estadística sol utilitzar-se el terme modalitats per a fer referència a cadascun dels diferents valors que adopta una variable. Per exemple, si la variable relativa a la millora de la qualitat de vida de l'exercici anterior va ser contestada "Sí" per 31 persones i "No" per 14, tindríem que les modalitats d'aquesta variable són dues: "Sí" i "No". En determinats contextos se sol utilitzar el terme nivells per a fer referència a les modalitats d'una variable. Nosaltres utilitzarem els claudàtors [ ], per a indicar les modalitats d'una variable. Per exemple, per a la variable *Z* d'abans:

Z: [0; 1; 2; 3; 5]

• Existeixen diferents maneres d'organitzar les dades recollides en un estudi, les quals és comú que siguen quantioses, atès que és freqüent obtenir dades de diverses variables per a un conjunt ampli de casos. Aquestes formes d'organització de les dades es denominen estructures de dades, de les quals la més utilitzada en la pràctica és la coneguda com a taula de dades (també denominada *matriu de dades*): es tracta d'una organització bidimensional en què les files representen les unitats d'observació i les columnes les variables, de manera que l'encreuament d'una fila i una columna qualssevol de la taula constitueix el valor observat (dada) corresponent a un cas concret en una determinada variable.

4.1. Tipus de variables

S'han plantejat diferents classificacions de les variables en funció del criteri considerat en la seua categorització. A continuació es presenten 3 d'aquestes tipologies que resulten especialment rellevants per a comprendre les implicacions pràctiques del concepte de variable.

(a) En funció de com les variables són mesurades –això és, de l'escala de mesura utilitzada en l'obtenció de les dades de la variable–, es poden diferenciar 3 tipus de variables:

• Variables quantitatives o numèriques: aquelles en què els valors resultants del mesurament són números que indiquen el grau o quantitat del que s'està mesurant. Les variables quantitatives es caracteritzen per tenir unitats de mesura.

Exemple: la variable "Pes (en grams)" mesurada en una ventrada de 8 rates (*X*).

X: {24,2; 39,6; 31,2; 27,8; 29,5; 36,5; 48,4; 42,0}

La unitat de mesura en aquest cas és el gram i les dades recollides ens mostren, per a cada rata, el nombre d'unitats de mesura que caracteritza a cadascuna d'elles.

De les variables quantitatives es fa una diferenciació addicional en funció que siguen contínues o discretes. Les variables quantitatives contínues són variables que poden prendre qualsevol valor numèric, això és, entre qualsevol parell de valors, pot donar-se un valor numèric intermedi. La variable "Temps emprat a completar una tasca orientada a avaluar la coordinació motriu" és un exemple de variable quantitativa contínua. Les variables quantitatives discretes només poden prendre valors concrets, per exemple, la variable "Nombre de fills". Així, una família pot tenir 1, 2, 3 fills..., però no pot tenir-ne 1,5.

En qualsevol cas, cal tenir en compte que la precisió limitada dels instruments de mesura provoca que, en la pràctica, totes les variables quantitatives siguen mesurades d'una manera discreta encara que algunes, per la seua naturalesa, siguen en realitat variables quantitatives contínues.

• Variables quasi-quantitatives o ordinals: aquelles en què els valors observats no són indicatius més que de l'ordre o posició de les unitats d'observació en el que s'estiga mesurant. La dada corresponent a un determinat cas tan sols representa en què grau s'és major o menor que un altre cas que té, respectivament, un valor inferior o superior en allò que s'estiga mesurant.

Exemple: la variable amb les dades recollides en un grup de 121 persones a partir de la següent pregunta d'un test: "Ansietat que sent quan es troba amb molta gent al voltant" (X).

X: {4; 2; 1; 2; 2; 3; 1; 3; 2; 4; 1; 3; 1; 2; 3; 2; 1; 1; 4; 1; 3...}

Les alternatives de resposta a aquesta qüestió eren: Gens; Poca; Bastant; Molta.

Codificació: 1=Gens; 2=Poca; 3=Bastant; 4=Molta.

• Variables categòriques (qualitatives, nominals): aquelles en què els valors no aporten cap informació de magnitud ni d'ordre, tan sols diferencien als casos en diferents categories de pertinença. Una classificació addicional de les variables categòriques diferencia a aquestes entre dicotòmiques (dos valors possibles) i politòmiques (més de 2 modalitats).

Exemple: la variable "Estat civil", recollida en un total de 50 persones de la ciutat de Castelló ($N = 50$):

X: {0; 0; 1; 2; 2; 0; 1; 3; 2; 0; 1; 0; 1; 2; 0; 2; 1; 1; 0; 1; 0...}

Codificació: 0=solter/a; 1=casat/ada; 2:=separat/ada o divorciat/ada; 3= vidu/a.

• A tenir en compte en relació amb els tres tipus de variables:

- Una variable no és d'un tipus o un altre *per se* sinó que dependrà de la manera en què és mesurada (p. ex., les variables "Edat" o "Consum de tabac" poden ser mesurades utilitzant diferents escales que donen lloc a variables de diferents tipus).
- El tipus de variable (quantitativa, quasi-quantitativa, categòrica) és determinant en la selecció del procediment estadístic a aplicar.
- Una característica (propietat, atribut) que adopta els mateixos valors per a totes les entitats es denomina constant. No es parla en aquest cas de variable, atès que no hi ha variabilitat en els valors observats. Cal destacar que el mesurament d'una mateixa característica pot donar lloc a una variable o a una constant en funció de l'estudi de què es tracte (p. ex., el 'sexe' en un estudi sobre l'ansietat i en un estudi sobre la depressió postpart, respectivament).

Exercici 2: De quin tipus són les següents variables en funció de la seua escala de mesura?

a) Temps en segons invertit a recórrer un laberint.
b) Nombre de cares reconegudes en una sèrie de 100 imatges de cares, de les quals 50 havien sigut presentades una hora abans.
c) Nivell d'ingesta alcohòlica mesurada com: Molt alta, Alta, Mitjana, Baixa i Nul·la.
d) Classificació d'un grup de subjectes en funció del seu lloc de residència (Urbana, Rural).
e) Capacitat de fer amics de l'alumnat, mesurada com l'ordenació del mateix realitzada pel professorat (1. per al més capaç, 2 per al següent, i així).
f) Quocient intel·lectual d'un grup de xiquets/es mesurat a partir de l'aplicació del test d'intel·ligència WISC (*Wechsler Intelligence Scale for Children*).
g) Dosi d'un fàrmac mesurada en mil·lilitres.
h) Tipus de col·legi al qual s'assisteix (Públic, Privat, Concertat).
i) Classificació feta per una psicòloga d'un grup de pacients que realitzen teràpia de grup en funció del grau de millora (1. per al que més ha millorat, 2 per al següent, i així).
j) Nacionalitat.
k) Seguretat laboral de les empreses mesurada com el nombre d'accidents laborals ocorreguts en l'empresa durant l'últim any.
l) Grup al qual es pertany en la realització d'una investigació (tractament vs. control).
m) Nombre de membres en la unitat familiar.
n) Rendiment acadèmic mesurat a partir de la puntuació en un examen.
o) Classificació de pel·lícules en funció del seu gènere (Comèdia, Terror, Musical…).
p) Estat d'ànim mesurat a través de la següent escala: Positiu, Neutre, Negatiu.

(b) En funció de què es mesura:

Aquesta tipologia diferencia entre aquelles variables que són pròpiament l'objecte d'interés del nostre camp de coneixement i aquelles que, encara que no de naturalesa psicològica, solen ser considerades en un gran nombre d'investigacions psicològiques atès l'interès que sol tenir plantejar anàlisis diferencials en funció d'aquestes variables.

• Variables psicològiques: són variables que fan referència a característiques relatives a la personalitat, la intel·ligència, hàbits, aptituds, actituds i habilitats, entre altres facetes de la psicologia.

• Variables sociodemogràfiques: aquelles que fan referència a aspectes demogràfics i sociològics com ara l'edat, el sexe, la nacionalitat, el lloc de residència, el nivell d'estudis aconseguit, la llengua principal que es parla, l'estat civil, el nivell d'ingressos econòmic, etc.

(c) En funció de quin rol exerceixen en el disseny de la investigació:

• És freqüent trobar estudis en què es proposen hipòtesis sobre la relació entre 2 variables (hipòtesis bivariades) i, dins d'aquestes, hipòtesis en què es planteja l'existència d'influència (o efecte) d'una variable sobre una altra –per exemple, una investigació pot incloure una hipòtesi sobre la influència de la intel·ligència emocional en la conducta solidària. En aquests casos es parla d'una variable explicativa que es planteja que és la causa d'una variable de resposta. Cal assenyalar que en la literatura apareix certa diversitat en la forma en què aquests dos tipus de variables són nomenades, tal com es posa de manifest en el següent esquema:

Variable explicativa		Variable de resposta
Variable independent (*VI*)	$\Rightarrow$	Variable dependent (*VD*)
Variable predictora		Variable criteri
Variable exògena		Variable endògena
Variable *X*		Variable *Y*

• Cal destacar que una mateixa variable pot aparèixer com a explicativa en un estudi, mentre que pot ser variable de resposta en un altre diferent. Per exemple, una variable com la intel·ligència emocional és la variable explicativa en l'exemple anterior de la conducta solidària, mentre que seria la variable de resposta en un estudi en què s'analitzara l'efecte de l'estil educatiu en la infància sobre la intel·ligència emocional.

• Si en una hipòtesi apareixen implicades més de 2 variables, es fa referència a la mateixa com a hipòtesi multivariada, Encara que poden donar-se diferents tipus de combinacions, un cas comú

d'hipòtesi multivariada és aquell en què una variable adopta el paper de variable de resposta, mentre que la resta de variables són variables explicatives.

• És comú que, en estudis en què es desitja analitzar la influència d'una variable sobre una altra, els valors de la primera, la variable explicativa, siguen assignats als subjectes per part de l'investigador/o investigadora. Aquest tipus de variable explicativa en què, en realitat, no hi ha un procés de mesura a partir del qual s'obtinguen les dades sinó que és el responsable de la investigació qui determina els valors que tindran els subjectes en aqueixa variable es denomina variable manipulada.

> **Exemple**: quan es vol estudiar l'efecte de diferents dosis d'una determinada substància psicoactiva sobre la conducta, una estratègia habitual és que s'apliquen diferents dosis d'aquesta, establides a priori per l'investigador (per exemple, 100, 200 i 300 mg.), a diferents grups de subjectes. Així, la variable "Dosi de la substància administrada" seria una variable manipulada. Un altre exemple: la variable "Grup en el qual es participa en un estudi", en què es vol comparar l'eficàcia d'una determinada teràpia psicològica (Grup A) enfront d'una altra (Grup B). Els subjectes són assignats a un grup o un altre per l'investigador.

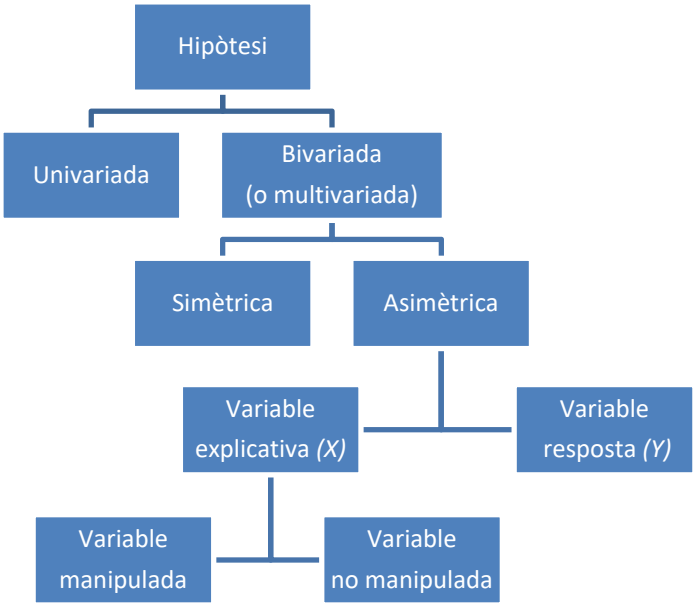
• Un altre tipus d'estudi també habitual en la investigació psicològica és aquell en què es planteja una hipòtesi sobre la relació entre dues (o més) variables, sense que s'establisca a priori que una variable siga explicativa i l'altra variable de resposta. En aquests casos es parla de relació simètrica entre les variables, enfront de la relació asimètrica que es donava en els exemples previs.

> **Exemple** de relació simètrica: un estudi en què aparegueren implicades les variables "Autoestima" i "Habilitats socials" en què l'investigador, per no tenir clara una relació de causalitat entre totes dues (pot donar-se en els dos sentits), es limitara a plantejar una hipòtesi d'existència de relació entre les dues variables.

• De la mateixa manera que no és aplicable la distinció entre variable explicativa i variable de resposta si ens trobem amb un cas de relació simètrica entre les variables, tampoc ho serà en aquells estudis que impliquen una única variable (*hipòtesi univariada*).

> **Exemple**: Un estudi en què es vulga comprovar la hipòtesi "la majoria de les i els joves d'entre 25 i 30 anys estan solters/es" implica l'obtenció de dades d'una única variable, l'"Estat civil", per la qual cosa en aquest estudi no té sentit parlar de variable predictora o variable criteri.

• A manera de síntesi gràfica d'alguns dels conceptes presentats fins ara:



Exercici 3: Per a cadascuna de les següents hipòtesis:

- ❑ Apliqueu l'arbre de decisió superior a fi de definir de quin tipus són les següents hipòtesis i les variables implicades en aquestes (excepte la branca de variable manipulada vs. no manipulada).
- ❑ Quan hi haja informació suficient (o es puga suposar), indiqueu de quin tipus són les variables implicades en funció de l'escala de mesura.

a) En delinqüents juvenils, la gravetat dels delictes comesos és inversament proporcional a l'edat en què van començar a delinquir.
b) Entre l'estudiantat que abandona la Universitat, la proporció d'homes i de dones és similar.
c) El nivell educatiu és un factor causal de l'aparició de demència en la gent gran.
d) Més del 50 % dels conflictes agressius entre xiquets/es en edat preescolar es relacionen amb la possessió d'un objecte.
e) La majoria dels xiquets i les xiquetes de 4 anys d'edat són capaces de nomenar els colors primaris.
f) Per a la joventut europea, l'atur és el problema social més important.
g) La situació dels pares influeix sobre l'estabilitat emocional de les i els fills, així les i els fills de pares casats són més estables emocionalment que les i els fills de pares separats i, aquests últims, més que les i els fills de pares solters.
h) L'ensenyament multimèdia de l'anglès en l'ESO és més efectiva que l'ensenyament tradicional.
i) La sobrecàrrega laboral és un factor desencadenant de l'aparició de la síndrome d'esgotament professional (*burn-out*).
j) L'absentisme laboral és una conseqüència rellevant de la síndrome d'esgotament professional.
k) Dos de cada tres fumadors/es a Espanya volen deixar de fumar.

l) Vora el 10 % de les persones treballadores europees del sector de serveis ha patit assetjament psicològic en el treball (*burn-out*) durant l'últim any.

m) La proporció de votants del PSOE en les pròximes eleccions generals a Espanya serà ≥ 0.30.

n) La integració dels xiquets i les xiquetes immigrants en els col·legis d'educació primària ve determinada per l'àmbit d'escolarització, i és millor en els col·legis públics que en els privats.

o) El benestar psicològic de la gent gran es troba afectat per l'entorn en què es viu, i és millor si es viu amb la família que si es viu en una residència geriàtrica.

p) Les qualificacions obtingudes en l'assignatura d'Estadística estan directament relacionades amb les obtingudes en Psicologia Social.

q) El nou fàrmac *A* contra l'ansietat ofereix millors resultats que els fàrmacs *B* i *C*.

5. Estadística descriptiva i estadística inferencial

• Una diferenciació tradicional en el camp de l'estadística ha sigut la que distingeix entre, d'una banda, l'interès d'aquesta disciplina per resumir les dades recollides d'una forma que resulte informativa, comprensible i permeta prendre decisions útils (estadística descriptiva) i, d'altra banda, l'interès per inferir sobre una població nombrosa, a partir d'un subconjunt reduït de membres d'aqueixa població (estadística inferencial). En la pràctica, l'aplicació d'ambdues no és excloent sinó, amb freqüència, complementària. Associada a aquesta distinció en el camp de l'estadística es troba la diferenciació entre les dues parelles de conceptes que, a continuació, es presenten.

5.1. Població i mostra

• "Se llama población estadística al conjunto de todos los elementos que cumplen una o diversas características o propiedades" (Botella *et al*., 2001). Una altra proposta de definició: Conjunt de casos objecte d'interés en un estudi.

• Com a conjunt, els seus elements tindran una o més característiques en comú que són les que determinaran la seua pertinença a aqueix conjunt. L'especificació de la població en un estudi ha d'expressar amb precisió aqueixes característiques, perquè representen el criteri de pertinença a la població, cosa que permet destriar amb claredat qui i qui no forma part de la població objecte d'estudi.

Exemples de poblacions:

- La població de dones d'entre 25 i 35 anys de la Comunitat Valenciana (CV). Per a aquest cas, els criteris de pertinença serien: ser dona, tenir entre 25 i 35 anys i pertànyer a la CV. Possibles ambigüitats: entre 25 i 35 anys, tots dos inclusivament?; què vol dir pertànyer a la CV, viure-hi?

- Els estudiants del grup D de primer curs de la Facultat de Psicologia de la UVEG del present curs acadèmic.
- Els col·legis privats o concertats de la ciutat de València.

• La població que s'especifique com a objecte d'estudi determinarà de qui es recolliran dades (siga de tots els seus elements o d'una part d'ells) i, també, sobre qui recauran les conclusions derivades de l'estudi. És per això que la seua especificació representa un aspecte crucial en el disseny de qualsevol investigació.

• En la difusió dels resultats d'un estudi han de detallar-se les característiques de la població objecte d'estudi. Això no és sempre el més habitual en les notícies sobre estudis o investigacions difoses en mitjans de comunicació de masses (premsa, ràdio, TV...). En alguns casos, l'ocultació està motivada per la urgència d'espai i temps que aquests mitjans tenen per a informar, però aquesta urgència acaba convertint-se a vegades en una excusa per a ometre o amagar informació de manera malintencionada.

> Els espectadors o lectors experts seran escèptics davant un informe en què s'ometa aquesta informació; no obstant això, no ocorrerà el mateix amb altres receptors no experts, que poden creure que els resultats fan referència a una població àmplia quan, en realitat, pot ser que no siga el cas. És un deure ètic ser transparent en un aspecte tan crucial com és el de la descripció de les característiques de la població objecte d'estudi i dels participants en l'estudi, perquè això permetrà valorar la capacitat de generalització dels resultats obtinguts a la població.

• Una possible font de confusió sobre el concepte de població és que aquesta siga concebuda com un conjunt necessàriament nombrós d'elements. A tall d'exemple, la següent frase extreta del llibre d'estadística de Pardo i San Martín (2001): "Precisamente el hecho de que las poblaciones, en general, sean infinitas o estén formadas por un gran número de elementos, hace que la descripción exacta de sus propiedades sea un objectivo prácticamente inaccesible."

La grandària de la població dependrà de quin és l'objectiu de l'estudi que es plantege dur a terme i, més específicament, de sobre qui es desitge extraure conclusions. Per exemple, la grandària de la població pot ser més que reduïda en l'estudi de l'eficàcia d'un mètode determinat de lectoescriptura que tinga com a població d'interès simplement els escolars d'una aula d'un col·legi concret, com podria ser el cas d'un estudi en un col·legi en el qual estiguem desenvolupant el nostre treball com a psicòlegs.

• "Una muestra es un subconjunto de los elementos de una población" (Botella *et al*., 2001). Tal vegada s'hi podria afegir: ... sobre el qual obtenir la informació que ens permeta inferir alguna característica o característiques d'aquesta població.

• Avantatges de treballar amb mostres en comptes de fer-ho amb les respectives poblacions:
- Economia de recursos: òbviament, com menor siga el nombre de casos dels quals calga recollir dades, menor serà el treball implicat.
- Qualitat de les dades recollides: si no es té la pressió de recollir dades d'una gran quantitat de casos, es pot prestar més atenció a fer-ho amb major cura i rigor.

• Un avantatge que se sol atribuir a treballar amb una mostra és la derivada del problema d'accessibilitat als elements de la població. A títol il·lustratiu, Pardo i San Martín (2001) afirmen: "Las poblaciones que habitualmente interesa estudiar en Psicología son infinitas o son tan grandes que normalmente resulta muy difícil (si no imposible) tener acceso a todos sus elementos. Bajo estas circunstancias, es de las muestras de donde podemos obtener la información necesaria...". En realitat, el problema no seria tant la *gran dificultat* o *impossibilitat* per a tenir accés als elements d'una població nombrosa –llevat que es tracte de poblacions que no tinguen una localització tangible en l'espai o en el temps–, sinó més aviat com de costós que pot resultar accedir-hi.

• Anàlogament al que ocorre amb el concepte de població, una font de confusió bastant freqüent en considerar el concepte de mostra és associar-lo a un conjunt d'elements de grandària reduïda. A tall d'exemple, la següent afirmació en Pardo i San Martín (2001): "Al contrario de lo que ocurre con las poblaciones, que suelen ser conjuntos de elementos de gran tamaño, las muestras suelen ser conjuntos de elementos de tamaño reducido." En realitat, una mostra d'una població nombrosa pot tenir una grandària considerable. Així, la mostra corresponent a un estudi en què la població siguen els habitants d'un país com Espanya pot perfectament tenir una grandària de tres o quatre mil casos, la qual cosa no és una grandària reduïda.

5.2. Paràmetres i estadístics

• "Un parámetro es un valor numérico que describe una característica de una población" (Botella *et al*., 2001). Una altra proposta de definició: És el valor d'un índex estadístic (mitjana aritmètica, variància, proporció...) obtingut a partir de les dades d'una població, i que descriu alguna característica d'aquesta.

• "Un estadístico es un valor numérico que describe una característica de una muestra" (Botella *et al*., 2001). Una altra proposta de definició: És el valor d'un índex (estadístic) obtingut a partir de les

dades d'una mostra que ens permetrà inferir el valor d'alguna característica de la població a la qual pertany la mostra.

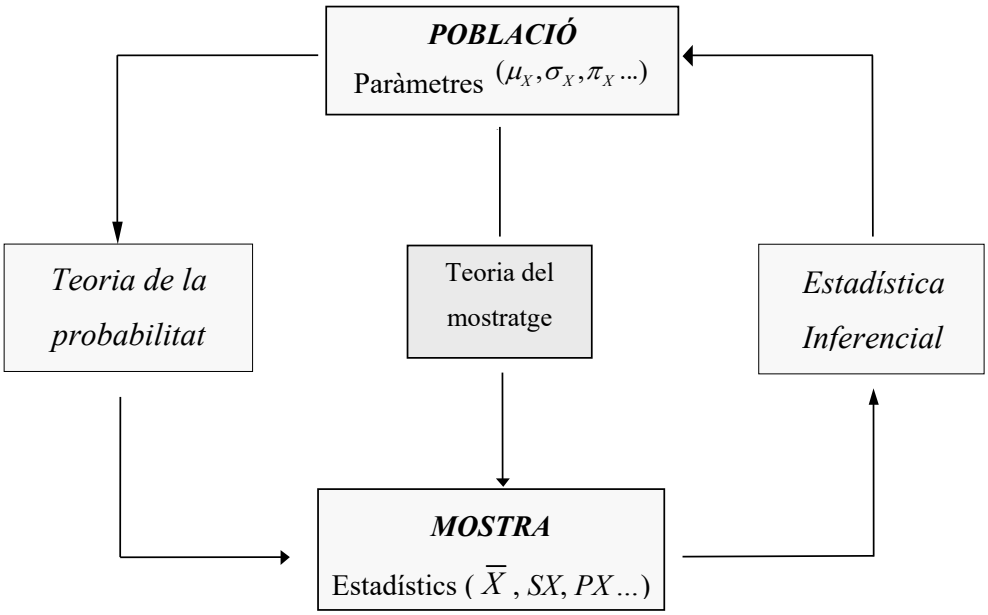
> **Exemple** amb l'índex estadístic de la mitjana aritmètica: si l'obtenim a partir de les dades d'una mostra, obtindrem un estadístic; mentre que si ho fem partir d'una població, un paràmetre.

• Mentre que el valor obtingut per a un paràmetre és una constant, l'obtingut per a un estadístic es pot concebre com una variable, atès que per a diferents mostres d'una mateixa població s'obtindran, normalment, diferents valors.

• És costum representar simbòlicament els paràmetres amb lletres gregues minúscules $(\mu_X, \sigma_X, \pi_X ...)$, mentre que els estadístics amb lletres llatines majúscules ($\bar{X}$, *Sx, Px* ...).

• Relacionant conceptes:

	Mostra	*Població*
Est. descriptiva	(→ Estadístics)	(→ Paràmetres)
Est. inferencial	(→ Estimacions de paràmetres)	×



Exercici 4: Per a cadascun dels següents estudis:

- Identifiqueu la població objecte d'estudi en cada cas.
- Indiqueu si les dades han sigut obtingudes d'una mostra d'aquesta població o de la població i, en conseqüència, dedueix si l'estudi seria únicament descriptiu, o bé descriptiu i inferencial.

a) Un empresari vol conèixer el nivell de satisfacció laboral dels 20 empleats de la seua empresa. Per a fer-ho els passa un qüestionari amb preguntes referides a la satisfacció respecte de diferents aspectes del seu treball.

b) A partir d'una mostra de dones embarassades, ens disposem a estudiar la relació entre el suport social rebut per la dona durant l'embaràs i l'estat de salut del bebé en nàixer (mesurat a partir de la puntuació en el test d'APGAR).

c) Un psicòleg d'un centre educatiu desitja augmentar la intel·ligència emocional de l'alumnat de primària d'aquest centre. Per a fer-ho aplica un programa d'intervenció i per a avaluar si ha sigut efectiu, compara el grau de conflictivitat de les interaccions de l'alumnat abans i després de l'aplicació del programa.

d) Volem conèixer la incidència de l'estrès laboral en la població del personal de la branca d'hostaleria de la Comunitat Valenciana. Així, seleccionem a l'atzar un grup d'empreses d'aquest sector de la Comunitat i enquestem als seus treballadors/es.

e) Un psicòleg de serveis socials d'un ajuntament vol conèixer l'actitud cap a la immigració dels habitants d'un barri conflictiu de la ciutat. Per a fer-ho realitza una enquesta telefònica a 100 persones censades en aquest barri.

f) Un psicòleg d'una empresa nacional vol avaluar si la implantació d'un nou sistema de qualitat en la producció ha augmentat la productivitat dels treballadors/es d'aquesta empresa. Per a fer-ho compara el nombre d'unitats fabricades abans i després de la implantació del sistema de qualitat en un subconjunt de centres d'aquesta empresa.

g) Un psicòleg escolar vol conèixer la satisfacció dels pares dels xiquets/es del col·legi amb les activitats extraescolars realitzades pels seus fills/es i convoca els pares i mares a una reunió i els passa un qüestionari sobre aquest tema.

h) Un psicòleg clínic vol avaluar l'eficàcia d'una teràpia de tipus conductual aplicada a un/a pacient amb claustrofòbia. Així, després de l'aplicació de la teràpia, mesura el seu nivell d'ansietat en estar dins d'un ascensor i el compara amb l'obtingut abans d'iniciar-se la teràpia.

i) Per estudiar si la integració dels xiquets/es amb discapacitats cognitives en els col·legis d'educació primària és millor en mitjans rurals que en mitjans urbans, se seleccionen a l'atzar dos col·legis (un de rural i un altre d'urbà) i pregunta als professors sobre el nivell d'integració dels xiquets/es amb alguna mena de discapacitat cognitiva a les aules.

6. Disseny d'investigació i estadística

• L'aplicació del mètode científic en la investigació s'ha plasmat a través de l'ús de diferents dissenys d'investigació en la pràctica. Una classificació clàssica dels dissenys d'investigació és la que diferencia 3 grans categories que varien en el nivell de control intern aplicat en el disseny de la investigació. Com més gran siga aquest control intern (grau d'intervenció en la manera de dur a terme l'estudi), amb major seguretat es pot arribar a afirmar que les diferències dels subjectes en la variable de resposta es deuen a la variable explicativa i no a altres variables (variables estranyes). En terminologia científica es parla d'una major validesa interna per a fer referència a aquest major control de totes les variables implicades en un estudi, de manera que les conclusions del mateix siguen tan robustes com siga possible.

• No obstant això, el major control intern sol estar associat a majors dificultats per a la realització de l'estudi, així com a estudis més artificials i allunyats de la realitat i, per tant, amb una menor representativitat del context al qual volem generalitzar les conclusions. En terminologia científica s'utilitza el terme validesa externa per a fer referència a la capacitat de generalitzar els resultats d'un estudi a altres contextos, a altres subjectes, i/o a altres moments temporals diferents dels presents en l'estudi realitzat. És desitjable que qualsevol investigació tinga tan alta validesa interna com externa, si bé és freqüent que tots dos aspectes es contraposen en la seua consecució en la pràctica.

• Classificació dels principals dissenys d'investigació i alguns dels seus trets bàsics:

– El disseny d'investigació experimental:
  - La variable explicativa és manipulada per l'investigador/a, qui definirà a priori els valors que aquesta pot prendre (→ variable manipulada).
  - L'investigador controla, en la mesura que siga possible, les variables estranyes, això és, variables que, sense ser la variable explicativa, poden tenir algun tipus d'efecte sobre la variable de resposta.
  - Assignació aleatòria dels participants en l'estudi als diferents subgrups definits per la variable explicativa.

– El disseny d'investigació quasi-experimental:
  - Les dues mateixes primeres característiques assenyalades per al mètode experimental.
  - Assignació no aleatòria dels participants als diferents subgrups definits per la variable explicativa, sinó determinada per condicionants pràctics com, per exemple, l'existència de grups naturals als quals assignar els diferents valors de la variable explicativa.

– El disseny d'investigació no experimental:

- No hi ha manipulació de la variable explicativa, els valors d'aquesta venen ja donats.

Dins d'aquesta última categoria se solen enquadrar dos tipus de dissenys d'investigació àmpliament utilitzats en l'àmbit de la investigació en les ciències del comportament: l'observacional i el selectiu (o d'enquestes).

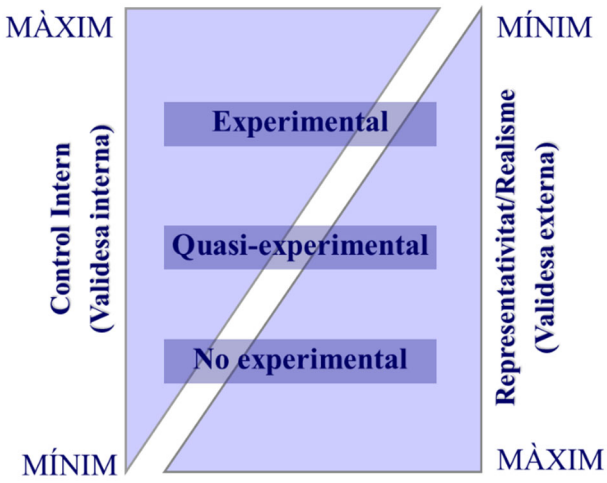


Figura resum de les característiques principals dels dissenys d'investigació (adaptat de Portell, Vives i Boixadós, 2003)

• Cal subratllar que podem trobar bastants exemples d'estudis en l'àmbit de la investigació psicològica en els quals no és possible manipular la variable explicativa i que, per tant, impliquen l'aplicació d'un disseny d'investigació no experimental. Aquest és el cas de tots els estudis en què la variable explicativa és alguna característica que els subjectes ja posseeixen a priori. Suposeu l'estudi de l'efecte de l'optimisme/pessimisme sobre la capacitat de recuperació d'una malaltia: no se li pot dir a un grup de subjectes que siga optimista i a un altre pessimista...

• Encara que no és sempre possible, en alguns estudis es pot optar per l'aplicació de qualsevol d'aquests tres mètodes d'investigació. Serà el responsable del disseny de la investigació qui, en funció dels recursos disponibles i de la mena de conclusions a què es vulga arribar, decidisca quin tipus de disseny d'investigació implementar. A manera d'il·lustració, suposeu que es vol comparar l'eficiència de l'aproximació farmacològica enfront de la psicològica en el tractament de l'ansietat (variable explicativa: tipus de tractament [farmacològic / psicològic]; variable de resposta: ansietat [puntuació en un test d'ansietat després d'aplicar el tractament]).

- **Exemple** de disseny d'investigació experimental → Es disposa d'un grup de subjectes amb problemes d'ansietat que desitgen rebre algun tipus de tractament. Es decideix crear dos subgrups, un que rebrà tractament farmacològic i un altre que en rebrà de psicològic. Els

subjectes són assignats aleatòriament, la meitat d'ells en un grup i l'altra meitat en l'altre. Es controla de manera acurada que al llarg de l'aplicació del tractament no hi haja altres variables que puguen influir de manera diferencial a un grup i a un altre en el seu nivell d'ansietat. Després de finalitzar el tractament, s'aplica un test als subjectes de tots dos grups per avaluar l'estat del seu trastorn d'ansietat; les puntuacions en aquest test són les que serviran de base per a comparar-los i extraure'n conclusions.

- Exemple de disseny d'investigació quasi-experimental → Es decideix crear dos subgrups, un que rebrà tractament farmacològic i un segon que en rebrà de psicològic. Al primer són assignats un conjunt de subjectes que han acudit a la secció de psiquiatria d'un hospital a rebre tractament, mentre que al segon, els procedents d'una clínica psicològica. Igual a l'anterior pel que fa al control de variables estranyes i a la mesura de la variable de resposta.

- Exemple de disseny d'investigació no experimental → En una sèrie de clíniques es recull informació de subjectes que hagen rebut tractament per a l'ansietat, ja siga psicològic o farmacològic, i que hagen passat després un determinat test d'ansietat per a avaluar el seu estat. Es comparen les puntuacions en el test entre aquells que han rebut el tractament psicològic o el farmacològic.

• L'anàlisi estadística de les dades derivades d'un estudi permetrà obtenir els resultats a partir dels quals extraure les conclusions oportunes. Ara bé, ha de ressaltar-se que la validesa d'aquestes conclusions dependrà fonamentalment del disseny d'investigació emprat, i no tant de l'anàlisi estadística de les dades realitzada. No és infreqüent que l'estadística reba unes atribucions que van més enllà dels beneficis que realment és capaç de proporcionar.

Per posar-ne un **exemple**, en l'estudi sobre el tractament de l'ansietat s'han obtingut les següents dades (tipus de disseny d'investigació utilitzat: qualsevol d'aquests):

Tract.Farm.	*Tract.Psic.*
6	4
8	9
11	4
...	...
$\bar{X}$=8	$\bar{X}$=6

Després de realitzar l'anàlisi de les dades recollides (obtenció de la mitjana aritmètica de tots dos grups), pot observar-se com la mitjana del grup del tractament farmacològic és major que la del tractament psicològic i, assumint que major puntuació en el test representa un pitjor resultat (major nivell d'ansietat), els resultats suggereixen concloure que el tractament

psicològic és més eficaç que el farmacològic. Ara bé, quina és la validesa d'aquesta conclusió?, es pot concloure que la causa de la diferència en ansietat entre tots dos grups es el tractament? La resposta a aquestes qüestions no ve determinada per l'anàlisi de les dades, sinó pel disseny d'investigació que s'haguera aplicat.

En el cas que s'haguera aplicat un disseny no experimental, tan sols podrem concloure que existeix relació entre la variable "Tractament" i la variable "Ansietat", en concret, que el tractament psicològic està associat amb nivells més baixos d'ansietat, mentre que el contrari succeeix amb el tractament farmacològic. Anar més enllà, definir una relació causal de la variable "Tractament" sobre la variable "Ansietat", és a dir, poder concloure que el tipus de tractament és la causa de les diferències observades en ansietat entre els dos grups només serà possible si s'ha aplicat un disseny d'investigació experimental. Finalment, en el cas d'aplicar un disseny quasi-experimental, la conclusió sobre la relació causal entre totes dues variables serà més robusta que si s'ha aplicat un disseny no experimental; no obstant això, aquesta conclusió estarà matisada per la incertesa □menor com millor es justifique l'equivalència dels grups en el moment d'iniciar-se el tractament.

Exercici 5: Indiqueu quina seria la metodologia més adequada per a contrastar empíricament cadascuna de les hipòtesis bivariades de l'exercici 3 en què hi ha una relació asimètrica, això és, una variable predictora i una variable de resposta. Una pista per a escollir el tipus de disseny: penseu si la variable explicativa es pot manipular o no.

7. L'informe d'investigació

• Un aspecte fonamental en el desenvolupament de qualsevol investigació és donar a conèixer com s'ha realitzat aquest estudi, així com els resultats i conclusions que se'n deriven. L'informe d'investigació representa el mitjà més reconegut de satisfer aquest objectiu, atès que fa possible que altres persones diferents a les que han participat en la seua realització puguen conèixer, valorar i replicar la investigació.

• El contingut d'un informe d'investigació pot adoptar diverses formes (lliurament per a una assignatura, treball de fi de grau, tesi de màster, comunicació en un congrés, article en una revista científica…) i cadascuna pot seguir uns patrons específics en la seua confecció. No obstant això, s'ha generalitzat entre la comunitat científica l'ús d'una sèrie de pautes relatives a què cal comptar i com cal explicar-ho, amb l'objectiu de facilitar la lectura dels informes d'investigació independentment de qui els haja escrit.

• La progressiva generalització de les pautes a seguir en l'elaboració d'un informe d'investigació s'ha traduït en la divulgació d'alguns manuals de publicació que especifiquen i il·lustren l'aplicació d'aquestes pautes. Així, quan es desitge publicar un informe d'investigació en una revista científica (i. e., un article), el primer que haurem de fer és mirar les recomanacions de l'editor de la revista sobre el manual de publicació concret en què hem de basar la redacció del treball. Així, en l'àmbit de la psicologia hi ha un nombre important de revistes en què es demana que els treballs que s'envien per a ser avaluats seguisquen en la seua redacció el *Manual de Publicació de l'APA* (American Psychological Association, 2019).

• Un aspecte en el qual existeix gran coincidència entre els diferents manuals de publicació és el que fa referència a l'estructura que ha de tenir un informe d'investigació i que es concreta en els següents apartats: (1) Títol; (2) Autor/s i filiació; (3) Resum; (4) Introducció; (5) Mètode; (6) Resultats; (7) Discussió; (8) Referències; (9) Apèndix, quan aquest siga oportú.

• A continuació es donen alguns detalls de quin és el contingut i la forma de cadascun d'aquests apartats d'acord amb el *Manual de Publicació de l'APA*:

- **Resum**: ha de sintetitzar el contingut del treball, per la qual cosa se sol suggerir que s'hi incloga el més essencial de les quatre seccions subsegüents. Si l'informe no està escrit en anglès, se sol demanar que el resum aparega també traduït a l'anglès. La longitud oscil·larà entre 150 i 250 paraules.
- **Introducció**: ha d'incloure una revisió de la temàtica objecte d'estudi que permeta entendre quina és la motivació i l'interès de l'estudi. La redacció de la introducció sol implicar la revisió de la literatura existent sobre el tema objecte d'estudi, raó per la qual és freqüent que en les introduccions apareguen bastantes cites a altres treballs. En la part final de la introducció s'ha d'especificar quin és l'objectiu de l'estudi i, quan siga el cas, enunciar quina hipòtesi es desitja posar a prova.
- **Mètode**: ha d'aportar una descripció de com l'estudi es va dur a terme en la pràctica, de manera que qualsevol lector siga capaç d'entendre què és el que s'hi va fer i fins i tot dur-ne a terme una rèplica. Entre altres aspectes, s'han de descriure les característiques dels participants en l'estudi i com s'hi va accedir, els mitjans utilitzats en l'obtenció de les dades (qüestionaris, registres d'observació, aparells…) i el procediment seguit en la recollida de les dades (el disseny d'investigació).
- **Resultats**: s'han d'oferir els resultats de l'anàlisi estadística de les dades recollides, en concret, totes les que donen suport a les conclusions que s'incloguen en la següent secció. És important utilitzar taules i gràfics que faciliten la lectura dels resultats.

- **Discussió**: ací es comentaran els resultats obtinguts en relació amb els objectius de l'estudi, es discutirà sobre les conseqüències teòriques i aplicades d'aquests resultats, es plantejaran les possibles limitacions i febleses de l'estudi, així com propostes de millora i vies de treball futures que es consideren d'interés.
- **Referències**: en aquesta secció s'inclouen les referències de tota la bibliografia que haja sigut citada en el contingut de l'informe d'investigació. Les referències bibliogràfiques han d'aparèixer llistades en ordre alfabètic i en cadascuna de les referències ha d'aparèixer tota la informació necessària per a poder localitzar el document al qual es faça referència. A continuació s'entra detalladament en alguns dels aspectes relatius a la bibliografia d'un informe d'investigació.

• Si alguna part del nostre treball es fonamenta en les idees o en el treball d'altres autors, hem de citar-los en l'informe en el mateix moment en què es fa referència en el text a les seues idees o resultats d'investigació. La manera de fer-ho es basa en la utilització de cites bibliogràfiques, un mètode abreujat que pretén que la lectura no siga difícil i que consisteix a utilitzar tan sols el primer cognom de l'autor/s i l'any en què va ser publicat el seu treball. És en la secció de Referències, al final de l'informe d'investigació, on s'inclourà un llistat de les referències completes de tots els treballs citats en el text.

• A continuació es mostren dos exemples de cites bibliogràfiques realitzades en el text d'un informe d'investigació d'acord amb les dues modalitats possibles de cita acceptades pel *Manual de Publicació de l'APA*:

Exemple de 1a modalitat de cita bibliogràfica:

> Tal com van assenyalar Amador i Sopeña (2009) i Romero (2006), existeix relació entre la lateralitat (esquerrà/destre) i l'execució en tasques visomotores, per la qual cosa en aquest treball hem optat per…

Exemple de 2a modalitat de cita bibliogràfica:

> Atesa la relació posada de manifest entre la lateralitat (esquerrà/destre) i l'execució en tasques visomotores (Amador i Sopeña, 2009; Romero, 2006), en el present treball hem optat per…

Tot seguit es mostren exemples de referències bibliogràfiques tal com apareixerien en la secció de Referències d'un informe d'investigació, d'acord amb el *Manual de Publicació de l'APA*. Es mostren exemples dels quatre tipus de referències bibliogràfiques més habituals:

Exemple de referència bibliogràfica d'article :

Scandura, J. M., i Wells, J. N. (1967). Advance organizers in learning abstract mathematics. *American Educational Research, 4*, 295-301.

Exemple de referència bibliogràfica de llibre :

Gómez, J. (1987). *Meta-análisis.* PPU.

Exemple de referència bibliogràfica de compilació o llibre :

Navas, M. J. (Ed.) (2001). *Métodos, diseños y técnicas de investigación psicológica.* UNED.

Exemple de referència bibliogràfica de capítol de llibre en una compilació o llibre:

Martínez, M. R. (1984). El anàlisis de los datos de diseños de caso único. En J. Major, i F. Labrador (Eds.), *Manual de modificación de conducta* (pp.155-202). Alhambra.

Exercici 6: Identifiqueu de quin tipus (llibre, article…) és cadascuna de les referències bibliogràfiques següents.

a) Rivière, A. (1993). Sobre objetos con mente: reflexiones para un debate. *Anuario de Psicologia*, *56*, 49-75.
b) American Psychological Association (2019). *Publication Manual of the American Psychological Association* (7ª ed.). APA.
c) Arnau, J. (1990). Metodologia experimental. En J. Arnau, M. T. Anguera, y J. Gómez (Eds.), *Metodología de la investigación en ciencias del comportamiento* (pàg. 9-122). Universidad de Murcia.
d) Major, J., i Labrador, F. (Eds.) (1984). *Manual de modificación de conducta*. Alhambra.

Exercici 7: Escriviu les referències bibliogràfiques següents d'acord amb el *Manual de Publicació de l'APA*:

a) Llibre: Autors: Francisco Xavier i Diego Macià / Títol: Modificación de conducta con ninos y adolescentes / Any de publicació: 1994 / Editorial: Pirámide.
b) Article: Autors: Daniel R. Weinberger / Títol: Evidence of dysfunction of a pre-frontal limbic network in schizophrenia / Any de publicació: 1992 / Revista: American Journal of Psychiatry / Volum: 149 / Pàgines: 890-897.
c) Capítol de compilació: Autor del capítol: Alfredo Fierro / Títol del capítol: Desarrollo social y de la personalidad en la adolescencia / Autors de la compilació: Mario Carretero, Jesús Palacios i Álvaro Marchesi / Títol de la compilació: Psicologia 3. Adolescencia, madurez y senectud / Any de publicació: 1991 / Editorial: Alianza / Pàgines: 95-142.

d) Compilació: Autors: Mario Carretero, Jesús Palacios i Álvaro Marchesi / Títol: Psicologia 3. Adolescencia, madurez y senectud / Any de publicació: 1991 / Editorial: Alianza.

Exercici 8: Contesteu les qüestions que es plantegen tot seguit referides a aquest article:
Suriá, R. & Villegas, E. (2020). Empoderamiento y rendimiento académico en estudiantes de educación secundaria obligatoria con y sin discapacidad. *Anuario de Psicología, 50*(1), 29-37.

Nota: L'article s'adjunta al final d'aquest tema, si bé, també es pot consultar i descarregar lliurement des de la pàgina de la revista *Anuario de Psicologia.*
(https://revistes.ub.edu/index.php/anuario-psicologia/index)

a) En quina universitat treballa Raquel Suriá?
b) Quins són els objectius de la investigació descrita en l'article?
c) Quines són les variables explicatives i les variables de resposta? De quin tipus són aquestes variables en funció de l'escala de mesura (nominal, ordinal o quantitativa)?
d) Quin tipus de metodologia s'ha utilitzat en aquest estudi (experimental, quasi-experimental, no experimental)? Per què?
e) Quins subjectes componen la mostra utilitzada? Quina seria la població d'estudi?
f) Identifiqueu si cadascuna de les taules presentades en l'article són d'estadística descriptiva o inferencial.
g) Comproveu que els apartats de l'article es corresponen amb els d'un informe d'investigació d'acord amb el *Manual de Publicació de l'APA*.
h) Cerqueu les següents referències en l'apartat de Referències i indiqueu de quin tipus són (article, llibre, etc.): 1. Cohen, 2. Moore, 3. Surià 4. Zimmerman
i) Comproveu si el contingut de cada apartat s'ajusta a les recomanacions de l'APA (escolliu la resposta correcta):
    1. Resum: SÍ/NO (en cas que NO, per què?)
    2. Introducció: SÍ/NO (en cas que NO, per què?)
    3. Mètode: SÍ/NO (en cas que NO, per què?)
    4. Resultats: SÍ/NO (en cas que NO, per què?)
    5. Discussió: SÍ/NO (en cas que NO, per què?)
    6. Referències: SÍ/NO (en cas que NO, per què?)
    7. Escriviu la referència bibliogràfica d'aquest article d'acord amb les normes APA.

Referències

American Psychological Association (2019). *Publication Manual of the American Psychological Association* (7ª ed.). APA.

Aaron, A., & Aaron, E. N. (2001). *Estadística para Psicología.* Prentice Hall.

Botella, J., León, O. G., San Martín, R., & Barriopedro, M. I. (2001). *Análisis de datos en psicología I: teoría y ejercicios.* Pirámide.

Cava, M. J., Murgui, S., & Musitu, G. (2008). Diferencias en factores de protección del consumo de sustancias en la adolescencia temprana y media. *Psicothema, 20,* 389-395.

Navas, M. J. (2001). *Métodos, diseños y técnicas de investigación psicológica.* Madrid: UNED.

Pardo, A., & San Martín, R. (2001). *Análisis de datos en psicología II.* Pirámide.

Portell, M., Vives, J., & Boixadós, M. (2003). *Mètodes d'investigació: recursos didàctics.* Servei de Publicacions de la UAB.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677-680.

Tema 2 – Organització i representació gràfica de les dades

1. **La distribució de freqüències**
2. **La representació gràfica d'una distribució de freqüències**
3. **Propietats de les distribucions de freqüències**

• La recollida de dades associada a la realització d'un estudi sol representar l'obtenció d'un conjunt quantiós de dades i, com a conseqüència, la interpretació d'aquestes a simple vista sol resultar poc intel·ligible en la majoria dels casos. L'estadística descriptiva ens ofereix eines per a organitzar i resumir les dades que hàgem recollit, de manera que puga ser extreta i interpretada la informació que s'hi continga i que siga del nostre interés. En aquest i en els set capítols que segueixen es presenten alguns dels mètodes més comuns de descripció estadística.

1. La distribució de freqüències

• La distribució de freqüències constitueix una de les formes més intuïtives de resumir les dades d'una variable: es basa en la creació d'una taula amb el recompte del nombre de casos (unitats d'observació, subjectes…) que hi ha en cadascuna de les possibles modalitats de la variable. És una tècnica estadística bàsica, però, tanmateix, molt informativa i rellevant en la pràctica de l'anàlisi de dades.

• L'elaboració d'una distribució de freqüències d'una variable (X) es basa en l'obtenció de:

(1) Les modalitats de la variable (X_i).

(2) El nombre de vegades que apareix cadascuna de les modalitats de la variable en el conjunt de les dades. Aquests recomptes es denominen freqüències absolutes (n_i) de les modalitats.

(2.1) Derivades de les freqüències absolutes es poden obtenir les freqüències relatives o proporcions (p_i):

$$p_i = n_i / n$$

(2.2) Les freqüències relatives també poden expressar-se en forma de percentatges ($\%_i$) si multipliquem el seu valor per 100:

$$\%_i = p_i \cdot 100$$

Exemple de distribució de freqüències per a la variable categòrica "Estat civil" (X), de la qual s'han recollit dades en una mostra de 50 persones de la ciutat de Castelló (n = 50):

X: {0; 0; 1; 2; 2; 0; 1; 3; 2; 0; 1; 0; 1; 2; 0; 2; 1; 1; 0; 1; 0;...}

Codificació: 0: solter/a; 1: casat/ada; 2: separat/ada; 3: vidu/a.

X_i	Freq. absoluta (n_i)	Freq. relativa (p_i)	Percentatge ($\%_i$)
0	15	0,3	30
1	20	0,4	40
2	11	0,22	22
3	4	0,08	8
	50	1,00	100

• En el cas de les variables quantitatives i les quasi-quantitatives, a més de l'anterior, es pot obtenir també la següent informació per a cadascuna de les modalitats:

- les freqüències absolutes acumulades (n_a),
- les freqüències relatives acumulades (p_a),
- i els percentatges acumulats ($\%_a$).

Exemple de distribució de freqüències per a la variable quantitativa "Nombre de fills/es" (X), amb dades per a una mostra de 20 famílies del barri de Velluters de la ciutat de València:

X: {2; 1; 0; 3; 2; 2; 3; 1; 1; 0; 1; 2; 1; 2; 0; 2; 4; 2; 3; 1}

X_i	Freq. absoluta (n_i)	Freq. relativa (p_i)	Percentatge ($\%_i$)	Freq. absoluta acumulada ($n_{a\,).}$	Freq. relativa acumulada ($p_{a\,).}$	Percentatge acumulat ($\%_{a\,).}$
0	3	0,15	15	3	0,15	15
1	6	0,30	30	9	0,45	45
2	7	0,35	35	16	0,80	80
3	3	0,15	15	19	0,95	95
4	1	0,05	5	20	1,00	100
	20	1	100			

• Algunes anotacions sobre les distribucions de freqüències:

(1) És costum situar els valors corresponents a la columna de les modalitats (1a columna de la taula) en sentit creixent de dalt cap avall.

(2) Per als valors de la variable que no hi haja cap cas és costum no dedicar cap fila en la taula de la distribució de freqüències a fi que aquesta ocupe menys espai.

(3) Les freqüències relatives o proporcions es caracteritzen per prendre valors entre 0 i 1, de manera que la suma de totes elles igual a la unitat. El mateix per als percentatges respecte a 100.

Exercici 1: Les dades següents són dels estudiants/es d'una classe en la qual un observador, durant el temps que dura una sessió de classe de 2 hores, ha anotat el nombre de vegades que ha participat cadascun dels estudiants dirigint-se a tot el grup en veu alta.

2 2 3 0 3 1 8 0 3 9 1 1 0 4 0 2 9 5 0 1

Obtingueu la distribució de freqüències completa (utilitzeu dos decimals). A partir d'aquesta distribució, contesteu les següents preguntes:

a) Quina proporció d'estudiants participà 2 ocasions o menys en la sessió de classe? Quants estudiants són?

b) Quin percentatge d'estudiants participà 5 vegades? Quants són?

c) Quina proporció d'estudiants participà 4 vegades o menys? Quants són?

d) Quin percentatge d'estudiants participà més de 4 vegades? Quants són?

e) Quina proporció d'estudiants participà almenys una vegada? Quants són?

f) Quin percentatge d'estudiants participà entre 2 i 5 vegades, ambdues inclosos? Quants són?

g) Quin percentatge d'estudiants participà 8 o 9 vegades? Quants són?

h) Quina proporció d'estudiants participà 4 vegades o més? Quants són?

(4) En el cas de les variables quantitatives contínues, atès que es pot obtenir un *nombre* gran de dades diferents si la mesura de la variable es realitza amb certa precisió, és pràctica habitual que en la columna de les modalitats (X_i) els valors s'agrupen en intervals d'igual amplada.

Exemple de la distribució de freqüències elaborada a partir de les dades de la variable "Pes (*kg*)" (X) dels 420 jugadors inscrits en la lliga professional de handbol masculí en la temporada 2008/09:

X: {82,5; 91,1; 90,6; 83,8; 92,1; 88,3; 93,6; 101,4; 91,7; 80,2; ...}

Pes (kg)	n_i
...	...
...	...
77	1
79	3
80	2
81	6
82	5
83	9
...	...
...	...
...	...

Així, per exemple, el valor 80 de la columna de les modalitats representa, en realitat, al conjunt de valors comprés entre 79,5 i 80,5 kg; el valor 81 a l'interval de 80,5 a 81,5 kg, i així successivament. Recordeu que en l'enumeració d'intervals que se solapen en un punt és habitual considerar que el primer valor de l'interval en forme part, mentre que el segon ja es considere del següent interval.

(5) Seguint amb el cas anterior, si el nombre de modalitats que pren la variable és molt ampli, una alternativa que permet generar una distribució de freqüències més compacta consisteix a organitzar la distribució de freqüències definint intervals de valors.

Exemple de la distribució de freqüències elaborada a partir de les dades de la variable "Alçada (*cm*)" en una mostra de 1436 subjectes adults de la població espanyola:

Alçada (cm)	n_i
140-150	15
150-160	131
160-170	345
170-180	623
180-190	267
190-200	42
200-210	13

En aquest cas, l'interval 140-150, per posar un exemple, representa a tots els valors compresos entre 140 i 150 cm.

Exercici 2: A partir de la distribució de freqüències de la variable "Alçada (cm)" de l'exemple previ, obtingueu les corresponents columnes de freqüències relatives, percentatges, freqüències absolutes acumulades, freqüències relatives acumulades i percentatges acumulats.

Exercici 3: En una enquesta sobre condicions psicosocials en el lloc de treball es va preguntar a una mostra de 3420 treballadors/es, entre altres qüestions, "en quina mesura el seu treball és desgastador emocionalment?" (X). Es van obtenir els següents resultats:

X_i	n_i	p_i	$\%_a$
Mai			10
Alguna vegada		0,10	
A vegades	513		
Moltes vegades			55
Sempre	1539	0,45	100
	3420	1	

Després d'emplenar els buits de la distribució de freqüències anterior, contesteu les següents qüestions:

a) Quina proporció de subjectes considera que el seu treball és desgastador emocionalment *moltes vegades*?

b) Quin percentatge de subjectes va contestar *mai*? I quin percentatge va contestar *alguna vegada* o *mai*?

c) Quants subjectes consideren que el seu treball és desgastador emocionalment *moltes vegades*? I quants consideren que *mai* ho és?

(6) Una distribució de freqüències condicionada mostra la distribució de freqüències d'una variable per als casos que en una segona variable tenen un determinat valor. És un concepte útil per a descriure com es comporta una variable en funció dels diferents valors que pren una segona variable.

Exemple. Tenim la següent distribució de freqüències de la variable "Qualificació examen" obtinguda en una mostra de 200 persones (n = 200):

Qualificació examen	n_i
Aprovat	130
Notable	41
Excel·lent	27
Matrícula honor	2
	200

A continuació es mostren les distribucions de freqüències de la variable "Qualificació examen" condicionada als valors de la variable "Sexe" [Dona; Home]:

Qualificació examen	(Sexe: Dona) n_i	(Sexe: Home) n_i
Aprovat	80	50
Notable	20	21
Excel·lent	14	13
Matrícula d'honor.	1	1
	115	85

Si la grandària dels subgrups definits per la variable condicionant no és igual (o bastant similar) és convenient presentar les distribucions de freqüències expressades en proporcions o percentatges a fi de poder comparar-les de forma correcta. Per exemple, les distribucions de freqüències de "Qualificació examen" condicionades a "Sexe", en percentatges, són:

Qualificació examen	(Sexe: Dona) $\%_i$	(Sexe: Home) $\%_i$
Aprovat	69,6	58,8
Notable	17,4	24,7
Excel·lent	12,2	15,3
Matrícula d'honor	0,9	1,2
	100	100

Exercici 4: En el context d'un estudi sobre la percepció de la ciència i la tecnologia a Espanya es va preguntar a una mostra de 7054 subjectes (1252 joves i 5802 adults) com valoraven la formació científica i tècnica rebuda. Els resultats van ser els següents:

Nivell de formació	Jove	Adult
Molt baix	7,99 %	22,50 %
Baix	28,51 %	33,30 %
Normal	45,53 %	32,80 %
Alt	14,30 %	9,00 %
Molt alt	3,67 %	2,40 %

a) Quin grup valora més positivament el nivell de formació rebut, el dels joves o el dels adults?
b) Expresseu les distribucions anteriors en freqüències absolutes.
c) Genereu la distribució de freqüències per al conjunt de subjectes de la mostra (n = 7054) en freqüències absolutes i en freqüències relatives.

• El programa SPSS: En obtenir la distribució de freqüències d'una variable amb aquest programa es mostra n_i, $\%_i$, $\%_i$ *vàlid* i $\%_a$, però no la informació referida a les freqüències relatives (p_i i p_a), tal com es pot observar en els dos exemples que es mostren a continuació.

Exemples: el primer exemple presenta la distribució de la variable "Satisfacció amb les instal·lacions d'un centre esportiu" (*sat_ins*) per a un grup de 106 usuaris. Les dades s'han recollit en una escala de 0 a 20 [0: totalment insatisfet; ... ; 20: totalment satisfet]; i el segon exemple presenta la variable "Religió" en una mostra de 407 estudiants de la UV (escala de resposta: Catòlica; No creient; Altra religió).

Satisfacció instalacions centre

		Freqüència	Percentatge	Percentatge vàlid	Percentatge acumulat
Vàlid	4	1	,9	,9	,9
	7	4	3,8	3,8	4,7
	8	11	10,4	10,4	15,1
	9	7	6,6	6,6	21,7
	10	26	24,5	24,5	46,2
	11	7	6,6	6,6	52,8
	12	24	22,6	22,6	75,5
	13	4	3,8	3,8	79,2
	14	11	10,4	10,4	89,6
	15	4	3,8	3,8	93,4
	16	4	3,8	3,8	97,2
	17	1	,9	,9	98,1
	18	2	1,9	1,9	100,0
	Total	106	100,0	100,0	

Religió

		Freqüència	Percentatge	Percentatge vàlid	Percentatge acumulat
Vàlid	Catòlica	128	31,4	34,7	34,7
	No creient	224	55,0	60,7	95,4
	Altra	17	4,2	4,6	100,0
	Total	369	90,7	100,0	
Perduts	Sistema	38	9,3		
Total		407	100,0		

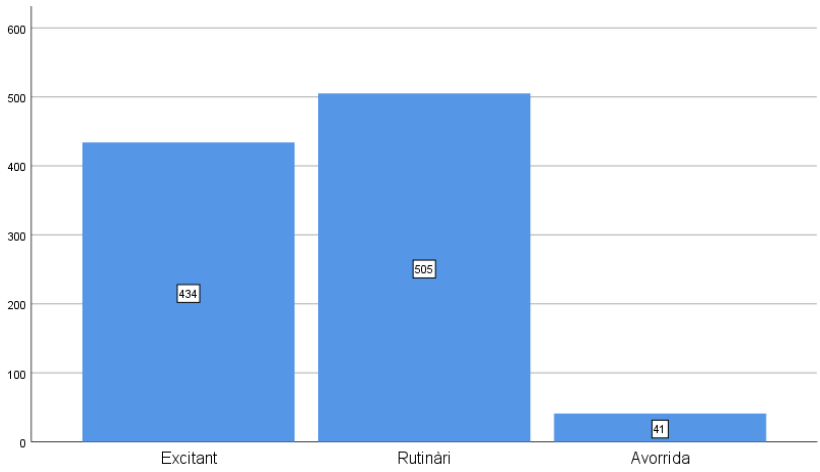
La diferència entre les columnes *percentatge* i *percentatge vàlid* és que el primer s'obté dividint cada freqüència absoluta entre el nombre total de casos (407, en l'exemple de Religió), mentre que el segon s'obté dividint entre el nombre de casos vàlids, es a dir, dels quals de fet s'ha recollit alguna dada en la variable, sense considerar els subjectes que no responen la pregunta (valors *perduts* o *missings*) (el total de casos vàlids és 369 en l'exemple de Religió).

2. La representació gràfica d'una distribució de freqüències

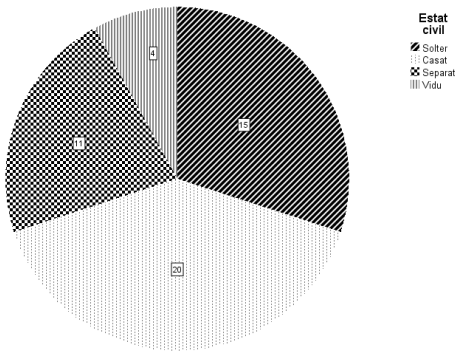
2.1. Per a variables categòriques

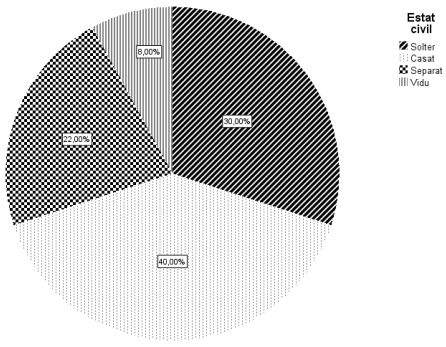
• El gràfic de barres: Les modalitats de la variable se situen sobre l'eix *X* (abscisses). Les barres tenen una altura igual a la freqüència absoluta de cadascuna de les modalitats de la variable. L'eix d'ordenades pot aparèixer expressat en freqüències absolutes, en freqüències relatives o en percentatges. Els gràfics de barres poden representar-se també de forma horitzontal.

> **Exemple** de gràfic de barres vertical per a la variable procedent de la següent pregunta d'un test: "Com és la seua vida?" (escala de resposta: Excitant; Rutinària; Avorrida):

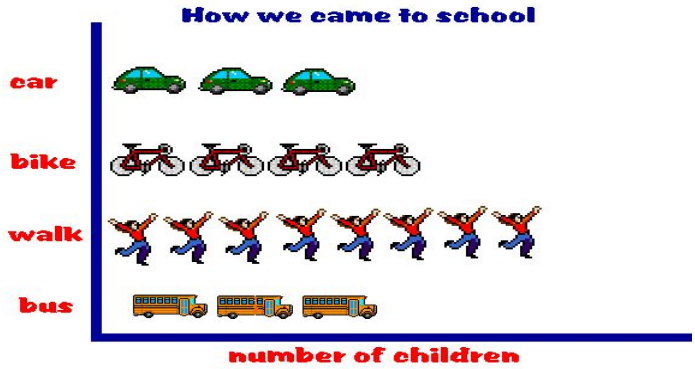


• El gràfic de sectors (pastissos o formatges): l'àrea de cada sector és proporcional a la freqüència o el percentatge de la modalitat a la qual representa. **Exemple** per a la variable "Estat civil":





• El pictograma: és una variació més vistosa dels gràfics de barres encara que també més procliu a generar confusions en la seua interpretació. **Exemple** per a la variable "Mitjà de transport per a anar al col·legi":

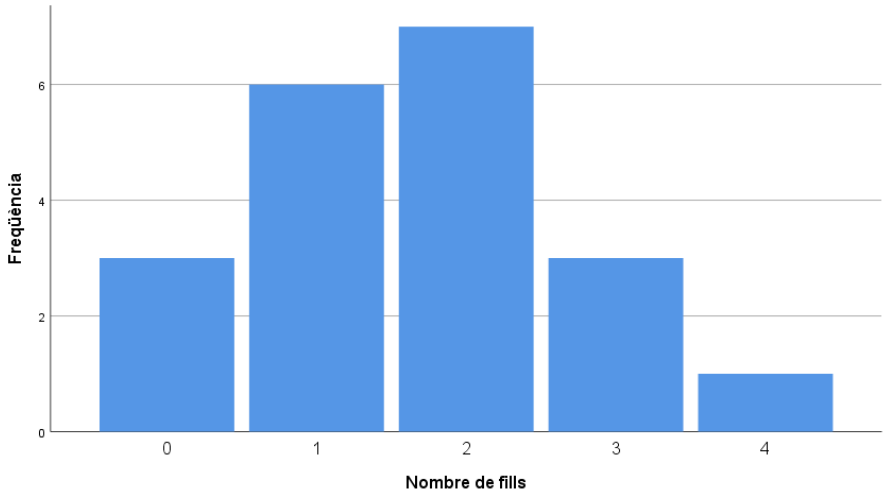


Exercici 5: A partir dels gràfics de les variables "Com és la seua vida?" i "Estat civil", obtingueu les corresponents distribucions de freqüències.

2.2. Per a variables quasi-quantitatives (ordinals) i quantitatives discretes

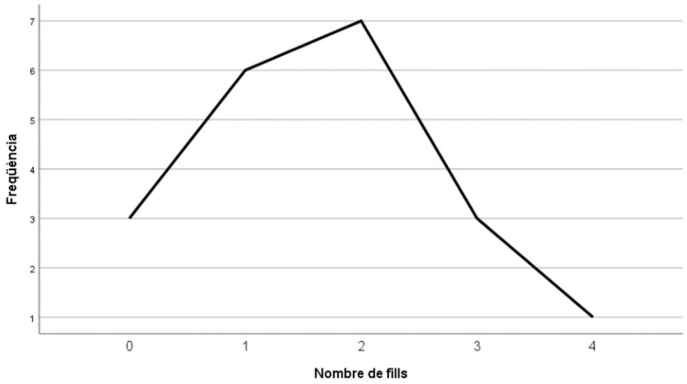
• És possible utilitzar els gràfics vistos en l'apartat anterior, si bé s'ha de tenir en compte un parell d'aspectes en la utilització del gràfic de barres: (1) no s'ha d'oblidar el buit entre les barres, perquè aquest serveix per a ressaltar que hi ha valors que no són possibles per a la variable representada; (2) a diferència de les variables categòriques, per a les variables ordinals i les quantitatives discretes sí que té sentit representar no sols les freqüències absolutes, les relatives i els percentatges, sinó també les respectives acumulades.

Exercici 6: A partir del gràfic de barres de la variable "Nombre de fills/es", dibuixeu el corresponent gràfic de barres de freqüències acumulades.



• El polígon de freqüències: polígon que resulta d'unir amb una línia els valors de les freqüències o percentatges (ja siguen acumulades o no) corresponents a les modalitats de la variable.

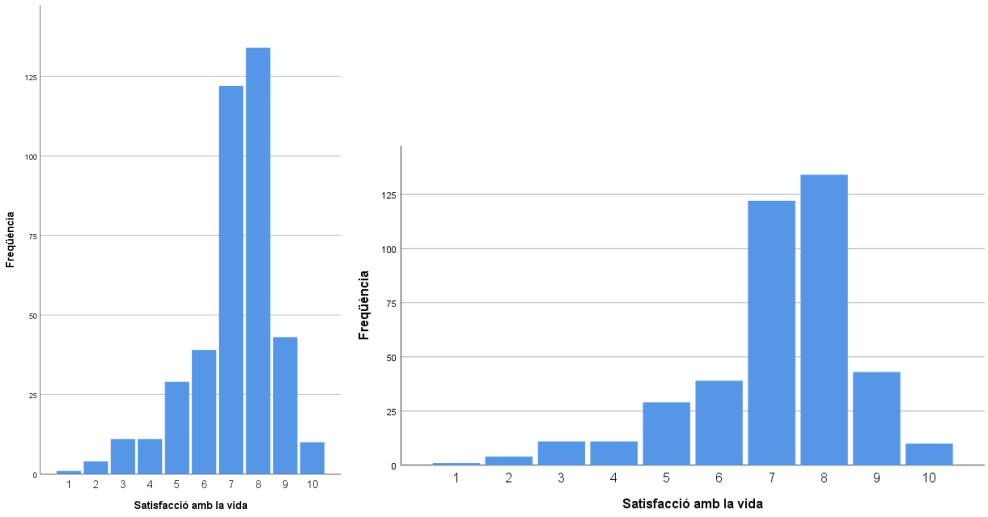
Exemple per a la variable "Nombre de fills/es":



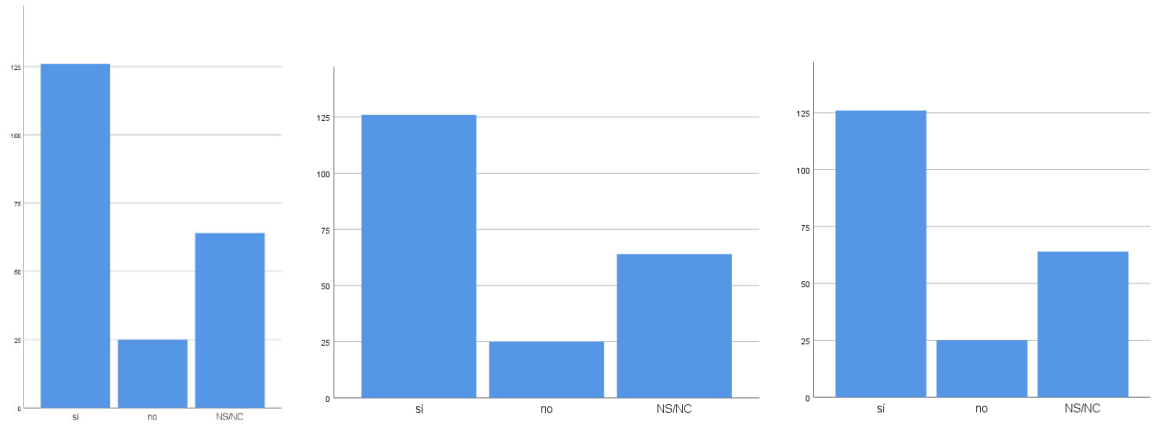
Exercici 7: A partir del polígon de freqüències de la variable "Nombre de fills/es", dibuixeu el corresponent polígon de freqüències acumulades.

Exercici 8: Feu, per a la variable "Nombre de vegades que es participa en classe" (vegeu l'exercici 1), els gràfics de barres corresponents a: les freqüències absolutes, les freqüències relatives, les freqüències absolutes acumulades i les freqüències relatives acumulades. Dibuixeu un polígon de freqüències a partir de qualsevol dels anteriors.

• Un aspecte que pot influir en la percepció d'una representació gràfica és la relació entre la longitud de l'eix X i de l'eix Y. Per exemple, els dos següents gràfics, encara que representen unes mateixes dades (variable "Satisfacció amb la vida" en una mostra d'estudiants), poden conduir a una percepció diferent de la informació proporcionada:

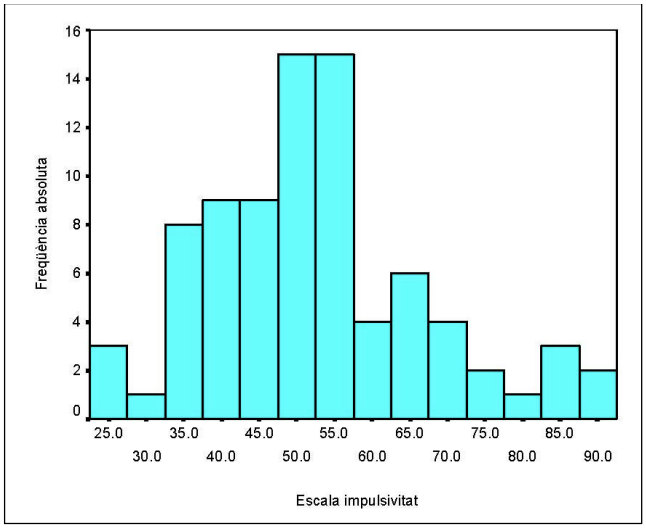


A fi d'evitar aquesta possible font de confusió, alguns autors recomanen que la relació entre l'amplària i l'altura del gràfic siga d'1,25 a 1. A tall d'exemple, quina de les següents representacions gràfiques, corresponents a un mateix conjunt de dades, s'ajusta a aquesta recomanació?

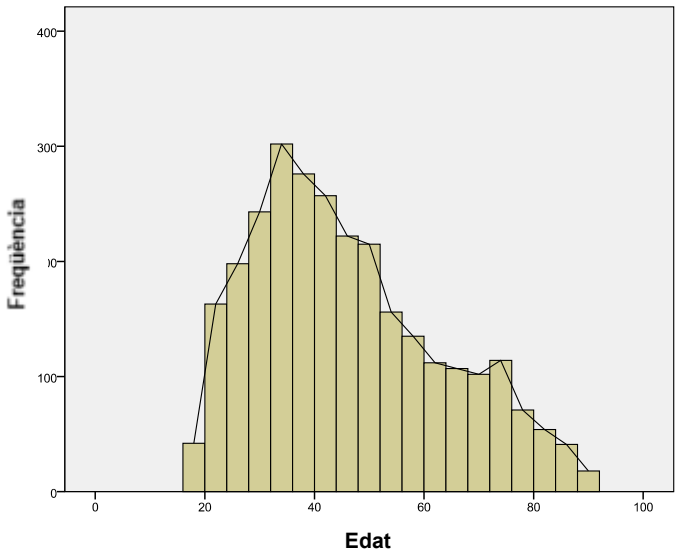


2.3. Per a variables quantitatives contínues

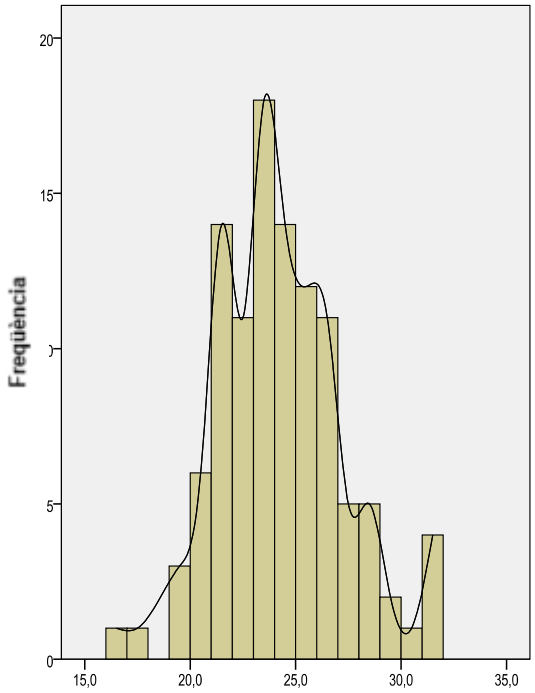
• Histograma: similar al gràfic de barres, si bé les barres són consecutives atesa la continuïtat de la variable. Cada barra representa ara no un valor sinó un interval de valors. Per a definir els intervals de valors (normalment, tots de la mateixa amplada) s'ha de tenir en compte que cap de les dades recollides per a la variable es quede fora dels intervals. Els intervals han de ser exhaustius i excloents. **Exemples** per a les puntuacions obtingudes per un grup de subjectes en una escala orientada a mesurar la impulsivitat.



• De la mateixa manera que amb les variables ordinals i les quantitatives discretes, també és possible dibuixar polígons de freqüències per a les variables quantitatives contínues unint amb una línia els valors de les freqüències o els percentatges (ja siguen acumulades o no) corresponents als intervals de valors creats. Vegeu l'exemple a continuació per a la variable "Edat de l'enquestat" superposat a l'histograma d'aquesta variable.



• Una variant del polígon de freqüències és la coneguda com a corba suavitzada. Per a la seua obtenció s'han proposat diversos procediments de suavització que pretenen eliminar les irregularitats en el polígon de freqüències que se suposa que són el resultat d'errors de mostratge. Vegeu l'exemple a continuació per a una variable obtinguda a partir de la mesura existent entre dos punts concrets del cervell.

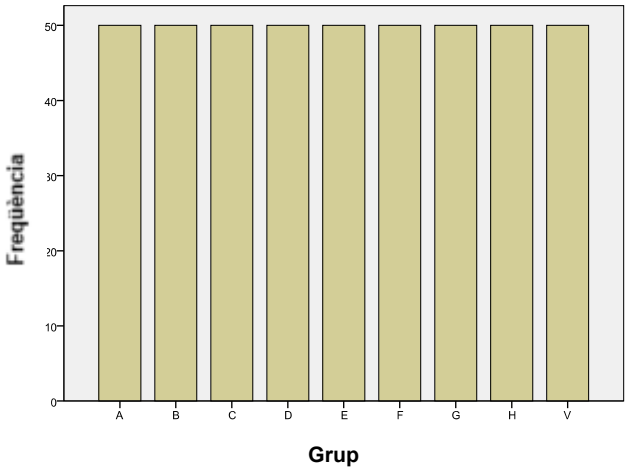


3. Propietats de les distribucions de freqüències

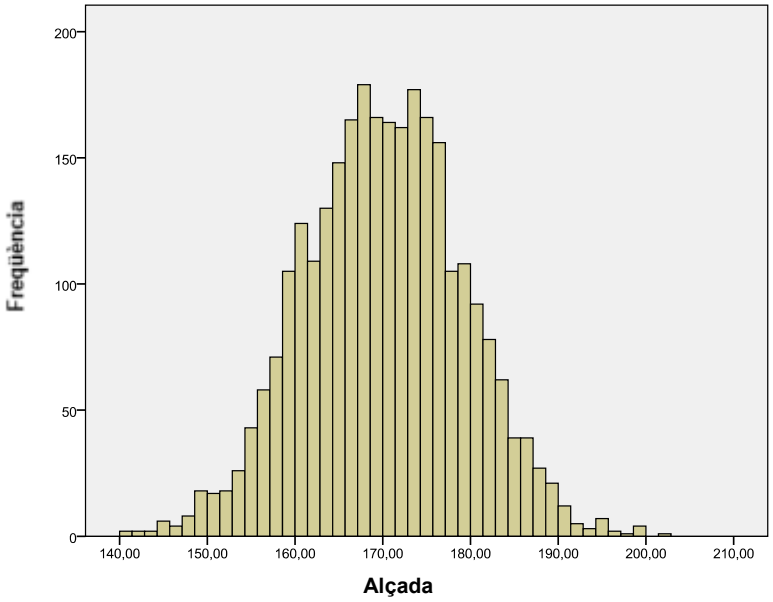
• Tot i que la representació gràfica d'una distribució de freqüències pot adoptar múltiples formes, hi ha alguns patrons de distribució que, per la seua singularitat i/o importància, han sigut denominats d'una manera concreta.

A tall d'exemple, les dues següents presentades en forma gràfica per a dues variables, "Grup al qual es pertany en l'assignatura d'Estadística" i "Alçada":

- La distribució rectangular o uniforme:



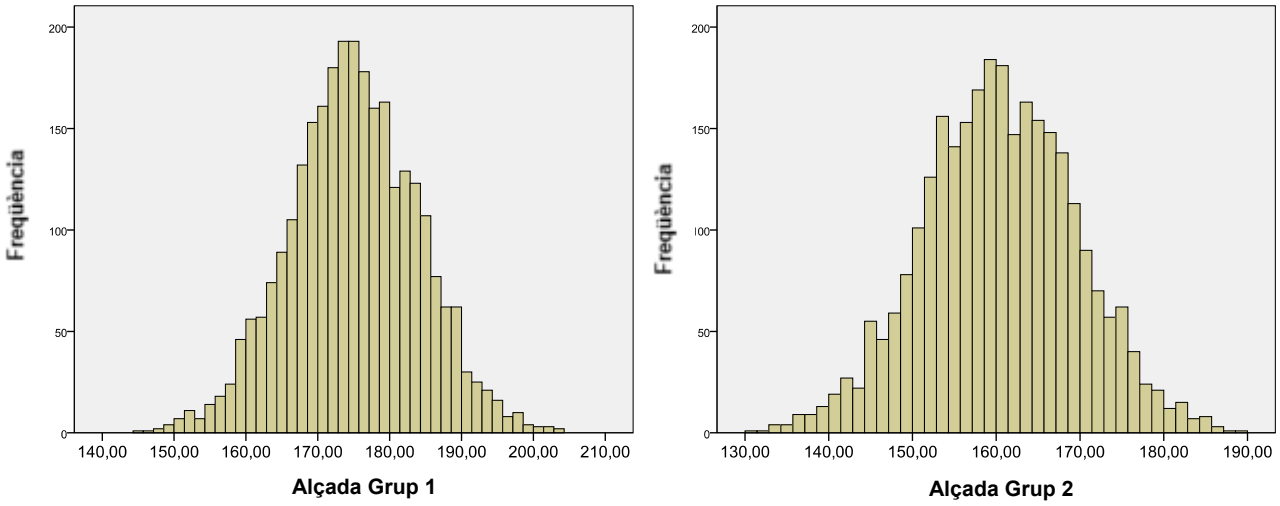
- La distribució normal:



• Sobre aquests dos patrons i uns altres que caracteritzen en el seu conjunt a la distribució de freqüències d'algunes variables s'aprofundirà en un tema posterior. Ara bé, per a descriure una distribució de freqüències podem atendre, més que a la forma en el seu conjunt, a diferents aspectes particulars d'aquesta. Així, els dos temes que segueixen se centren en alguns d'aquests aspectes que permeten sintetitzar la informació continguda en una distribució de freqüències. Es tracta de aspectes com els dos següents:

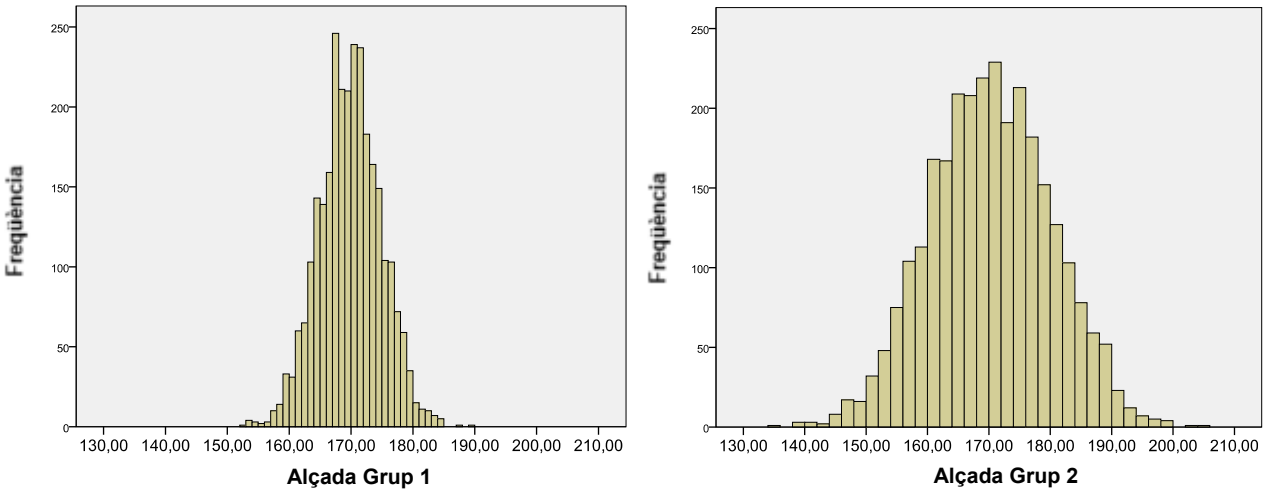
- La posició de la distribució

Exemple de la diferent posició de les dues distribucions de freqüències d'una mateixa variable, "Alçada (cm)", mesurada en dos grups de subjectes diferents:



- La dispersió o variabilitat de la distribució

Exemple de la diferent dispersió de la distribució de freqüències d'una mateixa variable, "Alçada (cm)", mesurada en dos grups de subjectes diferents –que, no obstant això, comparteixen una posició molt similar (la mitjana aritmètica d'ambdues és de 170 cm):



Tema 3.1 – Caracterització de grups: Estadístics de posició grupal

• Tant en aquest tema com en els dos següents es revisaran una sèrie d'índexs estadístics que ens permetran resumir la informació continguda en la distribució de freqüències d'una variable. Aquests índexs proporcionen valors numèrics que, de manera sintètica, resumeixen diferents característiques d'una distribució de freqüències, com ara la posició, la variabilitat i l'asimetria.

• Aquest tema, en concret, se centra en una sèrie d'índexs estadístics que permeten descriure numèricament quina és la localització o posició de la distribució de freqüències d'una variable.

• Els estadístics de posició grupal sintetitzen la informació de totes les dades d'una variable, això és, per a tot el grup de casos a partir dels quals aquestes dades han sigut recollides. Enfront d'aquests, en el tema 4 ("Estadístics de posició individual") es tractaran alguns procediments orientats a obtenir informació sobre la posició relativa d'un subjecte dins del seu grup.

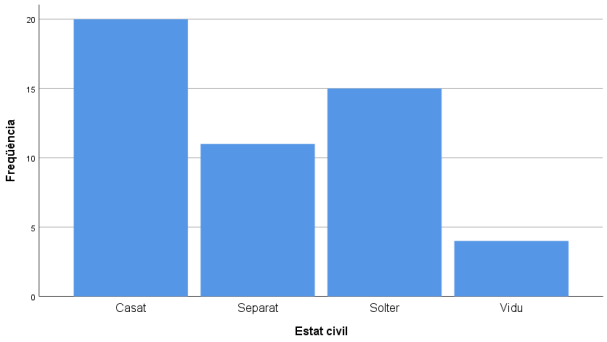
1. Estadístics de tendència central

• Els índexs que es revisaran en aquesta primera secció són indicatius de la tendència central de les dades d'una variable i, per tant, proporcionen un valor que expressa la posició entorn de la qual se centra la distribució de freqüències de la variable, això és, un valor que exercirà de representant de totes les dades recollides per la variable. Diferenciarem aquests estadístics en funció del tipus de variable (categòrica, ordinal, quantitativa) que es vol descriure.

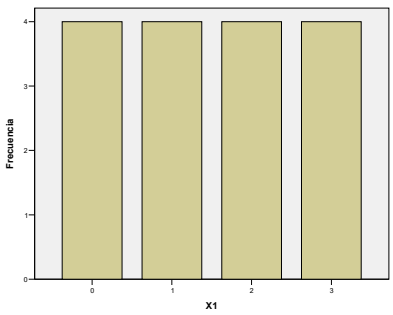
1.1. Variables categòriques: la moda

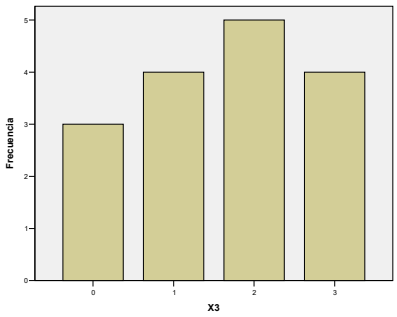
• La moda d'una variable X (Mo_X) és un estadístic de tendència central que s'obté com el valor que més es repeteix en el conjunt de dades corresponent a aquesta variable, això és, el que té la freqüència absoluta (n_i) més alta en la seua distribució de freqüències.

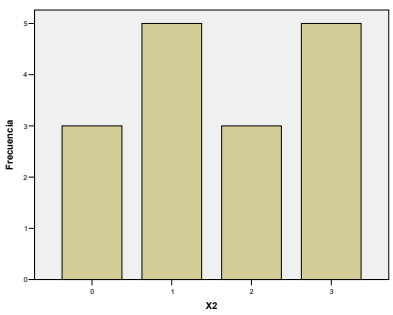
Exemple de càlcul de la moda per a les dades de la variable "Estat civil" obtingudes en una mostra de 50 persones de la ciutat de Castelló ($n = 50$):
Atès que el valor que més es repeteix és "Casat/ada" ($n_i = 20$), aquesta modalitat seria la moda de la variable.



• En funció del nombre de modes d'una distribució de freqüències, aquesta pot ser: amodal o uniforme, unimodal, bimodal i multimodal. Els següents gràfics de barres mostren exemples dels tres primers tipus:







Exemple d'obtenció de la moda en alguns casos particulars: Siguen les dades de la variable "Lloc de naixement" [0: Fora de la Comunitat Valenciana; 1: Província d'Alacant; 2: Província de València; 3: Província de Castelló] en dues mostres diferents de 16 subjectes residents a la Comunitat Valenciana:

X1: {0; 0; 1; 2; 1; 0; 1; 3; 2; 3; 1; 3; 0; 2; 3; 2} => $Mo_{X1} = \varnothing$ => Distribució amodal

X2: {1; 2; 0; 3; 1; 1; 0; 3; 3; 3; 1; 2; 0; 2; 1; 3} => Mo_{X2} = 1 i 3 => Distribució bimodal

• La moda, a més de descriure variables categòriques, es pot fer servir també amb variables ordinals i quantitatives, i el mateix es pot dir d'altres índexs estadístics que seran presentats per a variables categòriques en temes successius. En qualsevol cas, cal aclarir que aquests no són els que millor resumeixen la informació continguda en les variables ordinals i quantitatives.

1.2. Variables ordinals: la mediana

• Anàlogament al que s'assenyalava per a les variables categòriques, la mediana es pot obtenir també per a descriure variables quantitatives, si bé no és l'índex estadístic que millor resumeix la tendència central d'aquest tipus de variables.

• La mediana d'una variable *X* (Mdn_X) és el valor de la dada que, després d'ordenar totes les dades de la variable, de més petita a més gran, ocupa la posició del mig. Si hi ha un nombre parell de dades, s'obté com a mitjana aritmètica dels valors de les dues dades centrals.

Exemple:

X3: {6; 8; 1; 4; 2; 5; 6} => Mdn_{X3} {1; 2; 4; 5; 6; 6; 8} = 5

X4: {9; 6; 8; 1; 4; 2; 5; 6} => Mdn_{X4} {1; 2; 4; 5; 6; 6; 8; 9} = (5+6) / 2 = 5,5

Exemple d'obtenció de la mediana d'una variable amb les dades recollides a partir de la següent pregunta d'un test: "Ansietat que sent quan es troba amb molta gent al voltant". Les alternatives de resposta a aquesta qüestió eren: 1: Gens; 2: Poca; 3: Bastant; 4: Molta.

X_i	Freq. absoluta (n_i)	Percentatge ($\%_i$)	Freq. absoluta acumulada ($n_{a\,).}$	Percentatge acumulat ($\%_{a\,).}$
1	23	19,0	23	19
2	36	29,7	59	48,7
3	47	38,9	106	87,6
4	15	12,4	121	100
	121	100		

La mediana és el valor que, després d'ordenar les 121 dades, ocupe la posició central, en aquest cas, la posició 61. Així, en aquest exemple la mediana és igual a 3 (Bastant).

• En el cas, com en l'exemple anterior, d'una distribució de freqüències, el més senzill, abans que desagregar la distribució de freqüències per a veure quin és el valor central, és fixar-se en la columna de percentatges acumulats: la mediana serà el valor de la variable el %a del qual siga igual al 50 %, o bé el superior al 50 % més petit. Així, en l'exemple anterior de la pregunta d'un test, la mediana seria igual a 3 (Bastant), pel fet que el seu %a (87,6) és el superior al 50 % més petit.

1.3. Variables quantitatives: la mitjana aritmètica i altres alternatives resistents

1.3.1. La mitjana aritmètica.

• La mitjana [aritmètica] d'una variable X ($\overline{X}$ o μ_X) és un índex estadístic de tendència central que s'obté en sumar els valors de les dades de la variable i dividir pel nombre de casos (n):

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

Exemple: si X:{2; 3; 2; 7}, llavors $\overline{X}$ = (2+3+2+7) / 4 = 3,5

• Si tenim les dades agrupades en una distribució de freqüències, el càlcul de la mitjana suposa sumar el producte de cada valor per la corresponent freqüència absoluta i dividir el resultat pel nombre de casos:

$$\overline{X} = \frac{\sum X_i \cdot n_i}{n}$$

• Una fórmula equivalent a l'anterior consisteix a sumar el producte de cada valor de la variable per la seua freqüència relativa (proporció):

$$\overline{X} = \sum X_i \cdot p_i$$

Exemple de la variable "Temps emprat (en segons) a recórrer un laberint" per una mostra de 20 rates (n = 20).

Temps	n_i	p_i
9	3	0,15
10	8	0,4
11	6	0,3
12	2	0,1
13	1	0,05

$$\bar{X} = \frac{\sum X_i \cdot n_i}{n} = \frac{9 \cdot 3 + 10 \cdot 8 + 11 \cdot 6 + 12 \cdot 2 + 13 \cdot 1}{20} = 10,5$$

o, també, $\bar{X} = \sum X_i \cdot p_i = 9 \cdot 0,15 + 10 \cdot 0,4 + 11 \cdot 0,3 + 12 \cdot 0,1 + 13 \cdot 0,05 = 10,5$

• Si el que tenim és una distribució de freqüències amb els valors agrupats en intervals, podem calcular la mitjana a partir dels valors centrals de cada interval o "Marca de classe".

Exemple de la variable "Pes" a partir d'una distribució de freqüències en què les dades han sigut agrupades en intervals:

Pes (Kg.)	**Marca de classe**	n_i	n_a	p_i	p_a
40 – 50	45	5	5	0,086	0,086
50 – 60	55	10	15	0,172	0,258
60 – 70	65	21	36	0,362	0,620
70 – 80	75	11	47	0,190	0,810
80 – 90	85	5	52	0,086	0.896
90 – 100	95	3	55	0,052	0,948
100 – 110	105	3	58	0,052	1

$$\bar{X} = \frac{\sum X_i \cdot n_i}{n} = \frac{45 \cdot 5 + 55 \cdot 10 + 65 \cdot 21 + 75 \cdot 11 + 85 \cdot 5 + 95 \cdot 3 + 105 \cdot 3}{58} = 68,8$$

o també:

$$\bar{X} = \sum X_i \cdot p_i = 45 \cdot 0,086 + 55 \cdot 0,172 + 65 \cdot 0,362 + 75 \cdot 0,190 + 85 \cdot 0,086 + 95 \cdot 0,052 + 115 \cdot 0,052 = 68,8$$

Exercici 1: Obtingueu la mitjana, la mediana i la moda de la següent distribució de freqüències de la variable "Nombre de fills/es" obtinguda en una mostra de 200 famílies.

X_i	n_i
0	40
1	80
2	60
3	20

Exercici 2: Inventeu un conjunt de 7 dades que tinguen mitjana aritmètica igual a 9, mediana igual a 10 i moda igual a 11. Hi ha diferents solucions. Utilitzeu valors sencers, sense decimals.

• Si per a una mateixa variable disposem de mitjanes aritmètiques obtingudes en diferents grups de subjectes, es pot obtenir la mitjana total que resultaria d'ajuntar les dades d'aquests grups mitjançant el càlcul de la mitjana aritmètica ponderada:

$$\overline{X}_T = \frac{n_1 \cdot \overline{X}_1 + n_2 \cdot \overline{X}_2 + \cdots + n_k \cdot \overline{X}_k}{n_1 + n_2 + \cdots + n_k}$$

Exemple: Tenim 3 grups dels quals coneixem la nota mitjana en Avaluació Psicològica: la del grup A ($n = 160$) ha sigut de 7,8; la del grup B ($n = 110$), de 5,7, i la del grup C ($n = 148$), de 6,7. Quina seria la nota mitjana si s'ajuntaren els 3 grups?

$$\overline{X} = \frac{160 \cdot 7,8 + 110 \cdot 5,7 + 148 \cdot 6,7}{418} = 6,86$$

Exercici 3: Un grup de 50 estudiants d'Estadística està format per un 64 % d'estudiants de 1a matrícula, un 20 % de 2a matrícula i un 16 % de 3a matrícula o posterior. Les mitjanes dels 3 grups en l'avaluació final de l'assignatura van ser iguals a 7,2, 6,3 i 5,9, respectivament. Quina és la mitjana aritmètica del grup total?

1.3.2. Algunes anotacions sobre l'aplicació de la mitjana aritmètica

(1) La mitjana aritmètica també és aplicada en la pràctica a variables ordinals –fet que pot resultar qüestionat per raons teòriques sobre les quals no entrarem ací. D'altra banda, l'obtenció de la mitjana d'una variable ordinal suposarà obtenir, en no poques ocasions, valors que no es troben en el rang de valors de l'escala de mesura de la variable. En qualsevol cas, la seua aplicació és bastant freqüent en la pràctica perquè té algun avantatge com el que es comentarà en el següent epígraf i, d'altra banda, que obtinguem un valor que no coincidisca amb cap dels valors originals de la variable pot no representar un inconvenient important en moltes ocasions a l'hora de comunicar els resultats.

A tall d'exemple, el càlcul de la mitjana sobre la variable ordinal "Ansietat que sent quan es troba amb molta gent al voltant" dona com a resultat 2,45, que no és cap dels valors que pot tenir aquesta (1: Gens; 2: Poca; 3: Bastant; 4: Molta). No obstant això, és d'esperar que això siga entès per la majoria de les audiències a les quals ens podem dirigir com un nivell d'ansietat a mig camí entre "Poca" i "Bastant".

Com a contrapartida a aquest possible inconvenient, un avantatge de l'aplicació de la mitjana sobre les variables ordinals és que aquesta és capaç de captar amb més precisió que la mediana la informació relativa a la tendència central de les dades. Això és pel fet que té en compte totes les dades, no sols el valor de la dada que ocupa la posició central, com és el cas de la mediana.

Per exemple, siguen X i Y dues variables ordinals: X:{2; 3; 4; 4; 5; 5; 5; 6; 6} i Y :{2; 3; 4; 4; 5; 5; 5; 6; 7} que, com pot observar-se, només es diferencien en una dada. Si es calcula la mitjana i la mediana en totes dues variables s'observarà com la mediana n'és la mateixa per a les dues (Mdn = 5) i, contràriament, la mitjana sí que capta la diferència existent entre tots dos conjunts de dades ($\overline{X}$ = 4,44; $\overline{Y}$ = 4,55).

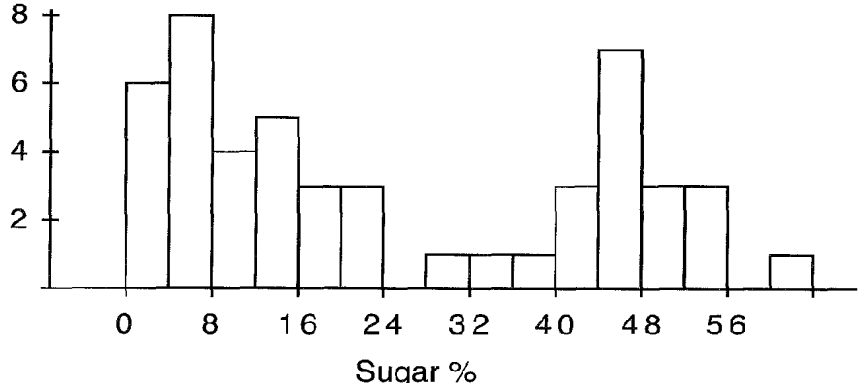
(2) El fet que la mitjana d'una variable ordinal puga ser un valor que no es trobe entre els valors possibles de la variable, pot succeir també amb variables quantitatives discretes.

Exemple: la mitjana de la variable quantitativa discreta "Nombre de fills/es" a partir de la distribució de freqüències que es mostra a continuació és igual a 1,65 fills/es.
=> És comprensible aquest valor per a un lector? – En cas afirmatiu, es pot reportar la mitjana; altrament, millor reportar-ne la mediana.

X_i	Freq. absoluta (n_i)	Freq. relativa (p_i)	Percentatge ($\%_i$)
0	3	0,15	15
1	6	0,30	30
2	7	0,35	35
3	3	0,15	15
4	1	0,05	5
	20	1	100

(3) Quan la forma de la distribució d'una variable quantitativa siga bimodal o multimodal, l'obtenció de la mitjana i la mediana no és adequada. En aquests casos, informar sobre les modes d'aquesta variable és l'opció més convenient.

Exercici 4: A partir de l'histograma de la distribució de la variable "Contingut de sucre en %", mesurada en 49 marques de cereals que es presenta més avall, reconstrueix la distribució de freqüències corresponent i calculeu la mitjana i la mediana d'aquesta. Són els dos estadístics un bon resum de la tendència central d'aquesta distribució?



1.3.3. Alternatives resistents a la mitjana aritmètica

• La mitjana és molt sensible a dades anòmales, atípiques o extremes, és a dir, dades que són de magnitud bastant diferent de la majoria. Generalment, en distribucions de freqüències molt asimètriques apareixen dades anòmales, és a dir, molt allunyades del centre de la distribució per una de les cues de la distribució.

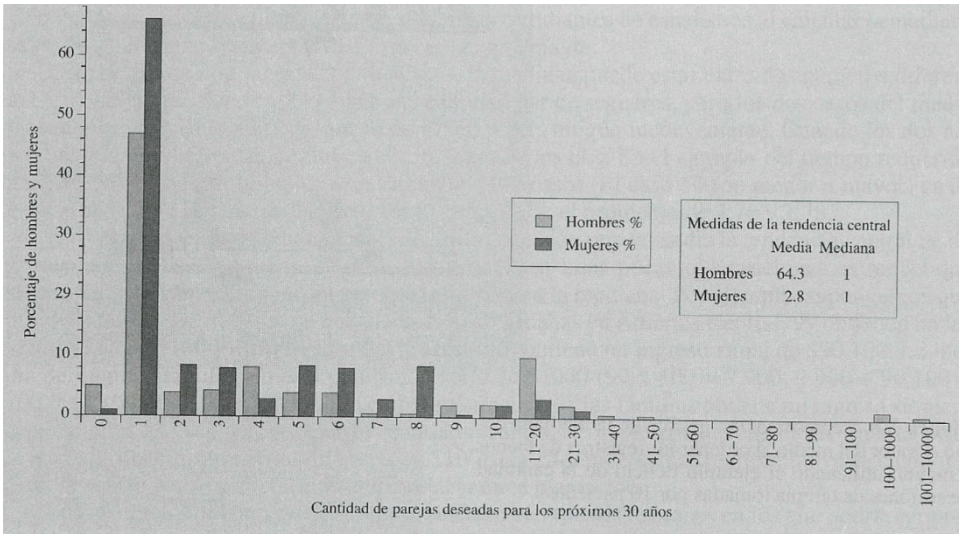
Exemple:

$$X5:\{8;\ 8;\ 9;\ 10;\ 10;\ 12;\ 14\} \rightarrow \quad \bar{X} = 10{,}14;\ \ Mdn = 10$$
$$X6:\{8;\ 8;\ 9;\ 10;\ 10;\ 12;\ 50\} \rightarrow \quad \bar{X} = 15{,}28;\ \ Mdn = 10$$

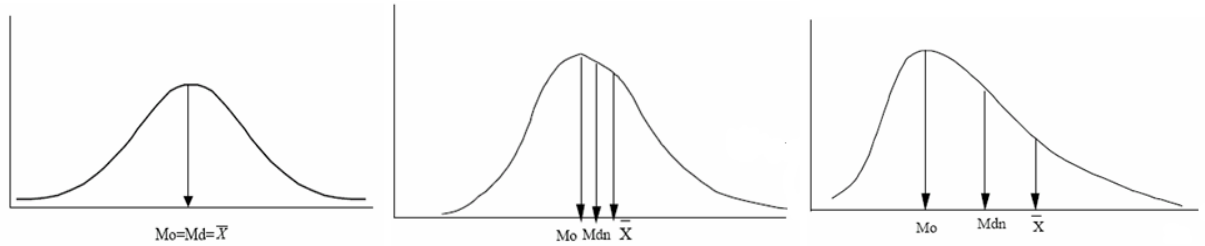
La variable *X6* conté una dada anòmala (el valor 50) (tal vegada motivada per un error en la recollida o en el processament de les dades) i, en conseqüència, la seua mitjana s'ha incrementat bastant, resultant poc representativa del gruix de les dades. La mediana, en canvi, no és afectada per aquesta dada atípica i el seu valor és el mateix en ambdues variables.

Exemple de l'estudi de Millar i Fishkin (1997) sobre la base evolutiva de l'elecció de la parella humana, sota la hipòtesi que els homes prefereixen tenir més parelles que les dones al llarg de la vida. Davant la pregunta de quin seria el nombre ideal de parelles al llarg d'un període de 30 anys, es van obtenir com a resultats una mitjana aritmètica de 64,3 parelles per als homes i de 2,8 per a les dones, resultats que evoquen immediatament conclusions dignes d'una primera pàgina en la premsa sensacionalista.

No obstant això, si observem detingudament un gràfic de barres de la distribució de freqüències de tots dos grups podem entendre el perquè d'una mitjana tan elevada en els homes...



• Com a conseqüència de la sensibilitat de la mitjana a les dades anòmales o atípiques, únicament coincidiran els valors de la mitjana, la mediana i la moda en una distribució simètrica unimodal. A mesura que la distribució siga més asimètrica, aquests índexs s'allunyaran i el valor de la mitjana es desplaçarà cap als valors anòmals o atípics de la distribució.



• En variables amb distribucions que continguen dades anòmales (=>distribucions asimètriques amb valors molt allunyats del centre de la distribució per una de les cues de la distribució) és més convenient aplicar índexs que no es vegen tan afectats per aquests valors atípics. Entre aquests índexs, anomenats estadístics resistents, es troba la mediana (ja vam estudiar anteriorment que el càlcul d'aquest estadístic no és afectat pels valors anòmals) i la mitjana retallada.

• La mitjana retallada consisteix en l'obtenció de la mitjana aritmètica excloent del càlcul un percentatge dels casos situats en els extrems superior i inferior de la distribució, és a dir, retallant en un percentatge determinat les cues de la distribució. Per exemple, la mitjana retallada al 10 % exclou del càlcul al 10 % inferior i al 10 % superior dels valors obtinguts.

Per exemple, els valors de X obtinguts en un grup de 10 subjectes són:

$$X:\{7;\ 8;\ 9;\ 10;\ 10;\ 12;\ 14;\ 16;\ 19;\ 57\} \quad \rightarrow \quad \overline{X} = 16{,}2$$

La presència d'una dada anòmala en la distribució (el valor 57) fa poc recomanable l'obtenció de la mitjana aritmètica. En aquest cas, és més aconsellable obtenir la mediana (11), o bé, la mitjana retallada. La mitjana retallada al 10 % s'obtindria eliminant del càlcul el 10 % inferior i el 10 % superior dels valors; en aquest exemple, un valor de cada cua de la distribució (els valors 7 i 57) i, d'aquesta manera, s'obtindria un valor de la mitjana igual a 12,25. Quin seria el resultat de la mitjana retallada al 20 % per a les dades anteriors?

• Altres estadístics resistents que no descriurem ací, però dels quals, almenys de moment, és interessant conéixer els noms són: la mitjana winsoritzada, la trimitjana, l'estimador M de Andrews i l'estimador M de Tukey.

2. Altres estadístics de posició grupal

• A continuació es presenta un altre conjunt d'estadístics que permeten descriure la posició o localització de les dades d'una variable, però que, a diferència dels estudiats en l'apartat anterior, no tenen per objecte proporcionar un valor que represente el centre de la distribució. Una característica comuna dels estadístics del present apartat és que podran ser obtinguts amb variables ordinals i quantitatives, però no amb variables categòriques.

2.1. El mínim i el màxim

• El mínim és el valor observat més baix de les modalitats que adopta una variable; complementàriament, el màxim és el valor observat més alt. Tots dos valors permeten fer-se una idea d'entre quins valors de l'escala de mesura d'una variable es localitzen les dades.

> **Exemple**: Per a la distribució de la variable "Temps emprat (en segons) a recórrer un laberint", presentada en l'apartat "La mitjana aritmètica", el mínim i el màxim serien 9 i 13 segons, respectivament.

Exercici 5: obtingueu el mínim i el màxim de les variables *X1*, *X2*, *X3* i *X4* dels exemples vistos en els apartats precedents.

2.2. Els quantils

• Es defineix com a quantil k (C_k) el valor de la variable tal que un $k\%$ dels subjectes tenen un valor inferior o igual a aquest valor (k pot ser qualsevol número entre 0 i 100, sencer o decimal).

• La mediana és un cas particular de quantil, en concret, el quantil 50 (C_{50}).

• El càlcul d'un determinat quantil k d'una variable resulta relativament senzill d'obtenir a partir de la distribució de freqüències d'aquesta variable si ens fixem en la columna dels percentatges acumulats: el C_k correspondrà al valor de la variable el percentatge acumulat del qual siga igual a k o, si escau, el percentatge acumulat major a k que siga més petit.

> Per posar-ne un **exemple** suposem que obtenim dades en una mostra de 200 treballadors/es a partir de la següent pregunta d'un test de cultura organitzacional: "Es valora en els treballadors/es la creativitat i la capacitat d'innovació". L'escala de resposta és tipus Likert,

des d'1 (Molt en desacord) a 7 (Molt d'acord). La distribució de freqüències de les dades obtingudes és la següent:

X_i	n_i	$\%_i$	$\%_a$
2	21	10,5	10,5
3	31	15,5	26
4	36	18	44
5	47	23,5	67,5
6	38	19	86,5
7	27	13,5	100
	200	100	

En aquest exemple:

- el quantil 67,5 d'aquesta variable és igual a 5 => $C_{67,5} = 5$

(és a dir, que un 67,5 % de subjectes van contestar amb el valor 5 o inferior).

- el quantil 40 d'aquesta variable és igual a 4 => $C_{40} = 4$

(és a dir, que un 40 % de subjectes van contestar amb el valor 4 o inferior).

• Interpretació dels quantils:

– El valor d'un determinat quantil indica quin percentatge de casos tenen valors iguals o inferiors a aquest valor en la variable. Per exemple, suposem que tenim les dades de la variable "Pes" en una mostra de ratolins, el C_{90} és el valor de pes tal que un 90 % dels ratolins tenen un pes inferior o igual a aquest valor.

– Complementàriament, el valor d'un determinat quantil indica el percentatge de casos que en tenen un valor per sobre d'aquest. Així, per a l'exemple anterior, sabem que un 10 % dels ratolins tenen un pes superior al pes que correspon al C_{90}.

– També és possible plantejar-se qüestions relatives a intervals de valors de la distribució de freqüències. Per exemple, entre quins valors es troba el 50 % central dels ratolins en la distribució de la variable Pes? => La resposta vindrà donada pels valors corresponents als quantils C_{25} i C_{75}.

Exercici 6: a partir de les dades de la pregunta "Ansietat que sent quan es troba amb molta gent al voltant" presentades anteriorment (vegeu l'apartat 1.2), obtingueu el quantil 48,7 ($C_{48,7}$), el quantil 50 (C_{50}), el quantil 19 (C_{19}), el quantil 20 (C_{20}) i el quantil 75 (C_{75}). Quin percentatge de subjectes hi ha per sobre del C_{75}? I entre el C_{20} i el C_{80}?

Exercici 7: A partir de les dades recollides amb la pregunta "Es valora en els empleats la creativitat i la capacitat d'innovació" (veure distribució de freqüències més amunt), obtingueu el quantil 5 (C_5), el quantil 25 (C_{25}), el quantil 45 (C_{45}) i el quantil 65 (C_{65}).

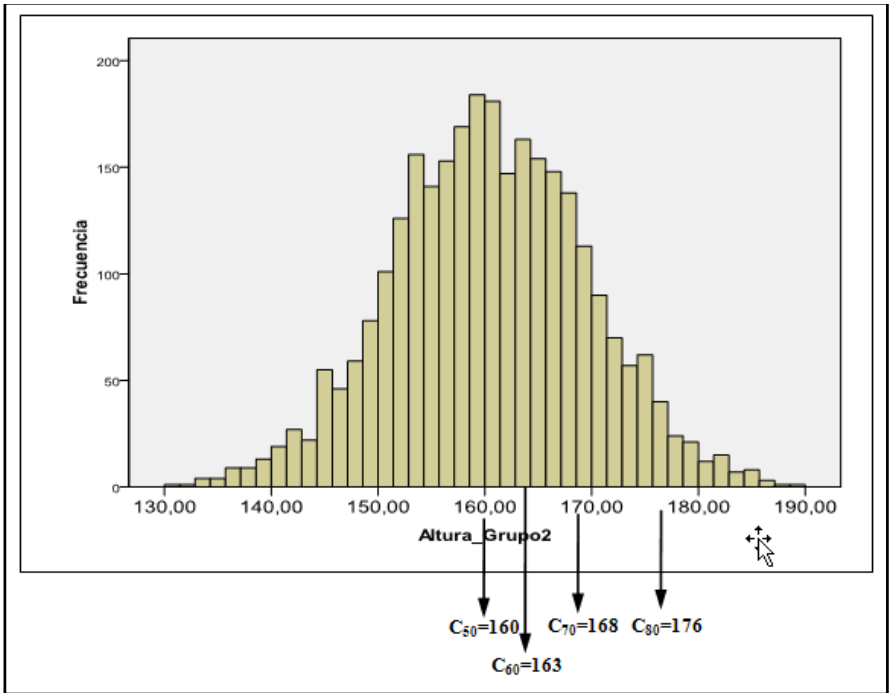
• Cal tenir en compte sobre la distància entre els quantils que entre parells de quantils equidistants existeix el mateix percentatge de subjectes. Per exemple, entre el C_{10} i el C_{20} existeix un 10 % de casos, el mateix que n'hi ha entre el C_{20} i el C_{30}, i això és cert en la distribució de qualsevol variable. No obstant això, les distàncies entre quantils no tenen per què ser constants en termes de distàncies entre els valors de la variable.

Prenent com a exemple la distribució de freqüències de la pregunta del test de cultura organitzacional, en el qual:

quantil 5 (C_5) → 2
quantil 25 (C_{25}) → 3
quantil 45 (C_{45}) → 5
quantil 65 (C_{65}) → 5

Es pot observar que entre els valors que corresponen al C_5 i al C_{25} la distància és d'1 unitat, i entre aquells que corresponen al C_{25} i al C_{45} és de 2 unitats, i entre aquells que corresponen al C_{45} i al C_{65} és de 0 unitats. En canvi, entre cadascun d'aquests quantils hi ha sempre un 20 % de subjectes.

Un altre **exemple**: distribució dels valors de la variable "Alçada (cm)" en una mostra de xiquets/es de 14 anys.



• Tipus específics de quantils molt utilitzats en la difusió de resultats:

- Els centils o percentils (P_k): fan referència als quantils 1 a 99, això és, el valor de k és un nombre sencer comprés entre 1 i 99. Percentils possibles: $P_1, P_2, P_3, P_4 \ldots P_{99}$ => divideixen la distribució de la variable en 100 parts, de les quals cadascuna conté l'1 % dels casos.

- Els quartils (Q_k): fan referència als quantils 25 (Q_1), 50 (Q_2) i 75 (Q_3). => divideixen la distribució de la variable en 4 parts, de les quals cadascuna conté el 25 % dels casos.

- Els decils (D_k): fan referència als quantils 10 (D_1), 20 (D_2), 30 (D_3) … a 90 (D_9) => divideixen la distribució de la variable en 10 parts, de les quals cadascuna conté el 10 % dels casos.

Taula d'equivalències entre quantils

P_{10}	→	D_1	→	→	→	C_{10}
P_{20}	→	D_2	→	→	→	C_{20}
P_{25}	→	→	→	Q_1	→	C_{25}
P_{30}	→	D_3	→	→	→	C_{30}
P_{40}	→	D_4	→	→	→	C_{40}
P_{50}	→	D_5	→	Q_2	→	C_{50}
P_{60}	→	D_6	→	→	→	C_{60}
P_{70}	→	D_7	→	→	→	C_{70}
P_{75}	→	→	→	Q_3	→	C_{75}
P_{80}	→	D_8	→	→	→	C_{80}
P_{90}	→	D_9	→	→	→	C_{90}

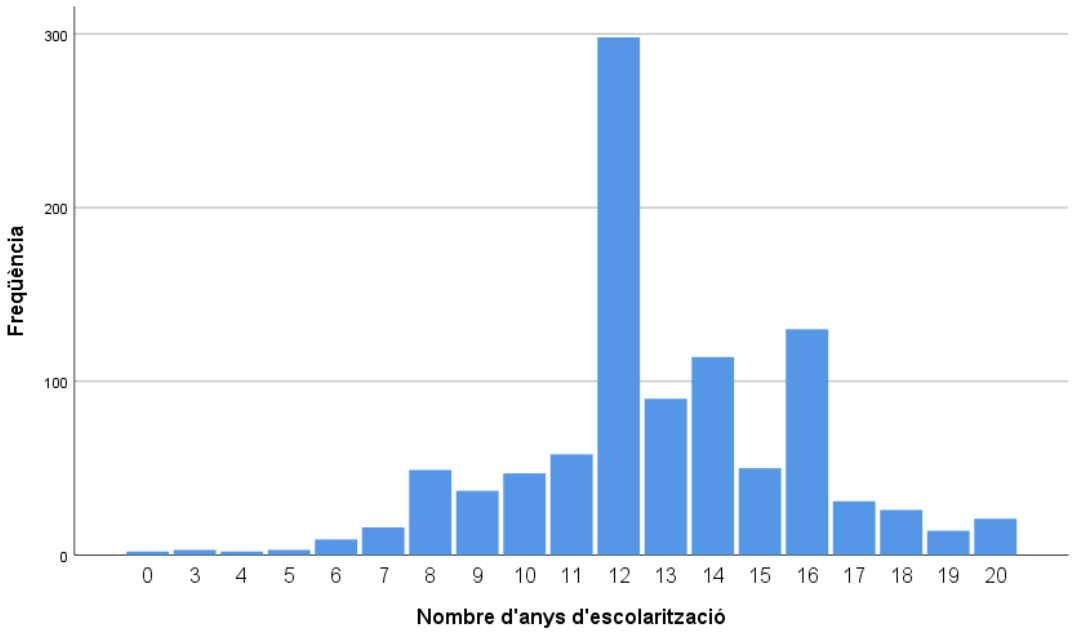
Exercici 8: A partir de les dades recollides amb la pregunta "Es valora en els empleats la creativitat i la capacitat d'innovació" (vegeu distribució de freqüències més amunt), calculeu el mínim i el màxim, la moda, la mediana, el Q_3, el $C_{10,5}$, el P_3 i el D_9.

Exercici 9: Tenim una variable X de la qual s'ha recollit informació de 500 subjectes. Quants subjectes estaran entre el Q_1 i el Q_3?, i entre el C_{10} i el C_{90}?, i entre el D_4 i el P_{60}?

Exercici 10: A partir de la distribució de freqüències de la variable "Nombre d'anys d'escolarització", en una mostra de 1000 subjectes adults de la ciutat d'Elx (n = 1000): (1) obtingueu els següents estadístics de posició: la moda, el mínim i el màxim, la mitjana aritmètica, la mediana, i els quantils P_{10}, D_2, P_{30}, D_9 i Q_3; (2) en funció del tipus de variable (categòrica, ordinal o quantitativa) i de la forma de la seua distribució de freqüències (vegeu diagrama de barres), raoneu quin estadístic de tendència central resultarà més convenient per a descriure-la.

Nombre d'anys d'escolarització

		Freqüència	Percentatge	Percentatge vàlid	Percentatge acumulat
Vàlid	0	2	,2	,2	,2
	3	3	,3	,3	,5
	4	2	,2	,2	,7
	5	3	,3	,3	1,0
	6	9	,9	,9	1,9
	7	16	1,6	1,6	3,5
	8	49	4,9	4,9	8,4
	9	37	3,7	3,7	12,1
	10	47	4,7	4,7	16,8
	11	58	5,8	5,8	22,6
	12	298	29,8	29,8	52,4
	13	90	9,0	9,0	61,4
	14	114	11,4	11,4	72,8
	15	50	5,0	5,0	77,8
	16	130	13,0	13,0	90,8
	17	31	3,1	3,1	93,9
	18	26	2,6	2,6	96,5
	19	14	1,4	1,4	97,9
	20	21	2,1	2,1	100,0
	Total	1000	100,0	100,0	



• Sobre l'obtenció de quantils en variables quantitatives contínues: Per tenir en compte el caràcter continu d'aquest tipus de variables se sol aplicar una fórmula que permet obtenir d'una manera precisa quin seria el valor exacte que correspon a un determinat quantil. La fórmula la podem trobar en el text de Botella *et al*. (2001, p. 70), si bé se'n pot estimar amb bastant precisió per interpolació a partir de la columna dels valors de la variable (X) i la dels percentatges acumulats ($\%_a$).

Exemple: Tenim la distribució de freqüències del "Temps (en segons) emprat a completar un laberint" en una mostra de 20 rates (n = 20). Com que el C_{85}=11 i el C_{95}=12, el valor del C_{90} seria igual a 11,5. Quins valors, aproximadament, corresponen als quantils 35, 25, 50, 70, 75 i 96?

Temps	n_i	p_i	p_a	$\%_a$
9	3	0,15	0,15	15
10	8	0,4	0,55	55
11	6	0,3	0,85	85
12	2	0,1	0.95	95
13	1	0,05	1	100

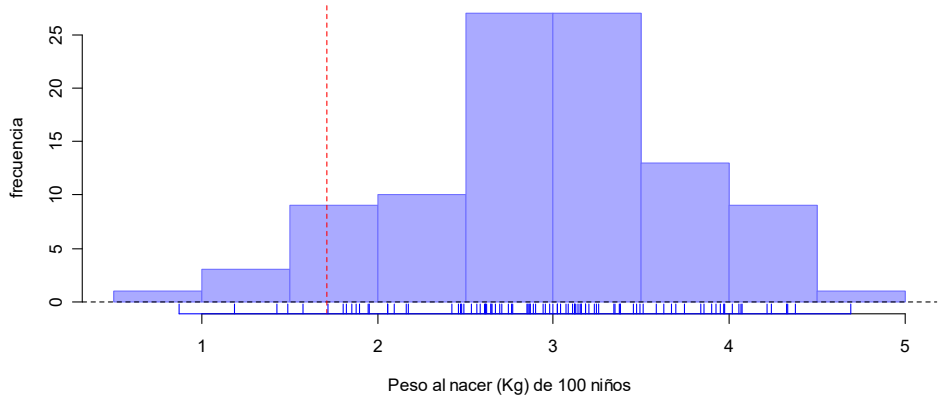
Exercici 11: A continuació es mostra la distribució de freqüències de la variable "Distància" que es va obtenir després de mesurar, en una mostra de 108 subjectes, la distància (en mil·límetres) del centre de la pituïtària a la fissura pterigomaxil·lar. Obtingueu, a partir d'aquesta, el valor dels següents quantils: $C_{68,5}$, P_4, D_6, Q_2, P_{87}, C_{90}, $C_{2,8}$ *i* P_{99}.

Distància

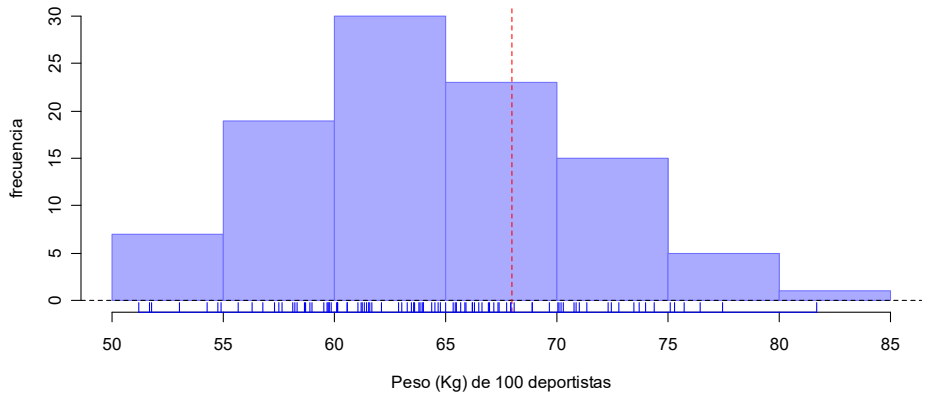
		Freqüència	Percentatge	Percentatge vàlid	Percentatge acumulat
Vàlid	16,5	1	,9	,9	,9
	17,0	1	,9	,9	1,9
	19,0	2	1,9	1,9	3,7
	19,5	1	,9	,9	4,6
	20,0	4	3,7	3,7	8,3
	20,5	2	1,9	1,9	10,2
	21,0	5	4,6	4,6	14,8
	21,5	9	8,3	8,3	23,1
	22,0	4	3,7	3,7	26,9
	22,5	7	6,5	6,5	33,3
	23,0	11	10,2	10,2	43,5
	23,5	7	6,5	6,5	50,0
	24,0	6	5,6	5,6	55,6
	24,5	8	7,4	7,4	63,0
	25,0	6	5,6	5,6	68,5
	25,5	6	5,6	5,6	74,1
	26,0	7	6,5	6,5	80,6
	26,5	4	3,7	3,7	84,3
	27,0	2	1,9	1,9	86,1
	27,5	3	2,8	2,8	88,9
	28,0	4	3,7	3,7	92,6
	28,5	1	,9	,9	93,5
	29,0	1	,9	,9	94,4
	29,5	1	,9	,9	95,4
	30,0	1	,9	,9	96,3
	31,0	3	2,8	2,8	99,1
	31,5	1	,9	,9	100,0
	Total	108	100,0	100,0	

Exercici 12: Tot seguit es mostren els histogrames de 3 variables quantitatives contínues (Baró-López, 2005). En cadascun apareixen una o vàries línies verticals discontínues que marquen un determinat quantil de la distribució. Contesteu les qüestions que es plantegen en cada cas.

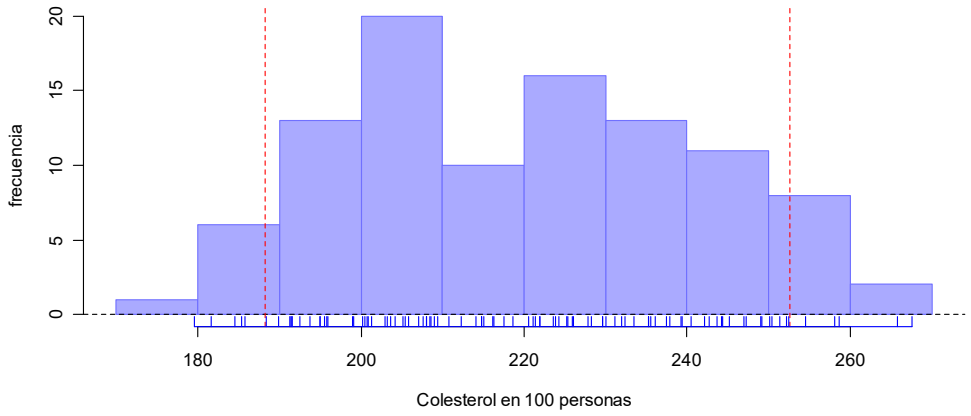
– A continuació es presenta l'histograma de la variable "Pes en nàixer (kg)" en una mostra de 100 xiquets/es. (El percentil 5 d'aquesta distribució està marcat per una línia vertical discontínua.) A quin valor de pes correspon, aproximadament, aquest percentil? Com s'interpreta aquest percentil?



– En el següent histograma de la variable "Pes" en una mostra de 100 esportistes, la línia vertical discontínua indica el valor de pes que es superat pel 25 % dels esportistes més pesats. Aquesta línia correspon, per tant, al percentil____ . Quin és, aproximadament, el valor de pes que correspon a aquest percentil? Quants esportistes de la mostra estan per sobre d'aquest pes?



– En un estudi s'ha obtingut la distribució de freqüències de la taxa de colesterol en sang en una mostra de 100 subjectes de la població espanyola i en l'histograma està representat el 90 % central dels valors obtinguts entre les dos línies verticals. El 90 % central de la distribució de dades es troba entre el percentil____ i el percentil____. A quins valors concrets de colesterol corresponen, aproximadament, aquests percentils (vegeu les línies discontínues)? Quants subjectes es troben fora d'aquest interval?

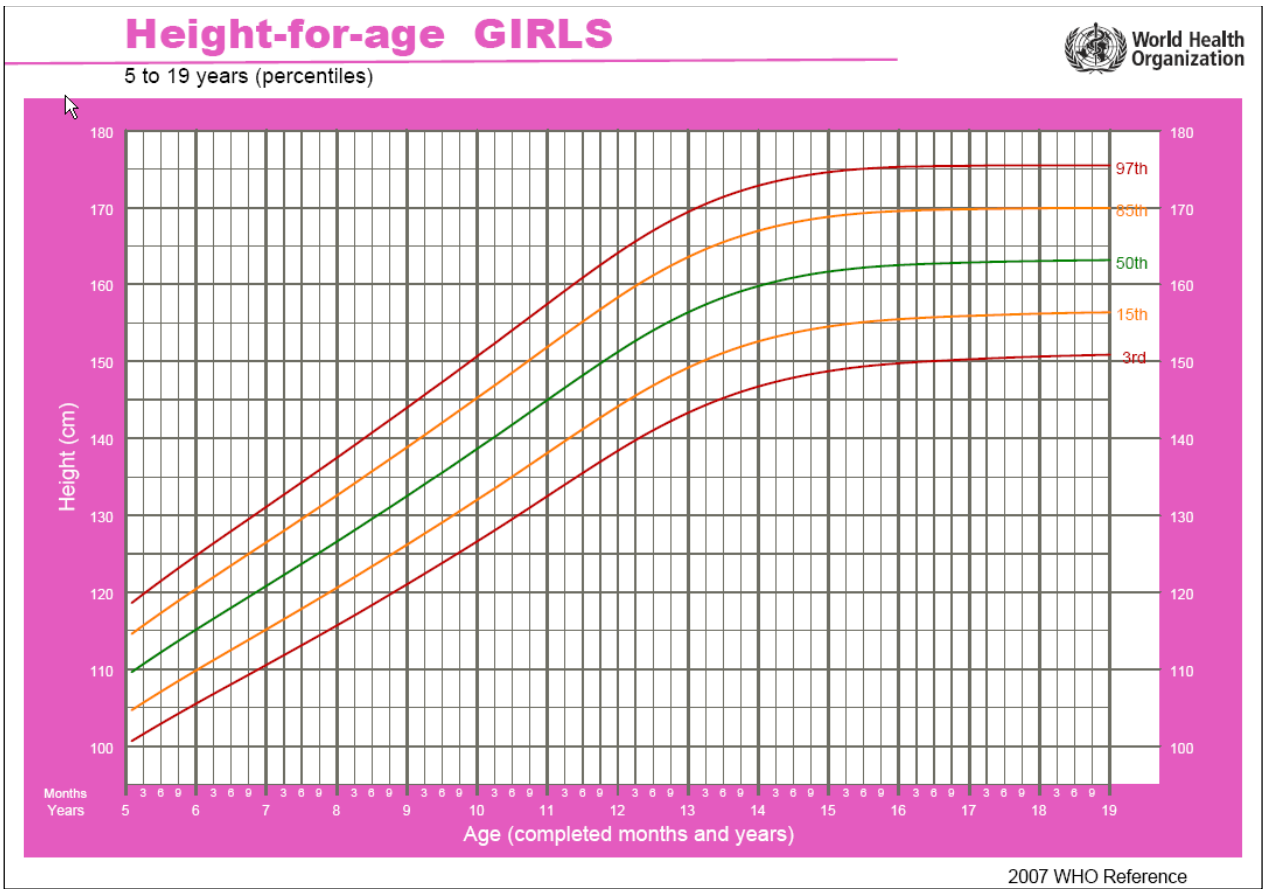


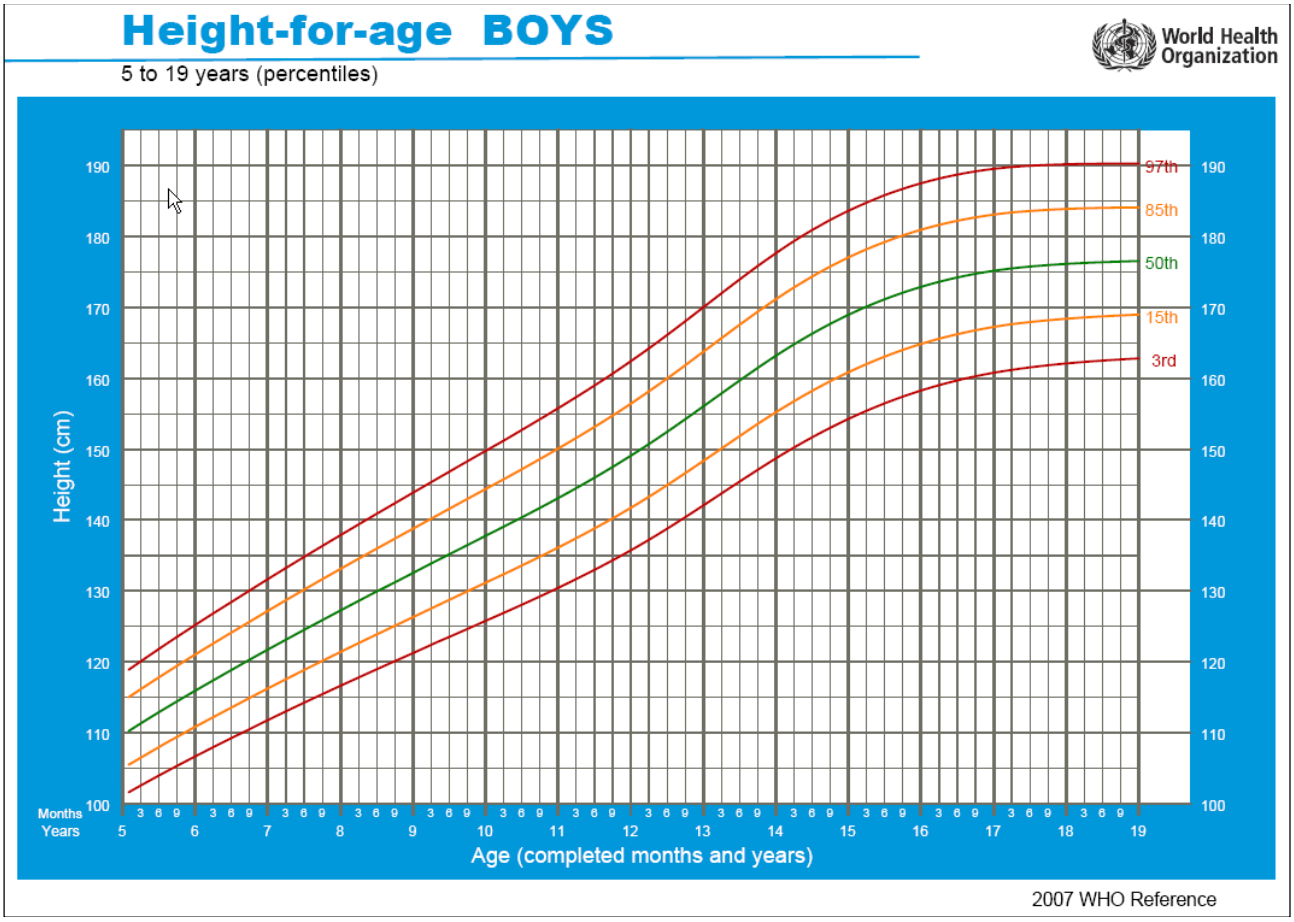
Exercici 13: A continuació es presenten les taules de creixement de l'OMS per a la variable "Alçada" de les xiques i xics de 5 a 19 anys. A partir d'aquestes contesteu les següents preguntes:

a) A continuació es mostren les dades d'alçada d'Ester entre els 5 i 19 anys. Marqueu (amb punts) sobre la taula corresponent les dades d'Ester i comenta com n'ha sigut l'evolució.

Edat (anys)	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Alçada (cm)	105	110	113	120	130	134	145	150	160	162	167	167	168	170	170

b) Carlos té 17 anys i la seua alçada és de 175 cm, quin percentil li correspon? Quin percentil ocupava quan tenia 8 anys?
c) Isabel té 15 anys i té una alçada de 160 cm, quin percentil li correspon?
d) Juan té 18 anys i a la seua alçada li correspon el P_{85}, quina es la seua alçada?
e) Una persona de 19 anys que té una alçada de 175 cm, quin percentil ocupa si és una xica? I si és un xic?
f) Quina és la mediana d'alçada de la població de xiques de 9 anys d'edat?
g) Entre quins valors d'alçada es troba el 70 % central de la població de xics de 16 anys?
h) Si es considera que una puntuació inferior al P_{15} indica retard en el creixement, es diagnosticaria retard en el creixement a un xic de 10 anys si la seua alçada fora inferior a _____ cm.


Height-for-age GIRLS
5 to 19 years (percentiles)
World Health Organization
97th
85th
50th
15th
3rd
Height (cm)
Months
Years
Age (completed months and years)
2007 WHO Reference


Height-for-age BOYS
5 to 19 years (percentiles)
World Health Organization
97th
85th
50th
15th
3rd
Height (cm)
Months
Years
Age (completed months and years)
2007 WHO Reference

Referències

Barón-López, J. (2005). *Bioestadística: métodos y aplicaciones.* Apunts i material disponibles en http://www.bioestadistica.uma.es/baron/apuntes/

Botella, J., León, O. G., San Martín, R. & Barriopedro, M. I. (2001). *Análisis de datos en psicología I: teoría y ejercicios.* Madrid: Pirámide.

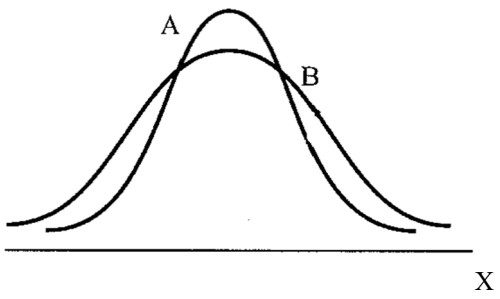
Millar, L. C. & Fishkin, S. A. (1997). Sobre la dinámica del éxito humano y el éxito reproductivo. En J. A. Simpson i D. T. Kendrick (Eds.): *Psicología Social Evolutiva.* Mahwah, NJ: LEA.

Tema 3.2 – Caracterització de grups: Estadístics de dispersió

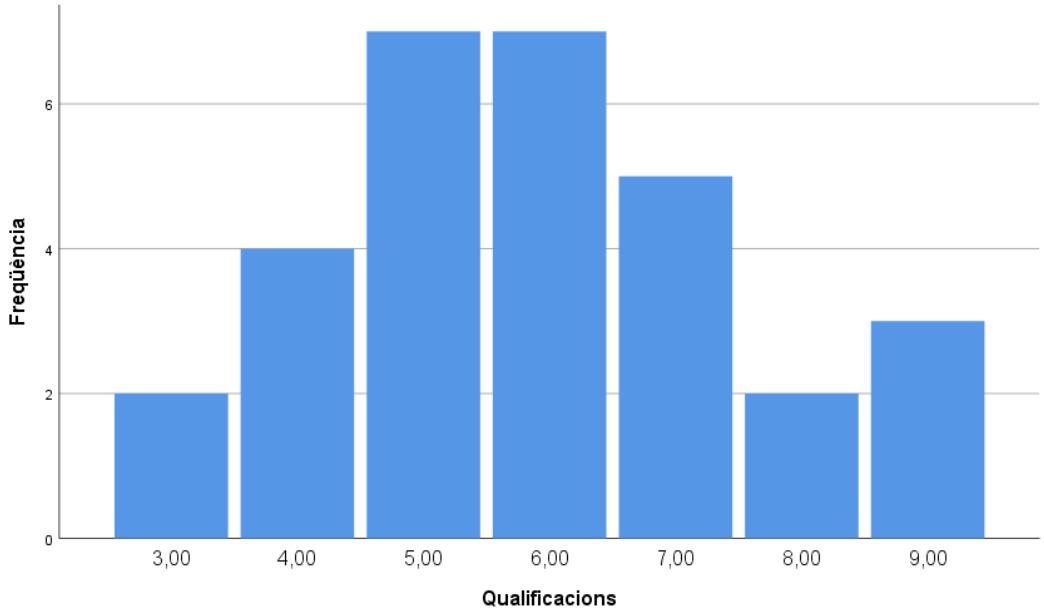
1. **Variables categòriques: l'índex de variació qualitativa**
2. **Variables ordinals: el recorregut i el recorregut interquartílic**
3. **Variables quantitatives: la variància, la desviació típica i el coeficient de variació**

• En diversos llibres d'Estadística es fa referència a la dispersió o variabilitat com la raó de ser d'aquesta disciplina; per exemple, de Veaux, Bock i Velleman (2003) afirmen de manera rotunda en el seu manual *Intro Stats* el següent: "Statistics is about variation". En efecte, si no existira heterogeneïtat o dispersió en les variables que estudiem no caldria utilitzar cap mètode estadístic: amb la dada d'un cas, coneixeríem el que ocorre per a la resta dels casos –en comptes de variables, tindríem, en realitat, constants.

• El concepte de dispersió fa referència al grau en què les dades d'una variable són més homogènies (=> menor dispersió o variabilitat) o més heterogènies (=> major dispersió o variabilitat). És un concepte que tal vegada puga captar-se d'una manera més intuïtiva gràficament: els següents polígons de freqüències suavitzats mostren gràficament la distribució d'una mateixa variable (*X*) en dos grups diferents de subjectes (A i B); observant-los, quin dels dos grups diríem que té major dispersió en aquesta variable?



• Origen de la variabilitat: la dispersió en els valors dels subjectes en una variable pot deure's a diferents causes, a les quals se sol fer referència com a fonts de variació de les dades. Per exemple, la variabilitat en les qualificacions d'Estadística dels estudiants del grup E d'un curs acadèmic recent (vegeu histograma), a què pot ser atribuïda? En aquest cas, una font de variabilitat fonamental serà el coneixement i domini de la matèria. És d'esperar que diferències individuals en aquest aspecte siguen la principal causa de la dispersió existent en les qualificacions de l'assignatura.



No obstant això, en el cas fictici que totes i tots els alumnes hagueren tingut el mateix domini i nivell de coneixements de l'assignatura, seria d'esperar que la nota haguera sigut la mateixa per a tots/es? –És més que probable que aquest no siga el cas. No és difícil pensar en altres possibles fonts de variació que tindran la seua influència sobre les qualificacions, per exemple, si s'ha dormit bé o malament la nit abans de l'examen, la capacitat per a afrontar situacions estressants, l'habilitat per a respondre el tipus de preguntes plantejades en l'examen (objectives, obertes…), la fiabilitat i la validesa de l'instrument de mesura (l'examen), etc.

• A continuació es presenten una sèrie d'índexs estadístics i representacions gràfiques orientats a descriure la dispersió d'una variable. Aquests apareixen diferenciats en 3 apartats que es corresponen amb la distinció plantejada per a les variables en funció de la seua escala de mesura.

1. Variables categòriques: l'índex de variació qualitativa

• L'índex de variació qualitativa (*IVC*) s'obté a través de la següent fórmula, on *k* és el nombre de modalitats de la variable i y p_i és la freqüència relativa associada a cadascuna d'aquestes:

$$IVC = \frac{1 - \sum p_i^2}{(k-1)/k}$$

• L'*IVC* expressa el grau en què els casos estan dispersos en les diferents modalitats de la variable; el seu màxim (*IVC* = 1) s'aconsegueix en cas que les freqüències relatives siguen iguals per a totes les modalitats de la variable (és a dir, una distribució uniforme, això és, de màxima dispersió). L'*IVC* seria igual a 0 quan la freqüència relativa d'una modalitat de la variable fora igual a 1, això és, quan tots els casos tingueren el mateix valor observat en la variable (dispersió nul·la).

> **Exemple**: Siga la variable "Religió que es professa" [Codificació: 0: Catòlica; 1: Protestant; 2: Una altra; 3: Cap] de la qual s'han obtingut dades en una mostra de 50 persones. La distribució de freqüències obtinguda és la següent:

X_i	*Freq. absoluta* (n_i)	*Freq. relativa* (p_i)
0	12	0,24
1	10	0,2
2	10	0,2
3	18	0,36
	50	1,00

El valor de l'*IVC* és igual a:

$$IVC = \frac{1 - (0.24^2 + 0.2^2 + 0.2^2 + 0.36^2)}{(4-1)/4} = 0.98$$

Exercici 1: Obtingueu l'*IVC* a partir de la distribució de freqüències de la variable "Estat civil" que es va presentar en un tema anterior i que apareix a continuació:

X_i	Freq. absoluta (n_i)	Freq. relativa (p_i)	Percentatge ($\%_i$)
solter/a	15	0,3	30
casat/a	20	0,4	40
separat/a	11	0,22	22
vidu/a	4	0,08	8
	50	1,00	100

Exercici 2: Inventeu dues distribucions de freqüències per a la variable "Estat civil" diferents a l'anterior (amb *n* = 50) en què l'*IVC* siga, respectivament, tan baix i tan alt com siga possible.

2. Variables ordinals: el recorregut interquartílic

2.1. El recorregut

• També denominat com a amplitud o rang, és la diferència entre el valor màxim i el valor mínim en una variable:

$$Recorregut = Màxim - Mínim$$

Exemple d'obtenció del recorregut per a les dades recollides amb la pregunta "Ansietat que sent quan es troba amb molta gent al voltant" (escala de resposta: 1: Gens; 2: Poca; 3: Bastant; 4: Molta).

X_i	n_i	$\%_i$	n_a	$\%_a$
1	23	19,0	23	19
2	36	29,7	59	48,7
3	47	38,9	106	87,6
4	15	12,4	121	100
	121	100		

$Recorregut = 4 - 1 = 3$

• El principal desavantatge del recorregut és que en basar-se en els valors mínim i màxim, si la distribució té valors atípics, el seu càlcul es veurà molt influït per aquests valors. Així doncs, el recorregut pot proporcionar valors que no siguen bons indicadors de la vertadera dispersió de les dades –per exemple, en la variable X :{8, 8, 9, 10, 10, 12, 50}, el recorregut és igual a 42 quan, en realitat, totes les dades, exceptuant-ne una, són bastant homogènies.

Exercici 3: Obtingueu el recorregut de la variable obtinguda a partir de les dades recollides, en una mostra de 200 treballadors/es, amb la següent pregunta: "Es valora en els treballadors/es la creativitat i la capacitat d'innovació". L'escala de resposta era tipus Likert des d'1 (Molt en desacord) a 7 (Molt d'acord).

X_i	n_i	$\%_i$	$\%_a$
2	21	10,5	10,5
3	31	15,5	26
4	36	18	44
5	47	23,5	67,5
6	38	19	86,5
7	27	13,5	100
	200	100	

• Pel que fa a la interpretació del recorregut, tant aquest com la resta d'índexs de variabilitat que es tractaran en els següents apartats (exceptuant, parcialment, el coeficient de variació) ofereixen resultats que no tenen una interpretació directa en termes absoluts. Així, què significa un recorregut de 4 o un recorregut de 10, molta o poca dispersió?

– L'únic cas en què la interpretació d'aquests índexs és inequívoca és quan són igual a 0, indicant l'absència de variabilitat en les dades –cas d'altra banda bastant excepcional. Valors majors que 0 indicaran dispersió en les dades; com més alt, major dispersió, però sense existir un sostre que ens permeta establir interpretacions en termes absoluts.

– La interpretació d'aquests índexs depèn de la naturalesa de la variable considerada i de l'escala de mesura utilitzada –per exemple, un recorregut de 10 en la variable *Pes* (kg) en una mostra de persones adultes sí que ens dona una idea de la dispersió d'aquesta: es tracta d'una variable amb molt poca dispersió atès que caldria esperar que, en una mostra de persones adultes, la diferència entre el valor màxim i el mínim fora bastant major de 10 kg. No obstant això, en altres casos la interpretació podria resultar més incerta, per exemple, un recorregut de 840 mil·lisegons en la variable "Temps de reacció per a reconéixer un determinat estímul visual" indica molta o poca dispersió? Aquest valor pot ser interpretat per algú amb experiència en experiments de temps de reacció amb estímuls visuals, però, en cas de no tenir-ne, pot resultar més que aventurat interpretar-lo.

– Ara bé, sí que serà sempre possible, amb els resultats de qualsevol dels índexs de dispersió, realitzar interpretacions en termes relatius, per exemple, si la mateixa variable l'hem mesurada en dues mostres, podem valorar quina de les dues mostres presenta una major dispersió o, també, podem comparar la dispersió que té una mateixa variable mesurada en dos moments temporals diferents. No s'ha d'oblidar que no tindrà sentit comparar aquests índexs de dispersió quan s'obtinguen per a variables diferents –hi ha només una excepció a aquesta última afirmació: quan es tracte de variables que estiguen expressades en les mateixes unitats i que tinga sentit comparar (per exemple, les variables ingressos i despeses mensuals per a una mostra de consumidors).

2.2. El recorregut interquartílic

• El recorregut (o amplitud o rang) interquartílic (*RIQ*) s'obté com a diferència entre el quartil 3 i el quartil 1:

$$RIQ = Q_3 - Q_1$$

Una variant d'aquest és el conegut com a recorregut (o amplitud) semi-interquartílic:

$$RSIQ = (Q_3 - Q_1)/2$$

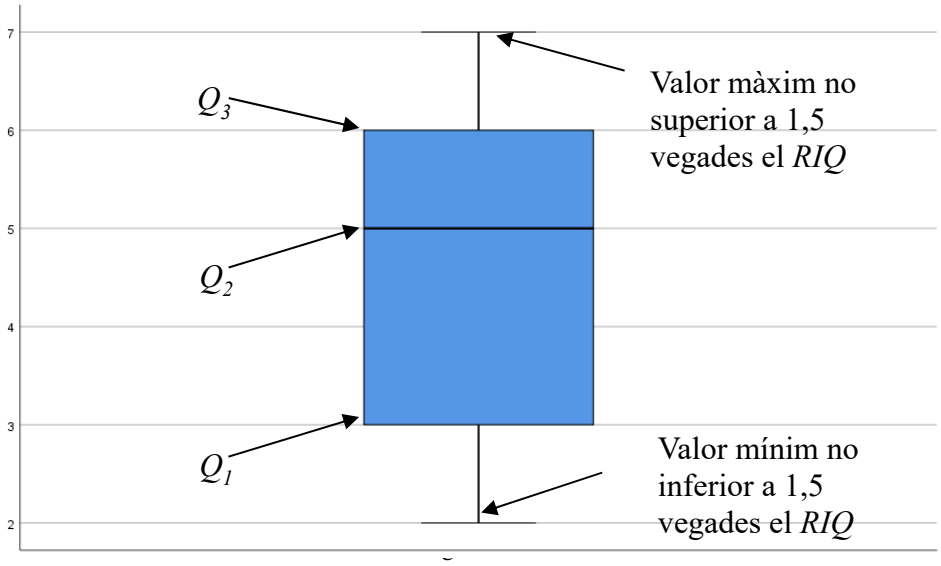
• Tots dos índexs tenen com a avantatge respecte al *Recorregut* que no es veuen afectats per l'existència de valors atípics en la variable, perquè no s'obtenen a partir dels dos valors més extrems de la variable sinó a partir de dos valors més centrats com són el Q_3 i el Q_1.

> **Exemple** d'obtenció del *RIQ* i del *RSIQ* per a la variable "Ansietat que sent quan es troba amb molta gent al voltant" (vegeu-ne la distribució de freqüències més amunt).
>
> $$RIC = 3 - 2 = 1 \qquad RSIC = (3 - 2)/2 = 0{,}5$$

Exercici 4: Obteniu el *RIQ* i el *RSIQ* de la variable "Es valora en els empleats la creativitat i la capacitat d'innovació" (vegeu-ne la distribució de freqüències més amunt).

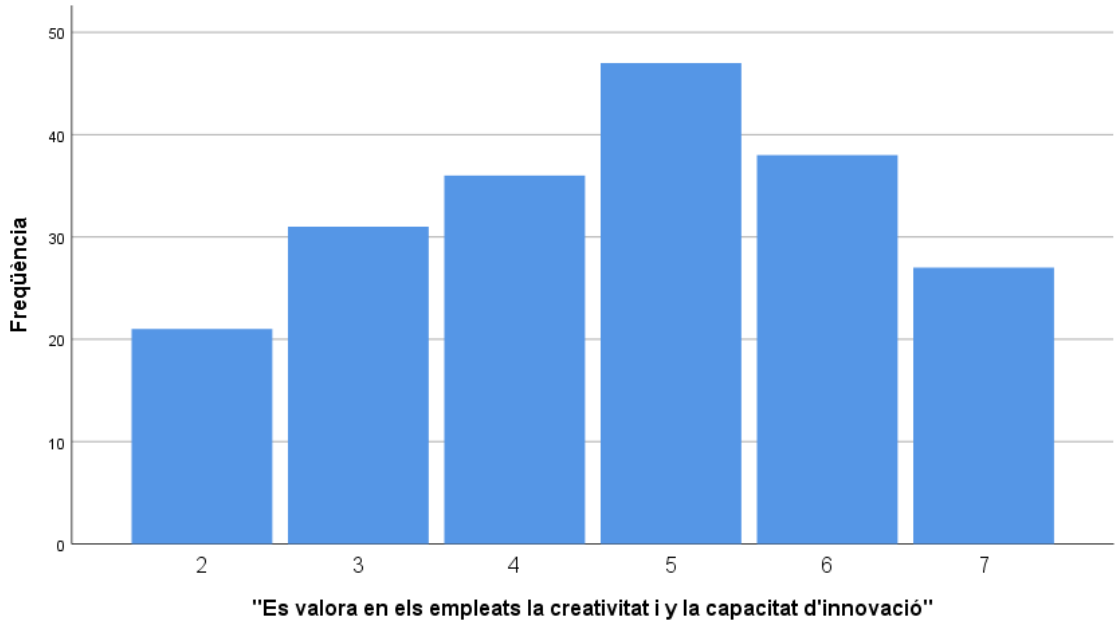
• Un gràfic basat en els Q_3 y Q_1 (i també en la mediana), que es cada vegada més utilitzat, és el conegut com a gràfic de caixa i bigots. Aquest gràfic ofereix informació simultània sobre la posició i variabilitat de la distribució de freqüències d'una variable. Com veurem més endavant, també n'ofereix sobre l'asimetria de la distribució, com també sobre la possible existència de valors atípics en la variable. A més, és un gràfic molt utilitzat amb la finalitat de comparar grups.

> Com a **exemple**, el gràfic de caixa i bigots de la variable "Es valora en els empleats la creativitat i la capacitat d'innovació" obtingut a partir d'una mostra de 200 treballadors/es:



Per fer-ho, a l'eix vertical hem de posar-hi els valors de la variable de manera ordenada i hem de dibuixar una *caixa* delimitada per la mediana i els quartils 1er i 3er (la distància entre tots dos és, precisament, el recorregut interquartílic). A partir de la caixa s'estenen els *bigots* fins als valors més extrems de la variable que es troben dins de 1,5 vegades el *RIQ*. Els valors més enllà de 1,5 vegades el *RIQ*, si existeixen, es consideren valors anòmals, atípics o extrems, a causa de la seua llunyania del gruix de les dades, i es representen per punts o asteriscos.

Es mostra a continuació el gràfic de barres de la mateixa variable a fi que puga comparar-se amb el gràfic de caixa i bigots anterior:



Exercici 5: Feu el gràfic de caixa i bigots de la variable "Ansietat que sent quan es troba amb molta gent al voltant".

3. Variables quantitatives: la variància, la desviació estàndard i el coeficient de variació

3.1. La variància i la desviació estàndard

• La distància dels valor d'una variable respecte a la seua mitjana aritmètica ofereix, de manera intuïtiva, el fonament per a la proposta d'un índex de dispersió, de manera que, com més gran aquestes distàncies, més dispersió hi haurà en les dades.

Aquesta distància d'una dada (X_i) respecte a la mitjana s'anomena en estadística *desviació* o puntuació diferencial (d_i), i és $d_i = X_i - \bar{X}$. Intuïtivament, l'índex de dispersió més senzill consistiria en l'obtenció de la mitjana de les desviacions ($\bar{d}$):

$$\bar{d}_i = \frac{\sum d_i}{n} = \frac{\sum (X_i - \bar{X})}{n}$$

Exemple de càlcul per a la variable *X*: {6, 7, 4, 2, 5, 6}:

$$\frac{\sum (X_i - \bar{X})}{n} = \frac{(6-5)+(7-5)+(4-5)+(2-5)+(5-5)+(6-5)}{6} = \frac{0}{6} = 0$$

El resultat obtingut (0) és poc creïble, en tant que la simple observació de les dades ens diu que la dispersió d'aquesta variable és qualsevol cosa menys nul·la.

• En efecte, la fórmula anterior ens plantejaria una contrarietat important si la utilitzàrem com a índex de dispersió: el seu resultat sempre serà 0, siga com siga el conjunt de dades que considerem. Així, unes altres variants d'aquesta han sigut proposades a fi de superar aquest inconvenient, entre les quals la més rellevant és la variància (S_X^2 *o* σ_X^2):

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

El numerador d'aquesta fórmula (el sumatori de totes les puntuacions diferencials elevades al quadrat) és anomenat en estadística *suma de quadrats* (*SC*), per la qual cosa l'anterior fórmula pot expressar-se així:

$$S_X^2 = SC_X / n$$

Exemple de càlcul per a la variable *X*: {6; 7; 4; 2; 5; 6}:

$$S_X^2 = \frac{\sum (X_i - \bar{X})^2}{n} = \frac{(6-5)^2+(7-5)^2+(4-5)^2+(2-5)^2+(5-5)^2+(6-5)^2}{6} = \frac{16}{6} = 2,67$$

• Si, en la fórmula anterior de la variància, el denominador (*n*) se substitueix per *n*-1, l'índex resultant s'anomena *quasi-variància* i tindrà molta rellevància en l'àmbit de l'estadística inferencial.

• Si la variància es calcula a partir d'una distribució de freqüències:

$$S_X^2 = \frac{\sum n_i \cdot (X_i - \bar{X})^2}{n}$$

Exemple de càlcul de la variància per a la variable "Temps (en segons) emprat a completar un laberint" en una mostra de 20 rates (n = 20):

Temps	n_i	p_i
9	3	0,15
10	8	0,4
11	6	0,3
12	2	0,1
13	1	0,05

$$\bar{X} = \frac{9 \cdot 3 + 10 \cdot 8 + 11 \cdot 6 + 12 \cdot 2 + 13 \cdot 1}{20} = 10,5 seg$$

$$s_x^2 = \frac{3 \cdot (9-10,5)^2 + 8 \cdot (10-10,5)^2 + 6 \cdot (11-10,5)^2 + 2 \cdot (12-10,5)^2 + 1 \cdot (13-10,5)^2}{20} = 1,05 seg^2$$

Una fórmula alternativa en el càlcul de la variància a partir de la informació d'una distribució de freqüències consisteix a sumar el producte de cada desviació al quadrat per la seua freqüència relativa:

$$S_X^2 = \sum p_i \cdot (X_i - \bar{X})^2$$

Exemple per a la variable "Temps emprat a completar un laberint":

$$S_x^2 = 0,15 \cdot (9-10,5)^2 + 0,4 \cdot (10-10,5)^2 + 0,3 \cdot (11-10,5)^2 + 0,1 \cdot (12-10,5)^2 + 0,05 \cdot (13-10,5)^2 = 1,05 seg^2$$

• La interpretació de la variància d'una variable és difícil, atès que aquesta s'expressa com el quadrat de la unitat de mesura de la variable. La desviació típica o estàndard (S_X *o* σ_X) és l'arrel quadrada de la variància i no té aquest inconvenient perquè la unitat en què s'expressa és la mateixa que la de la variable.

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

Exemple de càlcul de la desviació estàndard per a la variable "Temps emprat a completar un laberint":

$$S_X = \sqrt{1,05} = 1,02 seg$$

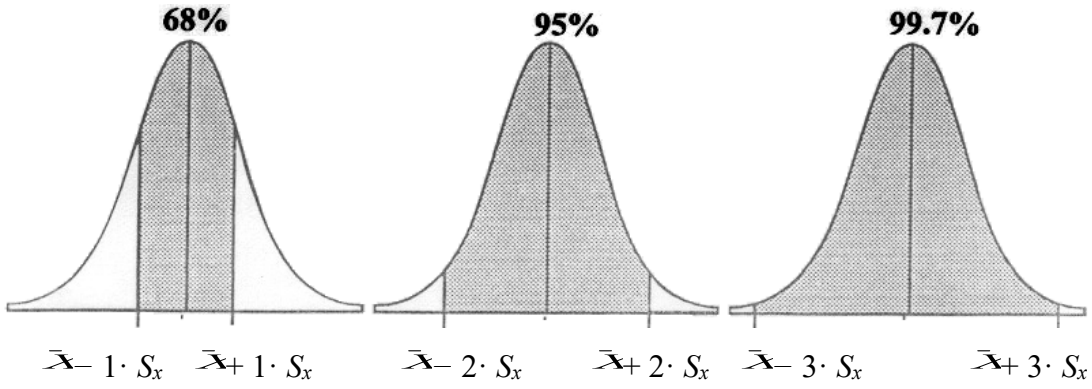
Exercici 6: Obtingueu la S_X^2 i S_X d'una variable quantitativa *X* per a la qual s'han obtingut les següents dades en un grup reduït de 7 subjectes: *X:* {6; 7; 4; 3; 5; 4; 6}

Exercici 7: Inventeu 2 conjunts de 6 dades (valors sencers entre 0 i 10, es poden repetir), cadascun amb $\overline{X} = 7$ però diferent S_x.

Exercici 8: Inventeu 5 dades (valors enters entre 0 i 10, es poden repetir), que tinguen la S_x més petita possible (diferent de 0).

Exercici 9: Inventeu 6 dades (enters entre 0 i 10, es poden repetir), que tinguen la S_x més alta possible.

• Una particularitat de la desviació estàndard és que si tenim una variable amb una distribució de freqüències que s'ajusta a la corba normal (campana de Gauss), llavors es pot deduir el percentatge de casos que es troben entre els valors $\overline{X} \pm k$ vegades la S_x. Per exemple, si $k = 1$ (és a dir, la mitjana ± una vegada el valor de la desviació típica), podem afirmar que el 68 % dels subjectes tindran els seus valors en aquesta variable entre els valors $\overline{X} \pm 1 \cdot S_x$. Gràficament, per a $k = 1$, 2 i 3, en una variable *X* distribuïda normalment:



Exercici 10: Després d'haver recollit dades de l'alçada en un grup de 500 subjectes, s'ha obtingut que la mitjana i la variància són iguals a 170 cm i 8 cm^2, respectivament. Sabent que la distribució de la variable s'ajusta a la corba normal: (1) entre quins valors d'alçada estan el 68 % central dels subjectes?; (2) el 99,7 % central dels subjectes té una alçada entre ____ i ____ ; (3) quants subjectes tenen una alçada entre 161 i 179 cm?

3.2. El coeficient de variació

• La variància o la desviació típica ens permeten comparar la dispersió de diferents distribucions de freqüències obtingudes per a una mateixa variable en diferents grups de subjectes. Per exemple, tenim dos grups de persones (*G1* i *G2)* i les desviacions típiques de la variable *Pes* en tot dos grups ($S_{Pes\_G1} = 4{,}18$ i $S_{Peso\_G2} = 14{,}55$). Aquests valors indiquen que la dispersió de la variable *Pes* és més

gran en el *G2*, tal com es pot veure també a simple vista si observem la taula de més avall seguida de les dades per les dues variables.

• La diferent dispersió també es pot observar en les dades dels dos grups de persones, *G5* i *G6*, en què va ser mesurada la variable *Alçada* ($S_{Alçada\_G5}$ = 0,036 i $S_{Alçada\_G6}$ = 0,227), posant-se de manifest com els valors de la desviació estàndard estan intrínsecament vinculats a l'escala de mesura de la variable considerada. Així, per a la variable *Alçada*, els valors de S_x són més baixos que els obtinguts per a la variable *Pes*, tot i que en el grup *G6* existeix una dispersió considerable en els valors de l'*Alçada*, tal com es pot observar en les dades. Sembla obvi que no resulta coherent comparar la dispersió de variables de diferent naturalesa amb coeficients que s'expressen en les mateixes unitats que les de les variables.

Nom variable	n	Mínim	Màxim	Recorregut	Mitjana	Desv. típ.	CV
Pes_G1	5	70	81	11	75,00	4,18	5,57
Pes_G2	5	59	94	35	75,20	14,55	19,35
Pes_G3	5	4800	5100	300	4960,00	119,37	2,40
Pes_G4	5	4200	6800	2600	5180,00	1028,1	19,85
Alçada_G5	5	1,68	1,77	0,09	1,72	,036	2,12
Alçada_G6	5	1,45	1,98	0,53	1,74	,227	13,04

Pes_G1 (kg): {73; 77; 81; 74; 70}
Pes_G2 (kg): {65; 94; 86; 72; 59}
Pes_G3 (kg): {4800; 4950; 5100; 4900; 5050}
Pes_G4 (kg): {4200; 5500; 6800; 4500; 4900}
Alçada_G5 (m): {1,70; 1,72; 1,77; 1,75; 1,68}
Alçada_G6 (m): {1,45; 1,56; 1,98; 1,91; 1,80}

• Fins i tot la comparació de la variabilitat per a diferents subgrups en una mateixa variable mitjançant la desviació típica no és recomanable en alguns casos: en concret, quan es tracta de subgrups amb mitjanes bastant diferents en aqueixa variable. La raó d'això és que sol haver-hi una associació entre la posició de les dades i la seua dispersió, de manera que com més alta és la mitjana d'una distribució, més alta n'és la dispersió. A tall d'exemple, si mirem en la taula les desviacions típiques per a la variable *Pes* mesurada en dos grups d'elefants *G3* i *G4* (S_{Pes_G3} = 119,4 i S_{Pes_G4} = 1028,1), s'observa que són valors molt elevats, almenys en comparació amb els obtinguts en els dos grups de persones per a la variable *Pes*. No obstant això, si ens fixem en les dades de la variable *Pes_G3* es posa de manifest com, en realitat, es tracta d'un conjunt de dades molt homogeni per al que seria d'esperar en una mostra d'elefants. Així doncs, si comparàrem les desviacions típiques corresponents a les variables *Pes_G3* i *Pes_G2* podríem arribar a conclusions totalment errònies.

• Aquest problema de la comparació de la variabilitat de subgrups amb mitjanes diferents pot resoldre's mitjançant un índex proposat per K. Pearson, el coeficient de variació (CV_X), el qual relativitza el pes de la desviació típica dividint-la per la mitjana (en conseqüència, no té unitats de mesura):

$$CV_X = \frac{S_X}{\bar{X}} \cdot 100$$

• En la pràctica, el *CV* pot ser qualsevol valor superior a 0; ara bé, tal com assenyalen Solanes *et al*. (2005), no sol prendre valors superiors a 100: valors per sobre posarien de manifest una dispersió excepcionalment alta en les dades. En aquest cas, s'aconsella indagar en les fonts de variabilitat de les dades, perquè podria existir algun error o biaix en la recollida de les dades que fora la causa d'una dispersió tan elevada. En aquest sentit, Marró, Ruiz i Sant Martí (2009) assenyalen que valors del *CV* superiors a 50 ja són indicatius de molta dispersió.

Com es pot observar en la taula d'estadístics per al nostre exemple, si es vol comparar la variabilitat de subgrups amb mitjanes diferenciades, s'obtenen conclusions correctes utilitzant el *CV*. A més, en tractar-se d'un coeficient adimensional (sense unitats de mesura), pot resultar també útil per a comparar la dispersió de variables diferents –quan això tinga sentit–, com podria ser el cas de les variables *Alçada* i *Pes* en el nostre exemple.

Exercici 11: Obtingueu totes les mesures de dispersió presentades en aquest tema per a la variable "Nombre de fills" a partir de la distribució de freqüències de les dades:

X_i	n_i
0	40
1	80
2	60
3	20

Exercici 12: Tenim dades sobre la despesa anual en noves tecnologies en els col·legis públics de dues ciutats. En quina de les dues ciutats presenta més dispersió aquesta variable?

Ciutat A	*Ciutat B*
$\bar{X}$ = 24000 €	$\bar{X}$ = 15000 €
Sx = 3300 €	*Sx* = 2900 €

3.3. Algunes anotacions sobre els índexs de dispersió orientats a variables quantitatives

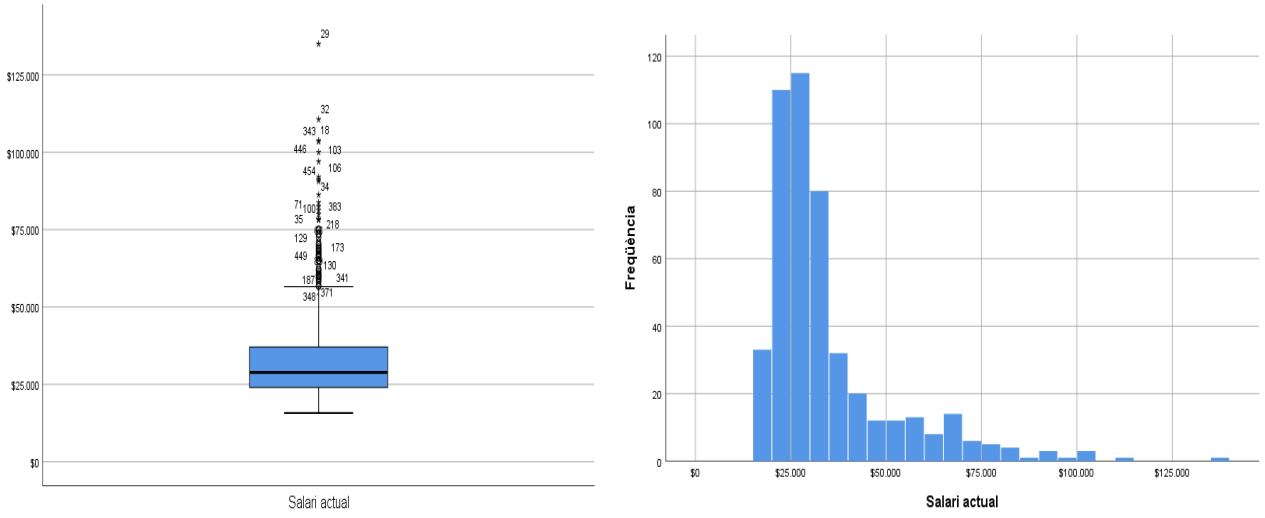
• Els índexs de dispersió basats en la mitjana aritmètica (variància, desviació típica i coeficient de variació) tenen el mateix problema que la mitjana: en concret, la seua sensibilitat a valors anòmals o atípics (valors que s'allunyen en excés del gruix de valors i que apareixen, per exemple, en distribucions de freqüències molt asimètriques). Així, en aquests casos es recomana aplicar el recorregut interquartílic en lloc d'aquests índexs.

• Cal recordar que, com ocorria en el tema precedent i ocorrerà en altres de successius, els índexs presentats per a un determinat tipus de variable, també son aplicables per a variables d'ordre superior –per exemple, els índexs presentats per a les variables categòriques es poden aplicar també a las variables ordinals i a les quantitatives.

3.4. Visualització gràfica de la dispersió amb variables quantitatives

• El gràfic de caixa i bigots resulta també adequat com a representació gràfica de la posició i dispersió de variables quantitatives. A més, aquest gràfic permet saber fàcilment si la distribució d'una variable presenta valor atípics. Aquests valors són els que són superiors a 1,5 vegades el *RIC* més el Q_3 o bé inferiors a 1,5 vegades el *RIC* menys el Q_1. Els valors atípics apareixen en el gràfic representats per punts o asteriscos.

Exemple de gràfic de caixa i bigots amb una distribució de freqüències amb valors atípics (variable "Salari actual" per als 474 empleats d'una empresa de serveis). Es mostra també l'histograma de la mateixa variable a fi que puga comparar-se amb el corresponent gràfic de caixa i bigots.

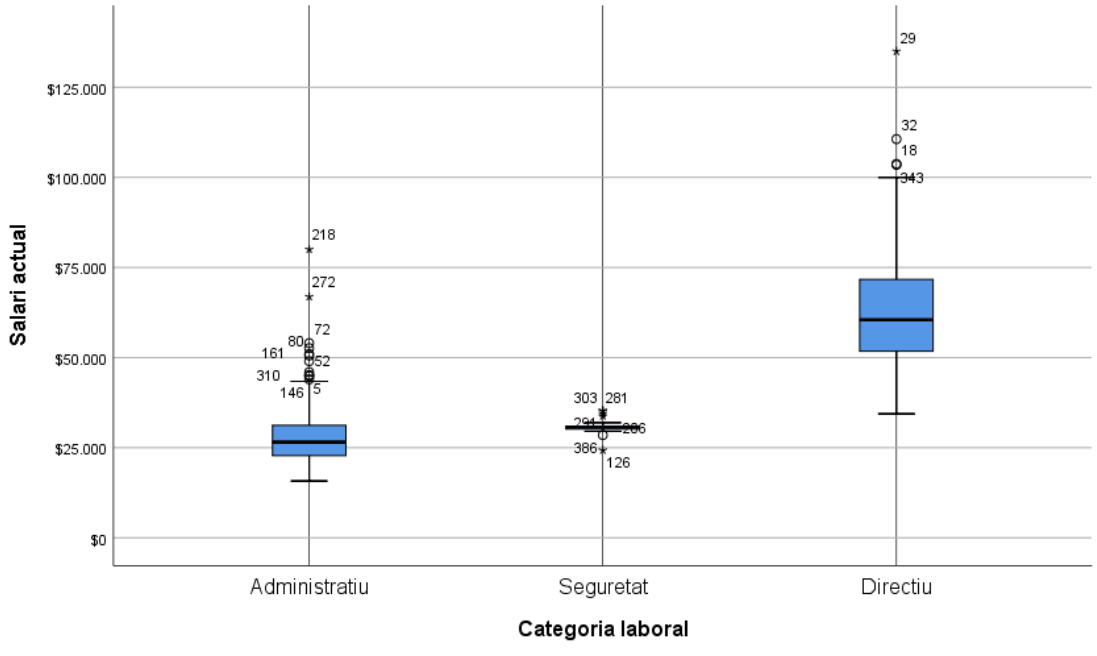


Com que hi ha valors atípics en la distribució d'aquesta variable, no seria adequat descriure la seua dispersió a partir dels índexs de dispersió per a variables quantitatives.

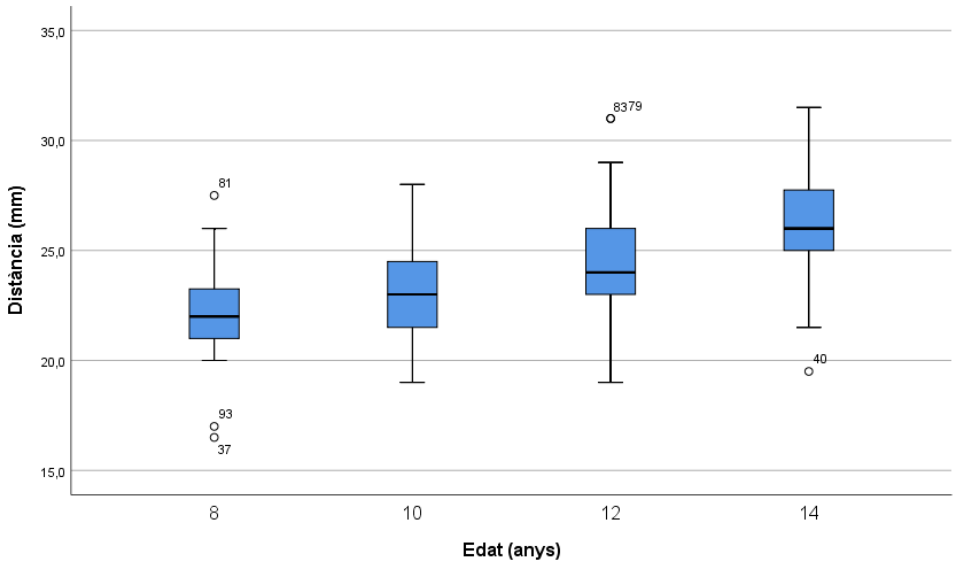
Exercici 13: Tot seguit es mostra la distribució de freqüències de la variable "Nombre de visites al servei d'urgències hospitalari durant l'any passat", mesurada en una mostra de 150 subjectes diagnosticats amb hipocondria. A partir d'aquesta: 1) dibuixeu el gràfic de barres i el gràfic de caixa i bigots (Nota: comproveu abans si hi ha valors atípics!); 2) decidiu quins serien els índexs de dispersió més adequats per aquesta variable i calculeu-los.

	X_i	n_i	%	$\%_a$
	0	11	7,33	7,33
	1	30	20	27,33
	2	41	27,33	54,66
	3	27	18	72,66
	4	19	12,67	85,33
	5	14	9,33	94,66
	6	5	3,33	98
	7	2	1,33	99,33
	10	1	0,67	100
		150	100	

• Una faceta de l'anàlisi estadística en què els gràfics de caixa i bigots són especialment convenients és per a <u>comparar</u> la posició i variabilitat, bé d'una mateixa variable mesurada en diferents subgrups de casos o bé d'una mateixa variable mesurada en diferents moments temporals. A continuació es mostra un **exemple** del primer cas; en concret, es tracta de la variable "Salari actual" per a cadascuna de les tres categories laborals diferenciades en una empresa de serveis:

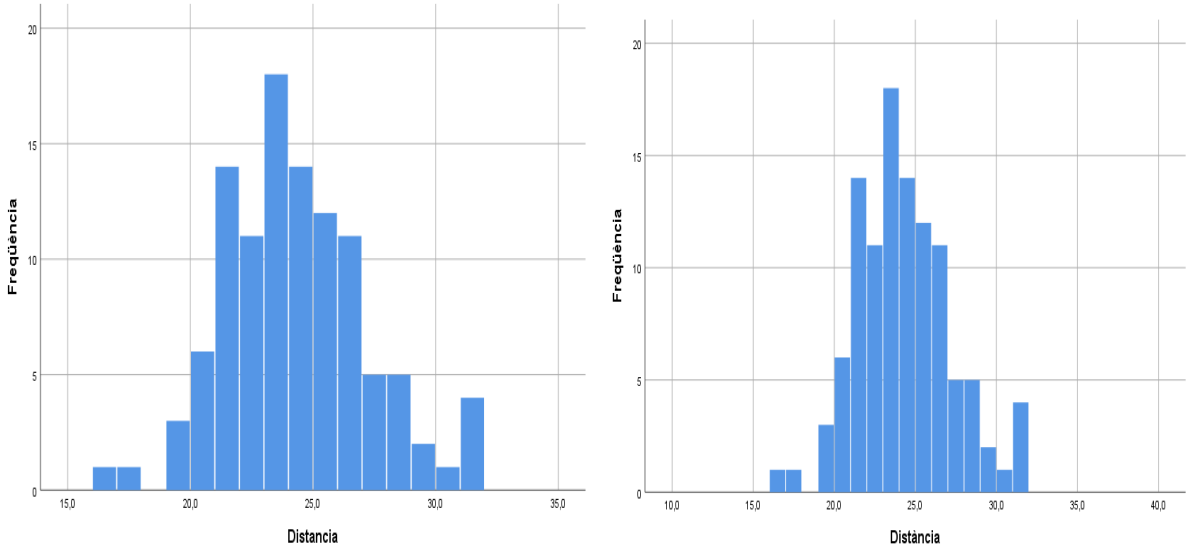


Un altre **exemple** en què es comparen 4 subgrups de subjectes definits en funció de l'edat (8, 10, 12 i 14 anys) en la variable "Distància en mm del centre de la pituïtària a la fissura pterigomaxil·lar":



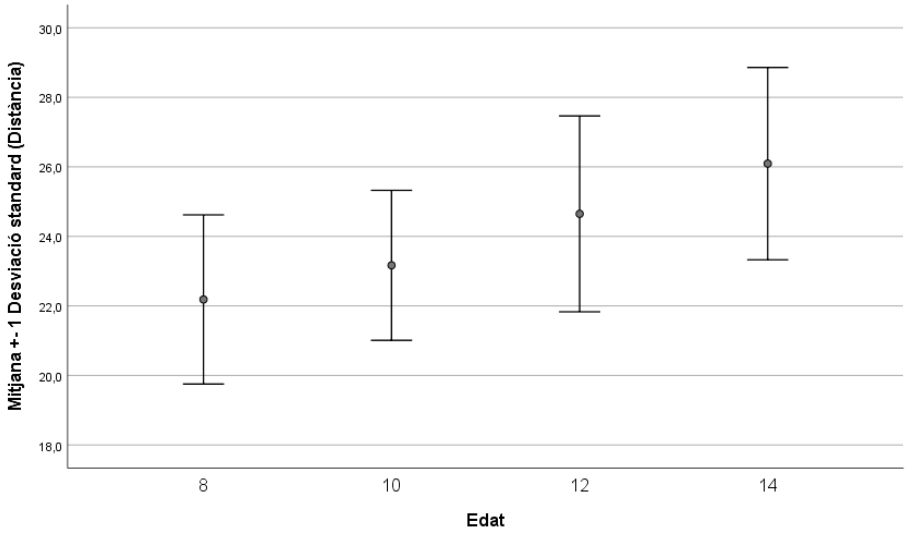
Seria possible representar cadascun dels 4 grups de subjectes mitjançant un histograma i comparar-los; no obstant això, el gràfic de caixa i bigots ofereix més avantatges pel que fa a l'aprofitament de l'espai gràfic. A més, s'evita el problema que cada subgrup puga estar representat en una escala diferent i els errors d'interpretació que puguen derivar-se'n.

Exemple en què es comparen dos subgrups de subjectes en una mateixa variable ("Distància en mm del centre ...") amb dos histogrames:

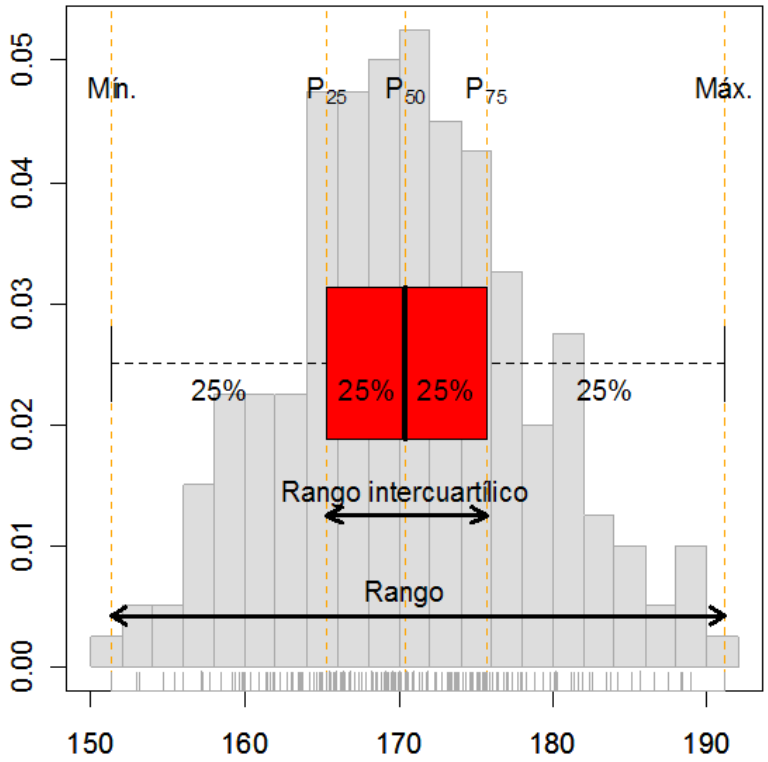


Si ens fixem bé, no es tracta, en realitat, de dos subgrups, sinó que són les mateixes dades en els dos gràfics, però l'escala sobre l'eix horitzontal és diferent. Ara bé, una primera impressió ràpida podria haver-nos conduït a concloure erròniament que els dos subgrups tenen una posició grupal similar, i amb una menor dispersió en el segon. Per tant, en fer ús d'histogrames per comparar grups en una mateixa variable, cal utilitzar la mateixa escala.

• Una variant del gràfic de caixa i bigots és el conegut com a gràfic de barres d'error, en el qual es representa amb un punt la mitjana de la variable i, a partir d'aquest punt, s'estenen dues línies rectes verticals de la mateixa longitud que poden representar diferents elements d'informació estadística, com ara la desviació típica de la variable, l'error estàndard o l'interval de confiança (els dos últims conceptes es tractaran en el mòdul d'estadística inferencial). Com a exemple, tot seguit es mostra un gràfic de barres d'error de la "Distància (mm) del centre de la pituïtària a la fissura pterigomaxil·lar", en què cadascuna de les barres representa la mitjana més/menys una desviació típica per un grup d'edat.



• Tant els gràfics de caixa i bigots com els gràfics de barres els podem trobar representats horitzontalment com es mostra a continuació. A més, en aquest exemple el gràfic apareix superposat sobre un histograma de la mateixa variable. Ambdós representen la distribució de freqüències de la variable "Alçada (cm)" en una mostra d'adults. Cal matisar que el gràfic de caixa i bigots que apareix en aquesta figura és una versió simplificada de la versió original proposada per John W. Tukey que pot trobar-se en alguns manuals d'anàlisi de dades. En aquesta versió original els bigots s'estenen fins als valors mínim i màxim de la distribució, sense tenir en compte la presència de valors atípics.



Exercici 14: A continuació es mostra la distribució de freqüències de la variable "Antiguitat en l'empresa", mesurada a partir del nombre de mesos des del contracte per als 474 empleats/des d'una empresa de serveis. A més, es mostra el gràfic de barres i alguns estadístics descriptius obtinguts amb SPSS. 1) Obtingueu, a partir de la distribució de freqüències, els següents índexs estadístics: el mínim i el màxim, el Q_1, la mediana, el P_{75}, la moda, el recorregut i el *RIQ*. 2) Comproveu que els valors obtinguts coincideixen amb els presentats en la taula de SPSS. 3) Dibuixeu el gràfic de caixa i bigots. 4) Decidiu quins serien els índexs més adequats per a descriure la tendència central i la dispersió d'aquesta variable.

Mesos des del contracte

N	Vàlid	474
	Perduts	0
Mitjana		81,11
Mediana		81,00
Moda		81[a]
Desv. típica		10,061
Variància		101,223
Amplitud		35
Mínim		63
Màxim		98
Percentils	25	72,00
	50	81,00
	75	90,00

a. Existeixen múltiples modes. Es mostra el valor més xicotet

Mesos des del contracte

		Freqüència	Percentatge	Percentatge vàlid	Percentatge acumulat
Vàlid	63	3	,6	,6	,6
	64	8	1,7	1,7	2,3
	65	15	3,2	3,2	5,5
	66	18	3,8	3,8	9,3
	67	15	3,2	3,2	12,4
	68	10	2,1	2,1	14,6
	69	21	4,4	4,4	19,0
	70	13	2,7	2,7	21,7
	71	3	,6	,6	22,4
	72	14	3,0	3,0	25,3
	73	12	2,5	2,5	27,8
	74	8	1,7	1,7	29,5
	75	7	1,5	1,5	31,0
	76	10	2,1	2,1	33,1
	77	13	2,7	2,7	35,9
	78	22	4,6	4,6	40,5
	79	14	3,0	3,0	43,5
	80	15	3,2	3,2	46,6
	81	23	4,9	4,9	51,5
	82	15	3,2	3,2	54,6
	83	19	4,0	4,0	58,6
	84	14	3,0	3,0	61,6
	85	10	2,1	2,1	63,7
	86	13	2,7	2,7	66,5
	87	11	2,3	2,3	68,8
	88	13	2,7	2,7	71,5
	89	7	1,5	1,5	73,0
	90	16	3,4	3,4	76,4
	91	13	2,7	2,7	79,1
	92	15	3,2	3,2	82,3
	93	23	4,9	4,9	87,1
	94	14	3,0	3,0	90,1
	95	6	1,3	1,3	91,4
	96	16	3,4	3,4	94,7
	97	11	2,3	2,3	97,0
	98	14	3,0	3,0	100,0
	Total	474	100,0	100,0	

Referències

De Veaux, R. D., Bock, D. E. i Velleman, P. (2003). *Intro Stats*. Boston: Addison–Wesley.

Pardo, A., Ruiz, M. A. i San Martín, R. (2009). *Análisis de datos en ciencias sociales y de la salud I*. Madrid: Síntesis.

Solanas, A., Salafranca, L., Fauquet, J. i Núñez, M. I. (2005). *Estadística descriptiva en Ciencias del Comportamiento*. Madrid: Thompson.

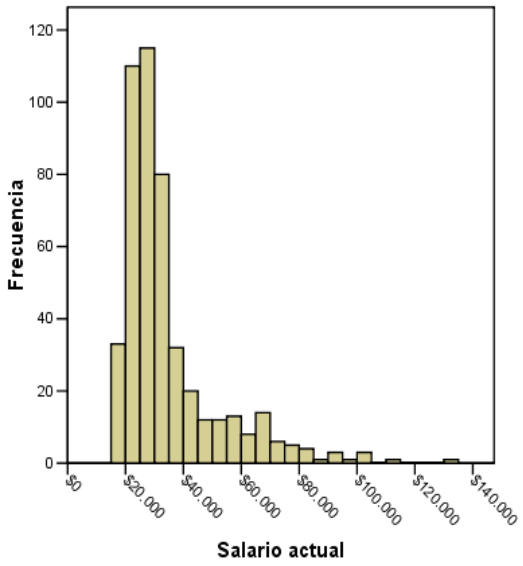
T. 3.3 – Caracterització de grups: Estadístics de forma de la distribució

1. **Simetria**
2. **Apuntament o curtosi**
3. **Descripció estadística d'una variable: taula resum**

• En temes precedents hem tractat l'anàlisi de la forma de la distribució de freqüències des d'una aproximació gràfica. De fet, aquesta és la forma més directa i intuïtiva de fer-se una idea sobre la forma de la distribució d'una variable.

• Tal com es va veure en el seu moment, conéixer la forma d'una distribució era important, per exemple, per tal de decidir quins estadístics de posició i dispersió utilitzar per descriure variables quantitatives. En qualsevol cas, l'examen de la forma de la distribució d'una variable aportarà informació rellevant per si mateixa amb vista a descriure aqueixa variable.

> **Exemple**: Què ens diu la forma de la distribució de la variable "Salari actual" que es mostra en el següent histograma?



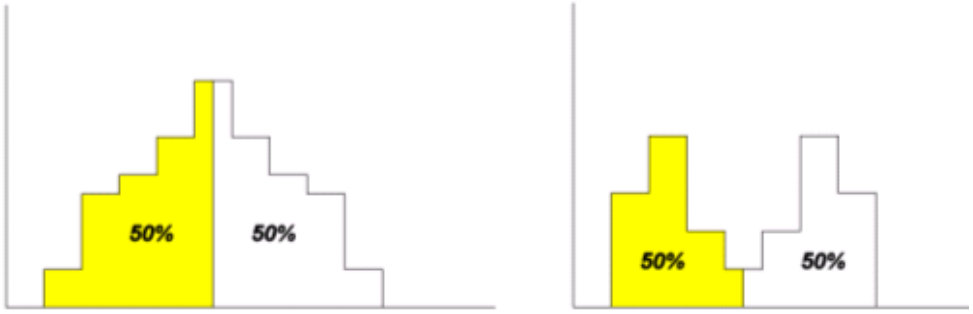
• En aquest tema es presenten diversos índexs que permeten descriure la forma d'una distribució, en concret, dues facetes d'aquesta: la simetria i l'apuntament (o curtosi).

1. Simetria

• La simetria d'una distribució de freqüències fa referència al grau en què valors de la variable, equidistants a un valor que es considere el centre de la distribució, posseeixen freqüències més o menys iguals. Com més similars siguen, més simètrica serà la distribució; com més diferents, més asimètrica.

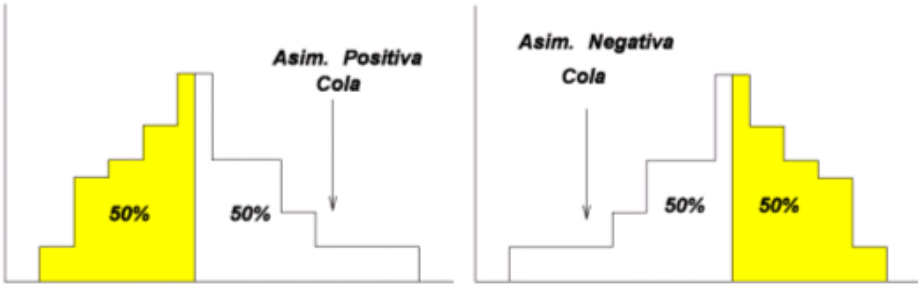
• És un concepte que resulta més intuïtiu de comprendre a nivell visual, en concret, observant una representació gràfica de la distribució de freqüències de la variable (gràfic de barres, histograma…). Aquesta serà simètrica si la meitat esquerra de la distribució és la imatge especular de la meitat dreta.

Exemples de distribució simètrica:



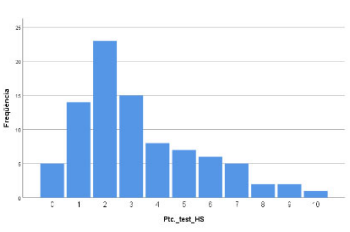
Mitjana aritmètica i mediana coincideixen en les distribucions simètriques. Si només hi ha una moda (distribució unimodal), el valor d'aquesta també serà igual a les dues anteriors.

• En distribucions unimodals, el nivell de simetria se sol descriure d'acord amb tres grans categories: distribucions simètriques, distribucions asimètriques positives (o asimetria a la dreta) i distribucions asimètriques negatives (o asimetria a l'esquerra). Si utilitzem la mode com a eix de referència, aquestes categories de asimetria venen definides pel diferent grau de dispersió de les dades a banda i banda (en les cues) d'aquest eix virtual. La cua més dispersa en el costat dels valors alts de la variable caracteritza a la asimetria positiva; si la major dispersió és en el costat dels valors més baixos, tenim asimetria negativa. Finalment, si la dispersió és igual o molt similar a banda i banda, la distribució de freqüències és simètrica.

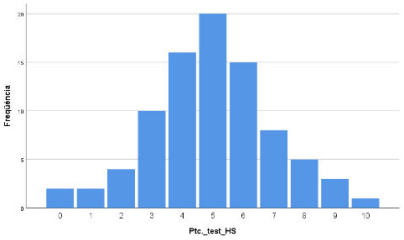


• Si la distribució es asimètrica, els valors de la mitjana, la mediana i la moda difereixen. En concret, si la asimetria és positiva: $\overline{X} > Mdn \geq Mo$; mentre que si és negativa: $\overline{X} < Mdn \leq Mo$.

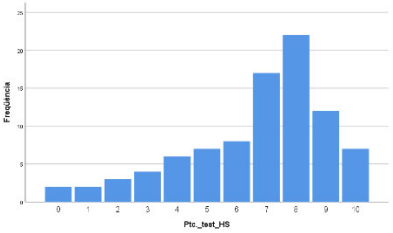
Exemple de les puntuacions d'un grup de subjectes en un test d'habilitats socials abans, durant i després de rebre 6 sessions d'entrenament en habilitats socials.



Abans ($\overline{X}$ =3,26; *Mdn*=3; *Mo*=2).



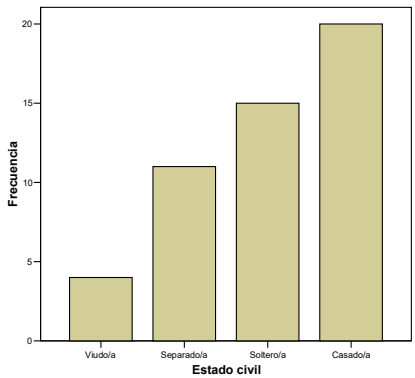
Durant ($\overline{X}$ =4,97; *Mdn*=5; *Mo*=5)

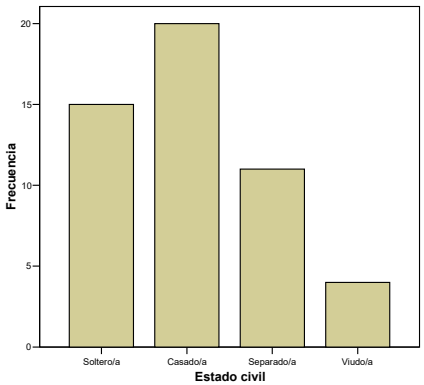


Després ($\overline{X}$ =6,67; *Mdn*=7; *Mo*=8)

• A continuació es presenten diferents índexs estadístics que permeten quantificar el grau de simetria d'una variable. Convé matisar que per a variables categòriques no té sentit obtenir aquests índexs, atès que no existeix un ordre intrínsec en els valors de la variable.

Veure, per **exemple**, dos gràfics de barres de la mateixa distribució de freqüències de la variable "Estat civil" que difereixen únicament en la posició de les modalitats de resposta:



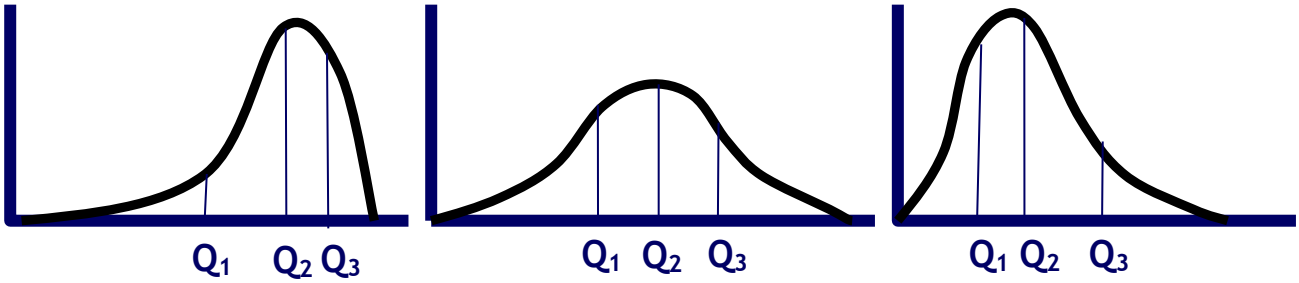


1.1. Variables ordinals: l'índex d'asimetria interquartílic

• L'índex d'asimetria interquartílic s'obté a partir dels quartils de la distribució com:

$$As_{Q_3-Q_1} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

• Interpretació: oscil·la entre -1 i 1, la qual cosa facilita la seua comunicació i comprensió.



$Q_3 - Q_2 < Q_2 - Q_1$	$Q_3 - Q_2 = Q_2 - Q_1$	$Q_3 - Q_2 > Q_2 - Q_1$
$As_{Q3-Q1} < 0 \rightarrow$Asimetria –	$As_{Q3-Q1} = 0 \rightarrow$ Simetria	$As_{Q3-Q1} > 0 \rightarrow$Asimetria +

Exercici 1: Calculeu l'índex As $_{Q3-Q1}$ per a les distribucions de freqüències de 3 grups de 100 casos cadascun (A, B i C). Els subjectes van emplenar un test que constava de 10 ítems i cadascun era valorat amb 1 punt si la resposta era correcta, i amb un 0 si n'era incorrecta. La puntuació en el test per a cada subjecte s'obtenia com a suma de les puntuacions dels ítems, per tant, podia oscil·lar entre 0 i 10 (Nota: encara que la variable podria considerar-se com a quantitativa, assumiu en aquest exercici que és ordinal). Obteniu també els gràfics de caixa i bigots per les 3 distribucions.

	Grup A	**Grup B**	**Grup C**
Puntuació	n_i	n_i	n_i
0	1	4	5
1	3	5	11
2	3	8	14
3	5	9	23
4	8	15	15
5	11	18	12
6	15	15	9
7	24	9	6
8	16	8	2
9	9	5	2
10	5	4	1
	100	**100**	**100**

1.2. Variables quantitatives: el coeficient d'asimetria de Fisher.

• El coeficient d'asimetria de Fisher es basa en les desviacions dels valors observats respecte a la mitjana. La fórmula per al seu càlcul és la següent:

$$As_F = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n \cdot S_X^3} \qquad \text{(versió per a distribució de freqüències: } As_F = \frac{\sum n_i(X_i - \bar{X})^3}{n \cdot S_X^3}\text{)}$$

• Interpretació: els valors menors que 0 indiquen asimetria negativa; els majors, asimetria positiva; i quan siga 0, o molt pròxim a 0, simetria. No està limitat a un rang de valors.

• El coeficient d'asimetria de Fisher és precisament l'únic que proporciona SPSS. El programa proporciona també, per defecte, el valor de l'error típic d'asimetria, un concepte que es tractarà més endavant (veure taula a continuació). No obstant això, convé assenyalar ja ara que tots dos valors ens permetran tenir un criteri més objectiu en la valoració de la asimetria d'una variable, de manera que, si es divideix el valor del coeficient d'asimetria de Fisher entre l'error típic d'asimetria, s'obtindrà un valor que si és inferior a –2, indica asimetria negativa; si oscil·la entre –2 i 2, indica simetria, i si és superior a 2 indica asimetria positiva.

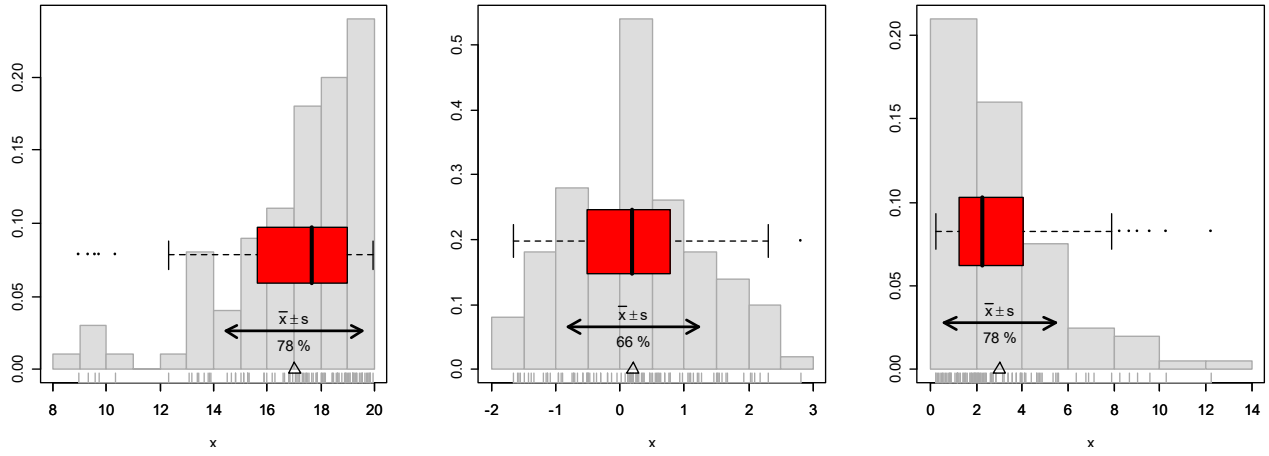
Estadístics

Nota mitjana d'accés

N	Vàlid	169
	Perduts	5
Mitjana		6,3885
Mediana		6,3000
Moda		6,60
Desv. Típica		,55429
Asimetria		1,139
Error estàndard d'asimetría		,187
Curtosi		2,733
Error estàndard de curtosi		,371
Recorregut		3,99
Mínim		5,06
Màxim		9,05

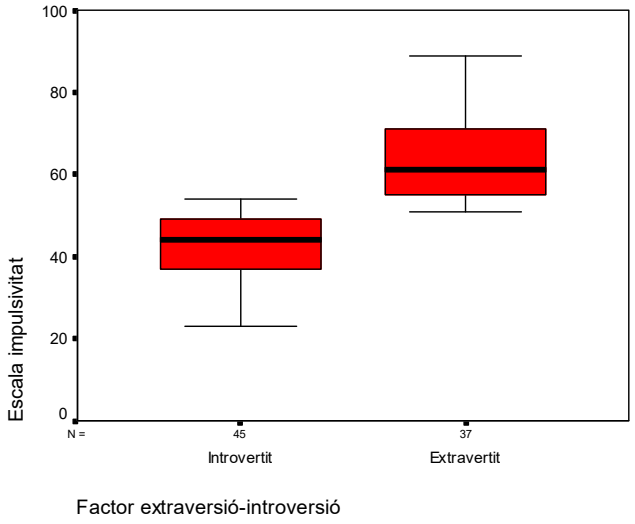
Exercici 2: Valoreu la simetria de la variable "Nota mitjana d'accés" a partir de la taula de SPSS anterior.

• L'histograma representa la millor opció per a la visualització de la simetria d'una variable quantitativa, si bé, el gràfic de caixa i bigots també constitueix una opció vàlida per a fer-ho. A continuació es presenta un exemple en què es mostren 3 distribucions amb diferent nivell de simetria amb tots dos tipus de gràfics superposats (Barón-López, 2005). Cal assenyalar que el gràfic de caixa i bigots apareix en horitzontal amb finalitats didàctics, la qual cosa no és habitual en la presentació d'aquest tipus de gràfic.



• Tal com ja es va destacar en el capítol previ, un avantatge important dels gràfics de caixa i bigots és la facilitat per a presentar-los conjuntament i, d'aquesta manera, poder realitzar comparacions entre diferents distribucions.

Exemple amb les puntuacions en un test d'impulsivitat en un grup de subjectes introvertits i un altre d'extravertits. En el grup d'introvertits la variable és lleugerament asimètrica negativa, mentre que és asimètrica positiva en el grup d'extravertits:



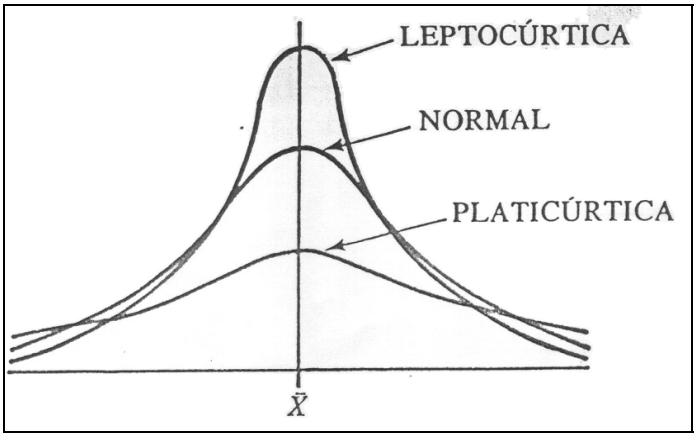
2. Apuntament (curtosi)

• L'apuntament o curtosi d'una distribució de freqüències no té un referent natural com en el cas de la simetria, sinó que se sustenta en la comparació respecte a una distribució de referència, en concret, la distribució normal o campana de Gauss. En conseqüència, la seua obtenció només tindrà sentit en variables la distribució de freqüències de les quals siga similar a la corba normal –en la pràctica això es redueix, bàsicament, al fet que siga unimodal i més o menys simètrica.

• L'apuntament expressa el grau en què una distribució acumula casos en les seues cues en comparació amb els casos acumulats en les cues d'una distribució normal, ambdues amb la mateixa dispersió (Marró i Ruiz, 2002). Així, de manera anàloga a la asimetria, es diferencien 3 grans categories d'apuntament:

- Distribució platicúrtica (apuntament negatiu): indica que en les cues hi ha més casos acumulats que en les cues d'una distribució normal.
- Distribució leptocúrtica (apuntament positiu): el contrari al cas anterior.
- Distribució mesocúrtica (apuntament normal): com en la distribució normal.

Exemples gràfics d'aquestes tres formes d'apuntament en la distribució:

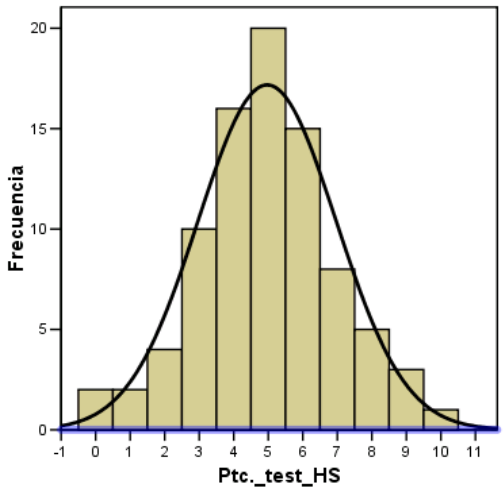


• El coeficient d'apuntament de Fisher permet descriure l'apuntament o curtosi d'una distribució de dades i es basa en les desviacions dels valors observats respecte a la mitjana. La fórmula per al seu càlcul és la següent:

$$Ap_F = \frac{\sum (X_i - \bar{X})^4}{N \cdot S_X^4} - 3$$ (versió per a distribució de freqüències: $Ap_F = \frac{\sum n_i (X_i - \bar{X})^4}{N \cdot S_X^4} - 3$)

• Interpretació: el valor d'aquest coeficient per a la distribució normal serà igual a 0, per tant, qualsevol distribució per a la qual s'obtinga un valor de Ap $_F$ igual o pròxim a 0 significarà que el seu nivell d'apuntament és com el de la distribució normal (mesocúrtica). Valors majors que 0 indiquen que la distribució és leptocúrtica, mentre que valors menors que 0 indiquen que la distribució és platicúrtica. No està limitat a un rang de valors.

Exercici 3: Valoreu la curtosi a partir de l'histograma de la distribució de la variable amb les puntuacions en el test d'habilitats socials. En el mateix apareix superposada la corba normal.



• El coeficient d'apuntament de Fisher és el que proporciona SPSS. Obtindrem també amb aquest programa, per defecte (veure figura a continuació), el valor de l'error típic associat a aquest estadístic. Ambdós valors ens permetran obtenir un criteri més objectiu en la valoració de la curtosi de la distribució d'una variable, de manera que, si es divideix el valor del coeficient d'apuntament de Fisher entre el corresponent error típic s'obtindrà un valor que si és inferior a –2, es considera com a indicatiu de distribució platicúrtica; si oscil·la entre –2 i 2 indica que la distribució és mesocúrtica, i si és superior a 2 indica que la distribució és leptocúrtica.

Estadístics

Nota mitjana d'accés

N	Vàlid	169
	Perduts	5
Mitjana		6,3885
Mediana		6,3000
Moda		6,60
Desv. Típica		,55429
Asimetria		1,139
Error estàndard d'asimetría		,187
Curtosi		2,733
Error estàndard de curtosi		,371
Recorregut		3,99
Mínim		5,06
Màxim		9,05

Exercici 4: Valoreu l'apuntament de la variable "Nota mitjana d'accés" a partir de la taula de SPSS anterior.

3. Descripció estadística d'una variable: taula resum

• En aquest tema i els precedents s'han presentat una sèrie de procediments estadístics, tant numèrics com gràfics, adequats per a descriure i/o resumir els valors obtinguts en mesurar una variable en un conjunt de casos. En la taula següent es resumeix aquesta informació classificada en funció de l'escala de mesura de la variable que es desitja descriure.

	Categòrica	Ordinal	Quantitativa simètrica	Quantitativa asimètrica
Gràfics	Gràfic de sectors Gràfic de barres Pictograma	Polígon de freqüències Gràfic de caixa i bigots Histograma (només per a variables contínues)		
Tendència central	Moda	Mediana	Mitjana aritmètica	Mediana Mitjana retallada
Variabilitat	Índex de variació qualitativa	Recorregut (Amplitud) Recorregut interquartílic	Variància Desviació estàndard Coeficient de variació	Recorregut interquartílic
Simetria		Índex d'asimetria interquartílic	Coeficient d'asimetria de Fisher	
Curtosi			Coeficient d'apuntament de Fisher	

Referencias:

Barón-López, J. (2005). *Bioestadística: métodos y aplicaciones.* Apuntes y material disponible en http://www.bioestadistica.uma.es/baron/apuntes/

Pardo, A. y Ruiz, M. A. (2002). *SPSS: Guía para el análisis de datos.* Madrid: McGraw-Hill.

Tema 4 – Estadístics de posició individual

1. **Els percentatges acumulats ("percentils")**
2. **Les puntuacions diferencials**
3. **Les puntuacions típiques**
   3.1. Les escales derivades

• Fins ara s'ha abordat la descripció de les dades d'una variable; en aquest capítol, en canvi, l'objectiu principal és la descripció de casos particulars, en concret, veurem estadístics que ens oferiran informació sobre la posició que ocupa un valor concret dins d'un conjunt de valors observats (variable).

• En tant que es tracta d'estadístics que ofereixen informació sobre la posició d'un valor respecte a un grup de referència, ens permetran establir una interpretació relativa dels valors observats.

Exemple: Ens diu un amic que els han passat a tots els treballadors de la seua empresa un test d'aptituds verbals i que ell ha obtingut una puntuació igual a 134. A continuació, sense més detalls, ens pregunta si aquesta puntuació significa que és bo o dolent en aptituds verbals.

- Quina informació addicional podria ser-nos d'utilitat a fi de poder oferir-li algun tipus de valoració d'aquesta puntuació?
- Com podríem transformar aquesta puntuació per tal que fora més informativa?

En els apartats successius s'ofereixen algunes respostes a aquesta segona qüestió.

1. Els percentatges acumulats ("percentils")

• El percentatge acumulat ($\%_a$) d'un valor concret d'una variable és el percentatge de casos que obtenen un valor inferior o igual a aquest en la variable en qüestió, informació que pot obtenir-se a partir de la distribució de freqüències de la variable.

• Aquests percentatges acumulats són anomenats més habitualment, encara que d'una manera equívoca, percentils (terme ja utilitzat en el context dels estadístics de posició grupal). Així, és comú escoltar expressions com ara "Crec que serà molt alt, ara està en el percentil 90 dels de la seua edat" o "Ha obtingut un mal resultat en la prova de coordinació, està en el percentil 5". Una altra expressió que també s'utilitza en la literatura per a fer referència al percentatge acumulat d'una determinada puntuació és la de rang centil, si bé, el seu ús no està molt estés.

Exemple d'obtenció de percentatges acumulats (percentils): tenim la següent distribució de freqüències de les puntuacions en un test d'intel·ligència *(CI)* que va ser administrat a una mostra de 250 persones.

- Quin és el percentatge acumulat (rang centil o "percentil") corresponent a una puntuació de 97 en aquest test?, com l'interpretem?
- I si la puntuació fora igual a 103?
- I si fora igual a 91? => interpolació del $\%_a$

CI	n_i	n_a	p_i	p_a	$\%_a$
89	1	1	0,004	0,004	0,4
90	2	3	0,008	0,012	1,2
92	3	6	0,012	0,024	2,4
93	5	11	0,02	0,044	4,4
94	8	19	0,032	0,076	7,6
95	10	29	0,04	0,116	11,6
96	14	43	0,056	0,172	17,2
97	17	60	0,068	0,24	24
98	24	84	0,096	0,336	33,6
99	29	113	0,116	0,452	45,2
100	36	149	0,144	0,596	59,6
101	33	182	0,132	0,728	72,8
102	26	208	0,104	0,832	83,2
103	19	227	0,076	0,908	90,8
104	12	239	0,048	0,956	95,6
105	7	246	0,028	0,984	98,4
107	2	248	0,008	0,992	99,2
110	1	249	0,004	0,996	99,6
114	1	250	0,004	1	100
	250		1		

• Els $\%_a$ o "percentils" són molt utilitzats en la interpretació de les puntuacions dels tests. Així, una taula amb la correspondència entre els possibles valors observats en un test (puntuacions del test) i els corresponents percentatges acumulats, constitueix el que s'anomena com el barem d'aqueix test. Un barem sol ser elaborat a partir d'una mostra representativa de la població a la qual pertanyen els subjectes als quals es té intenció d'aplicar aqueix test.

2. Les puntuacions diferencials

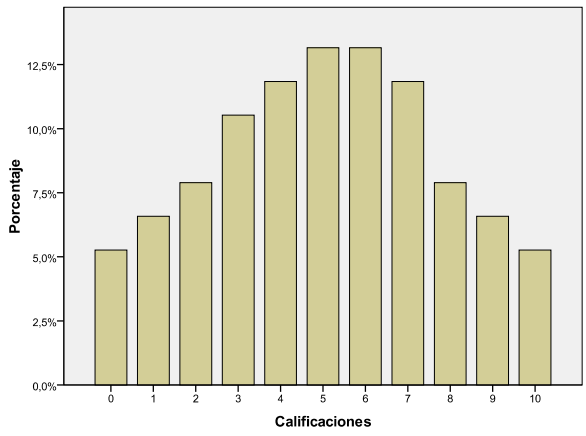
• Per començar, val la pena recordar que en el context dels estadístics de posició individual és bastant habitual que s'utilitze el terme puntuacions, o també puntuacions directes, per a fer referència als valors observats d'una variable.

• Una aproximació intuïtiva a la interpretació d'una puntuació d'un subjecte en una variable consisteix a obtenir la corresponent puntuació diferencial, això és, la diferència entre la puntuació directa i la mitjana d'aqueixa variable: $x_i = X_i - \overline{X}$. Si la puntuació diferencial obtinguda és major [menor] que 0 és que es tracta d'una puntuació que està per damunt [davall] de la mitjana, més allunyada de la mitjana com més gran siga aqueixa diferència en valor absolut.

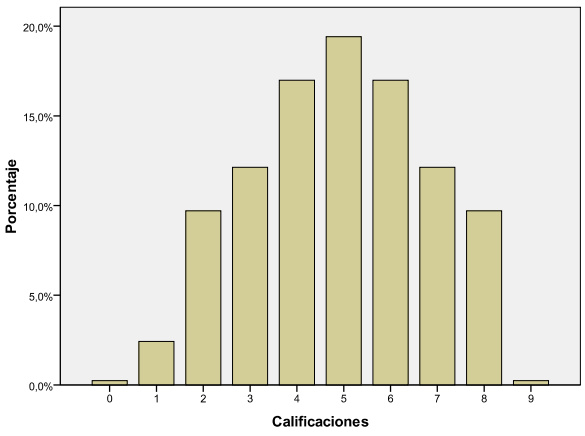
• Un inconvenient es pot plantejar quan es comparen puntuacions diferencials procedents de grups diferents, perquè aquestes no tenen en compte la possible diferent dispersió dels grups.

> **Exemple**: Carmen s'ha presentat a l'examen de les oposicions per a professora d'ensenyament de secundària en dues comunitats autònomes diferents (Andalusia i Comunitat Valenciana) i ha obtingut una puntuació de 8 en tots dos exàmens. Si ens diuen que la mitjana de les puntuacions en l'examen en ambdues comunitats ha sigut de 5, què podem afirmar respecte al rendiment de Carmen en les dues comunitats? Una resposta ràpida a l'anterior pregunta pot donar lloc a una interpretació errònia del resultat de Carmen en tots dos exàmens. Així, en observar tot seguit les distribucions de freqüències de les puntuacions dels exàmens en Andalusia i la CV, què podem dir respecte a aquesta interpretació?, en quina comunitat es posiciona millor Carmen?, en quina comunitat va obtenir, per tant, un millor resultat?

Andalusia ($\overline{X} = 5; S_x = 3$)



CV ($\overline{X} = 5; S_x = 2$)



3. Les puntuacions típiques

• Una alternativa al problema plantejat en la utilització de les puntuacions diferencials és l'ús de les puntuacions típiques (estàndard o z), una transformació de les puntuacions directes que té en compte tant la tendència central (mitjana) com la dispersió (desviació típica) de la distribució de freqüències de la variable. La fórmula per a obtenir-les és la següent:

$$z_i = \frac{X_i - \bar{X}}{S_X}$$

Exemple: quines seran les puntuacions típiques corresponents a les puntuacions directes obtingudes per Carmen en totes dues comunitats? ($z_{Andalusia}$ = ?; z_{CV} = ?)

$$z_{Andalucía} = \frac{8-5}{3} = 1 \qquad z_{CV} = \frac{8-5}{2} = 1{,}5$$

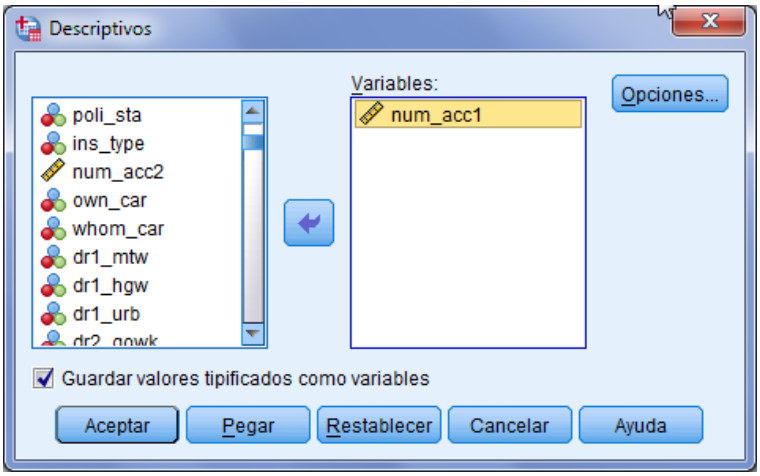
• La puntuació *z* corresponent a un determinat valor expressa el nombre de desviacions típiques que aqueix valor dista de la mitjana del conjunt de les observacions. Així, si Carmen té una puntuació típica igual a 1 a Andalusia, això significa que la seua puntuació directa està 1 desviació típica per damunt de la mitjana d'aqueix grup, així és que, la seua puntuació directa és (encara que ja la sabíem) igual a 1·3 + 5 = 8. Per altra banda, si Carmen té una puntuació típica igual a 1,5 en la CV, això significa que la seua puntuació directa està 1,5 desviacions típiques per damunt de la mitjana d'aquest grup, així és que la seua puntuació directa és igual a 1,5·2 + 5 = 8 (tal com ja sabíem). En conclusió, encara que la puntuació directa és la mateixa en totes dues comunitats, Carmen obté una millor puntuació (en termes relatius), és a dir, està més ben posicionada en la CV que en Andalusia.

Exercici 1: Quina seria la puntuació típica (*z*) d'un opositor que es va presentar en la CV i va obtenir una puntuació directa de 2?; i la d'un altre opositor amb una puntuació de 5? Interpreteu aquestes puntuacions típiques. Quina seria la puntuació (directa) d'un opositor d'Andalusia que té una puntuació *z* igual a 1,5?, i la d'un altre amb una *z* = –1?, i la d'un tercer amb una *z* = 0?

Exercici 2: Completeu la taula amb les puntuacions directes, diferencials i típiques de 4 casos, en una variable X de la qual tenim dades en una mostra de 1250 subjectes ($\bar{X} = 18$; $S_X = 4$).

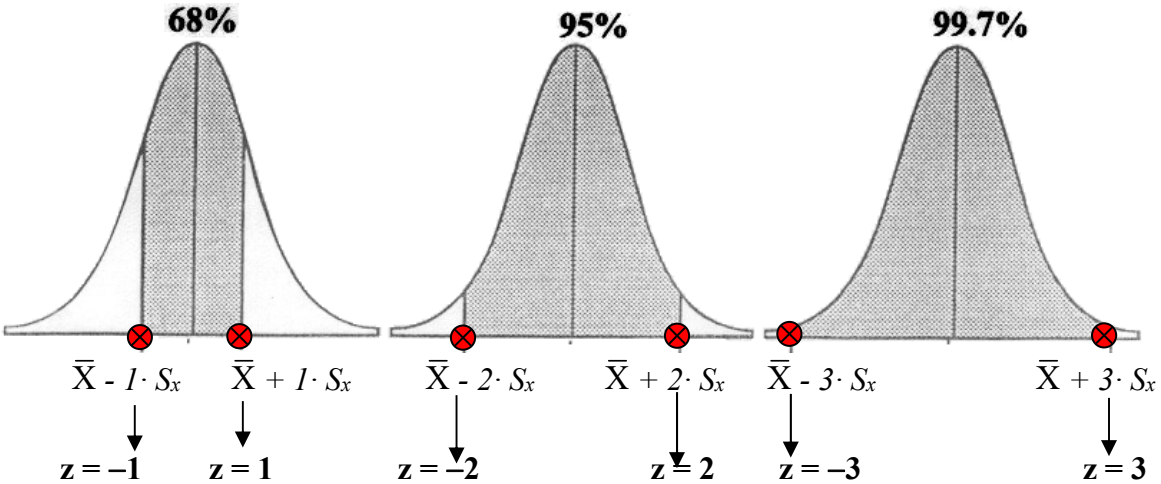
Cas	X_i	x_i	z_i
1	20		
2		–3	
3			3
4			–2

• Amb SPSS és possible obtenir la puntuació típica corresponent a cadascun dels casos d'una determinada variable del nostre arxiu de dades. Això es fa mitjançant el quadre de diàleg que apareix en seleccionar el comandament 'Descriptius' (menú Analitzar > Estadístics descriptius). En aquest quadre cal seleccionar l'opció 'Guardar valors tipificats com a variables'. En executar aquesta funció, es crearà una nova variable en l'arxiu de dades que contindrà les puntuacions *z* corresponents a la variable seleccionada en el citat quadre de diàleg.



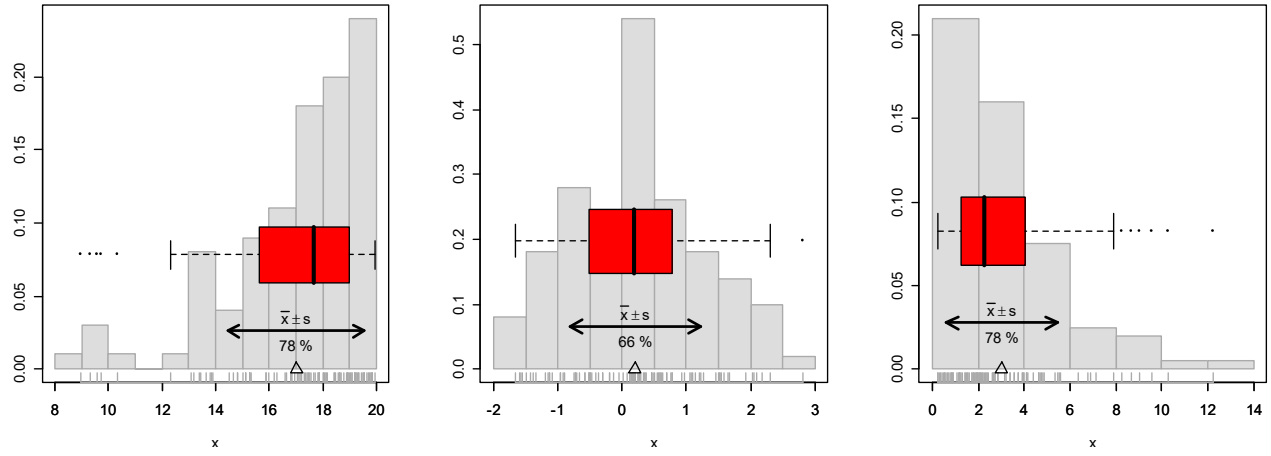
• Algunes característiques de les puntuacions típiques:

(1) Tal com ja es va veure en el tema sobre la dispersió, si una distribució de freqüències s'ajusta a la corba normal, es compleix que entre la mitjana ± una desviació típica (és a dir, entre z = –1 i z = 1) es trobaran el 68 % dels casos. Si es considera la mitjana ± 2 desviacions típiques (entre z= –2 i z = 2), el 95 %; i si la mitjana ± 3 desviacions típiques (entre z = –3 i z = 3), el 99,7 %. Cal destacar també que aquests són 3 casos particulars en què els valors de *z* són valors sencers, però és possible conèixer quin és el percentatge de casos que es trobaran entre qualsevol parell de valors *z* (o per damunt o per davall d'un determinat valor *z*) a partir de la consulta de la taula de la distribució normal tipificada que serà presentada en un tema posterior.



En conseqüència, a nivell interpretatiu, puntuacions típiques majors d'1 o menors de –1, poden considerar-se ja com bastant altes i baixes, respectivament (només un 16 % de casos tindrà una puntuació *z* major de 1 i només un 16% menor de –1); si són majors/menors de 2/–2, es tractaria de valors molt alts/baixos (només un 2,5% de casos tindrà una puntuació z major de 2 i només un 2,5% menor de –2); i si són majors/menors de 3/–3, ja mereixerien el qualificatiu d'extrems (només un 0,15 % de casos tindrà una puntuació z major de 3 i només un 0,15% menor de –3). Aquesta és una característica rellevant de les puntuacions típiques, perquè permet oferir una interpretació d'una puntuació (valor observat) en relació al seu grup de referència.

Finalment, cal subratllar que en el cas de distribucions que no s'ajusten a la corba normal, els percentatges són diferents als presentats i, a més, desconeguts a priori. N'és un bon exemple la següent figura en què es presenta el percentatge de subjectes entre els valors *z* -1 i 1 per a 3 variables amb distribucions de freqüències marcadament diferents. Es pot observar que, en el cas de les dues distribucions més asimètriques, aquest percentatge (78 %) és superior al 68 %:



(2) Si transformem els valors d'una variable qualsevol en puntuacions típiques, els nous valors tindran sempre una mitjana igual a 0 i una desviació típica igual a 1.

$$\bar{z}_X = 0$$
$$S^2_{z_X} = S_{z_X} = 1$$

En conseqüència, transformar 2 (o més) variables a l'escala de les puntuacions típiques suposa fer-les directament comparables entre si, perquè es trobaran en una mateixa escala amb mitjana igual a 0 i desviació típica igual a 1, la qual cosa permetrà comparar les puntuacions d'un mateix subjecte en variables diferents.

Exercici 3: Tenim una variable X:{3, 6, 5, 2}. Transformeu els valors observats (puntuacions directes) en puntuacions típiques (és a dir, tipifiqueu la variable). Calculeu després la mitjana i la desviació típica de les puntuacions típiques obtingudes.

Exercici 4: Una psicòloga especialitzada en psicologia clínica infantil va observar un xiquet de 5 anys mentre jugava amb altres xiquets en una situació estàndard d'observació. Així, la psicòloga va registrar que, durant el temps d'observació, el xiquet havia interactuat verbalment 6 vegades amb els altres xiquets. Com s'interpretaria aquest valor tenint en compte que en estudis previs similars la mitjana d'interaccions verbals dels xiquets és 12 i la variància 9? Si assumim que aquesta variable es distribueix normalment, quin percentatge de xiquets tindrà una puntuació igual o inferior a la d'aquest xiquet (6)?

Exercici 5: Mariona i Lucia, després d'acabar els seus estudis de grau en Psicologia i Economia, respectivament, reben ofertes de treball amb la següent remuneració econòmica neta: Mariona, 11.420 €; Lucía, 12.320 €. Sabem per estudis estadístics a nivell nacional, que els salaris per a primera ocupació en tots dos graus tenen en l'actualitat les següents característiques:

Psicologia (Mariona)	*Economia* (Lucía)
$\bar{X} = 10217\,€$	$\bar{X} = 10818\,€$
$Sx = 510\,€$	$Sx = 901\,€$

A partir de les dades anteriors, quina jove es pot dir que té una oferta millor en relació als salaris dels respectius graus?

3.1 Les escales derivades

• Una dificultat amb les puntuacions típiques és comunicar els resultats, per raó dels decimals i valors negatius que hi són inherents. Per aquest motiu, s'han proposat algunes transformacions lineals de les puntuacions típiques que pretenen fer-les més intuïtivament interpretables.

• Totes aquestes escales derivades de l'escala de les puntuacions típiques es basen en una transformació genèrica del tipus:

$$D_i = a \cdot z_i + b,$$

A conseqüència d'aquesta transformació, les noves puntuacions D passaran de tenir una mitjana 0 i una desviació típica 1 a tenir una mitjana b i una desviació típica a.

• Diverses propostes d'escales derivades han sigut plantejades sense que s'haja generalitzat l'ús concret de cap d'aquestes. Les propostes que han tingut més repercussió són les següents:

- L'escala T → $T_i = 10 \cdot z_i + 50$ $(\overline{T} = 50 s_T = 10)$
- L'escala S o d'estanins → $S_i = 2 \cdot z_i + 5$ $(\overline{S} = 5\ s_S = 2)$
- L'escala CI → $CI_i = 15 \cdot z_i + 100$ $(\overline{\mathrm{CI}} = 100\ s_{\mathrm{CI}} = 15)$

Exercici 6: Transformeu les dades de la variable X: {3, 6, 5, 2} a les 3 escales derivades presentades. Nota: aquestes dades ja van ser transformades a puntuacions típiques en l'exercici 3.

Tema 5.1 – Associació: organització i representació gràfica de dades multivariades

• Després d'abordar en temes previs el tractament individualitzat de les variables (estadística univariada), en aquest i en els temes successius es descriuran una sèrie de procediments associats al tractament conjunt de dues o més variables. Aquests procediments estadístics permetran obtenir informació sobre la relació entre aquestes variables. L'exposició es limitarà, majoritàriament, al cas bivariat (dues variables) per a ser més senzill en la seua presentació i deixa l'anàlisi multivariada (més de dues variables) per a cursos més avançats.

1. La distribució conjunta multivariada

• De manera anàloga al cas univariat exposat en el tema 2, un resum bàsic de la informació d'un grup de 2 o més variables consisteix en la distribució conjunta de freqüències, la qual es basa en el recompte del nombre de casos (freqüències) que presenten les diferents combinacions de valors que empíricament s'hagen observat per a aqueixes variables. Les modalitats d'una distribució conjunta de freqüències consisteixen, no en els valors d'una variable concreta, sinó en totes les possibles combinacions dels valors de les variables que es consideren –exceptuant-ne aquelles combinacions que no s'hagen presentat empíricament i que, per tant, no té sentit incloure en la distribució de freqüències.

Exemple: La següent taula de dades procedeix d'un estudi sobre les relacions de parella en què es va obtenir informació en una mostra de 71 subjectes de les 3 variables següents: Sexe (1: Home; 2: Dona); Nombre de parelles estables al llarg dels últims 5 anys; i Situació emocional actual (1: Satisfactòria; 2: Ni satisfactòria ni insatisfactòria; 3: Insatisfactòria).

ID	Sexe	Nre_parelles	Sit_actual
1	1	1	3
2	1	4	2
3	2	1	1
4	2	2	1
5	2	1	3
6	1	0	1
7	2	3	2
...	...	...	...
71	1	1	1

L'organització de les dades de l'anterior taula en forma de distribució conjunta de freqüències absolutes seria la següent (sent X = Sexe; I = Nre parelles i Z = Sit_actual):

X_i ; I_i ; Z_i	n_i
1 ; 0 ; 1	4
1 ; 0 ; 2	3
1 ; 0 ; 3	2
1 ; 1 ; 1	12
1 ; 1 ; 2	8
1 ; 1 ; 3	6
1 ; 2 ; 1	5
1 ; 2 ; 2	1
1 ; 2 ; 3	2
1 ; 4 ; 2	1
2 ; 0 ; 1	6
...	...
2 ; 3 ; 2	1
	71

• La distribució conjunta de freqüències relatives o proporcions (p_i) i la de percentatges ($\%_i$) poden obtenir-se a partir de les freqüències absolutes en dividir cada freqüència absoluta pel nombre de casos (n), i en multiplicar les freqüències relatives per cent, respectivament.

• L'ordenació de les modalitats en una distribució conjunta de freqüències no té sentit, si bé se solen situar en ordre alfabètic/numèric creixent a fi de poder localitzar més fàcilment qualsevol combinació de valors de les variables.

• L'obtenció de les freqüències acumulades, ja siguen absolutes, relatives o percentatges, tampoc té sentit, atès que les modalitats de la distribució no representen un continu –igual que ocorria amb les distribucions de freqüències de les variables categòriques.

• Inconvenients: Si el nombre de variables és ampli o si alguna de les variables té molts valors, el nombre de combinacions de valors possibles pot arribar a ser molt nombrós, cosa que fa que la distribució de freqüències no siga un bon resum de les dades. Existeixen algunes alternatives que poden ajudar a resoldre aquest problema en algunes situacions:

(1) En el cas d'una variable (o més) amb molts possibles valors (com és el més habitual amb variables quantitatives), una opció és col·lapsar aqueixos valors en intervals. D'aquesta manera es perd en precisió de la informació, però es redueix el nombre de combinacions de valors possibles.

A tall d'**exemple**, suposem que tenim dues variables (X i Y), consistents en el temps (en segons) emprat per un grup de persones a executar dues tasques procedents d'un test d'aptituds mecàniques. Si els valors mínim i màxim són 0 i 20 segons, respectivament, en ambdues variables, una possible agrupació per crear la distribució conjunta de freqüències podria ser:

X_i, Y_i	n_i
0-5 ; 0-5	...
0-5 ; 5-10	...
0-5 ; 10-15	...
0-5 ; 15-20	...
5-10 ; 0-5	...
5-10 ; 5-10	...
...	...
15-20 ; 15-20	...

Quantes files tindrà l'anterior distribució conjunta de freqüències? Quantes files tindria la distribució conjunta de freqüències sense agrupar les dades en intervals?

(2) En el cas de moltes variables, una alternativa consisteix a aplicar algun dels mètodes estadístics que se solen englobar sota el qualificatiu de "tècniques de reducció de dades" (per exemple, l'anàlisi factorial, l'escalament multidimensional o l'anàlisi de correspondències), mètodes que escapen als continguts de la present assignatura.

1.1. La taula de contingència

• En el cas de dues variables, una forma molt convenient de visualitzar la distribució de freqüències conjunta és en forma de taula de contingència, això és, una taula de doble entrada en què les modalitats d'una variable ocupen les files i les de l'altra n'ocupen les columnes. En les caselles interiors de la taula apareixen les freqüències conjuntes (ja siguen absolutes, relatives o percentatges) corresponents a cadascuna de les possibles combinacions de les modalitats de files i columnes.

Exemple: es va dur a terme un estudi per avaluar si l'estat d'ànim dels majors de 65 anys podia estar influït pel fet de viure en una residència geriàtrica o no. Es van recollir dades d'una mostra de 500 persones de les variables "Estat d'ànim" [negatiu (−); neutre (±); positiu (+)] i "Viure en residència" [Sí; No]. La distribució conjunta de freqüències de totes dues variables en forma de taula de contingència és la següent:

	Sí	**No**
−	48	70
±	42	105
+	60	175

Com s'ha construït aquesta taula de contingència? Realitzant, a partir de la matriu de dades original, un recompte del nombre de casos que presenten cada combinació de les modalitats de les dues variables.

Cas	*Residència*	*Estat ànim*
1	Si	−
2	No	±
3	Si	−
4	Si	+
...	...	...
500	No	±

• També és possible obtenir a partir d'aquesta distribució conjunta de freqüències:

- La distribució de cada variable per separat (= distribucions marginals):

Residència (X)

X_i	n_i	p_i
Sí	150	0.30
No	350	0.70
	500	1

Estat ànim (Y)

Y_i	n_i	p_i
−	118	0.236
±	147	0.294
+	235	0.470
	500	1

- La distribució conjunta de freqüències:

X_i ; Y_i	n_i	p_i
Sí ; −	48	0.096
Sí ; ±	42	0.084
Sí ; +	60	0.120
No ; −	70	0.140
No ; ±	105	0.210
No ; +	175	0.350
	500	1

• En les taules de contingència és habitual afegir en els laterals dret i inferior, les sumes de les caselles corresponents a cada fila i columna, respectivament. Aquestes son són les anomenades com a distribucions marginals, és a dir, les distribucions univariades o simples de les dues variables.

	Sí	**No**	**Total**
−	48	70	118
±	42	105	147
+	60	175	235
Total	150	350	500

Exemple de la taula de contingència de les dues variables anteriors tal com és presentada en el programa SPSS:

Taula de contingència Estat ànim * Viure residència

Recompte

		Viure residència		
		Sí	No	Total
Estat ànim	Negatiu	48	70	118
	Neutre	42	105	147
	Positiu	60	175	235
Total		150	350	500

• En les taules de contingència es poden presentar també les freqüències relatives o percentatges:

	Sí	**No**	**Total**
−	0,096	0,140	0,236
±	0,084	0,210	0,294
+	0,120	0,350	0,470
Total	0,300	0,700	1

	Sí	**No**	**Total**
−	9,6	14	23,6
±	8,4	21	29,4
+	12	35	47
Total	30	70	100

El següent *output* mostra la taula de contingència obtinguda amb SPSS en el cas de sol·licitar que en les caselles de la taula apareguen, a més de les freqüències absolutes (que es mostren per defecte), el percentatge de casos en cada casella respecte del total de casos (Nota: les freqüències relatives o proporcions no es poden obtenir en SPSS):

Taula de contingència Estat ànim * Viure residència

			Viure residència		
			Sí	No	Total
Estat ànim	Negatiu	Recompte	48	70	118
		% del total	9,6%	14,0%	23,6%
	Neutre	Recompte	42	105	147
		% del total	8,4%	21,0%	29,4%
	Positiu	Recompte	60	175	235
		% del total	12,0%	35,0%	47,0%
Total		Recompte	150	350	500
		% del total	30,0%	70,0%	100,0%

• Pel que fa a la disposició de les variables en les files i columnes de la taula de contingència, si considerem que la relació entre totes dues variables és simètrica, llavors és indiferent quina variable es posa en les files i quina en les columnes. Al contrari, en el cas que la relació entre totes dues variables siga asimètrica, s'acostuma a situar en les files la variable de resposta i en les columnes la variable explicativa, com és el cas en el nostre exemple, en què "Estat d'ànim" (variable de resposta) se situa en les files de la taula, mentre que "Viure en una residència" (variable explicativa) se situa en les columnes.

• Les files i columnes interiors (sense la columna i fila de les distribucions marginals) d'una taula de contingència són anomenades distribucions condicionals. Per exemple, la primera columna de la nostra taula d'exemple (48, 42, 60) és la distribució condicional de la variable "Estat d'ànim" per a aquells subjectes que *Sí* que viuen en una residència. La segona columna (70, 105, 175) és la distribució condicional de la variable "Estat d'ànim" per a aquells subjectes que *No* viuen en una residència. Anàlogament, per a la variable "Viure en una residència" es poden diferenciar 3 distribucions condicionals: (48, 70), (42, 105) i (60, 175).

• La comparació de les distribucions condicionals d'una variable en funció d'una segona variable és fonamental per a valorar si hi ha o no relació entre aqueixes dues variables. En el següent tema es concreta com dur a terme tal comparació a fi d'analitzar el grau d'associació existent entre dues variables categòriques o ordinals.

Exercici 1: Tenim les variables *X* (Aplicació d'un programa d'intervenció per a afavorir la interacció social [Sí (1), No (0)]) i *Y* (Grau d'interacció en l'hora de l'esbarjo [Baix (1), Mitjà (2), Alt (3)]), de què tenim dades en un grup de 20 alumnes d'una classe en què es va avaluar l'eficàcia d'aquest programa d'intervenció.

ID	*X*	*Y*
1	1	2
2	1	3
3	0	2
4	1	2
5	1	1
6	0	1
7	0	2
8	0	2
9	1	3
10	0	2
11	1	2
12	1	1
13	1	3
14	0	2
15	0	1
16	1	2
17	0	3
18	0	1
19	0	2
20	1	2

a) Organitzeu les dades de les variables *X* e *Y* mitjançant una distribució conjunta de freqüències.

b) Organitzeu les dades mitjançant una taula de contingència de freqüències absolutes i una de proporcions.

Exercici 2: S'ha obtingut amb SPSS la següent taula de contingència entre les variables "Disfrutar amb les explicacions en classe" i "Motivació cap als estudis" en un grup de 174 estudiants de Psicologia (dades procedents de l'enquesta sobre la vida acadèmica). Empleneu els interrogants que apareixen en la taula.

			Motivació estudis psicologia			
			alta	mitja	baixa	Total
Disfrutar amb les explicacions	sempre o quasi sempre	Recompte	???	8	0	24
		% del total	9,2%	???	0,0%	13,8%
	algunes vegades	Recompte	74	65	5	???
		% del total	42,5%	37,4%	2,9%	82,8%
	mai o quasi mai	Recompte	1	4	1	6
		% del total	0,6%	2,3%	0,6%	3,4%
Total		Recompte	91	???	6	174
		% del total	52,3%	44,3%	3,4%	100,0%

Exercici 3: A partir d'una enquesta sobre condicions psicosocials en el lloc de treball realitzada a una mostra de 1000 treballadors, trobem que un 8 % va manifestar haver patit assetjament psicològic en el treball i, d'aquests, un 20 % treballen en una petita empresa, un 50 % en una empresa mitjana i un 30 % en una gran empresa. A més, sabem que dels 1000 treballadors, 200 treballen en petites empreses, mentre que 300 treballen en empreses mitjanes. A partir d'aquesta informació, obtingueu la taula de contingència de les dues variables implicades, tant en freqüències absolutes com en freqüències relatives.

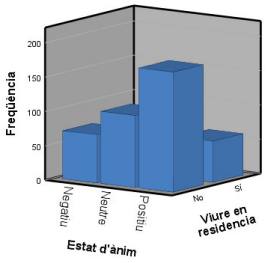
2. Representacions gràfiques

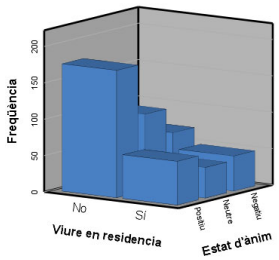
• Tot seguit es presenten un conjunt de gràfics per a representar la distribució conjunta de dues o més variables, en funció de l'escala de mesura d'aquestes variables. No es presentaran gràfics específics per a les variables ordinals, ja que es poden utilitzar qualsevol dels orientats a variables categòriques, o bé els orientats a les variables quantitatives, si s'assumeix la naturalesa quantitativa d'aquestes variables ordinals.

2.1. El cas de dues variables categòriques

• El gràfic de barres tridimensional o 3-D

Exemples de gràfic de barres 3-D amb la distribució conjunta de freqüències absolutes de "Estat d'ànim" i "Viure residència" i amb la posició de totes dues variables intercanviada.

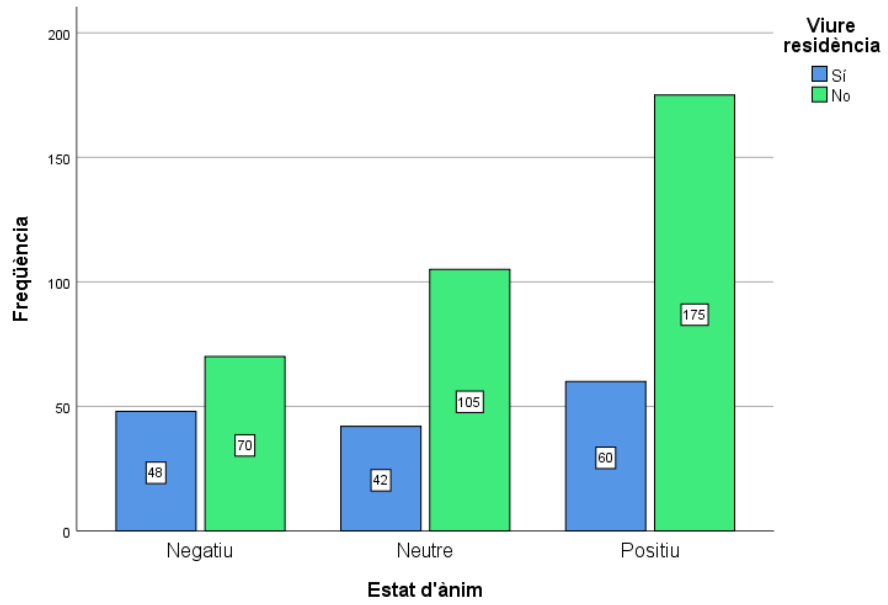


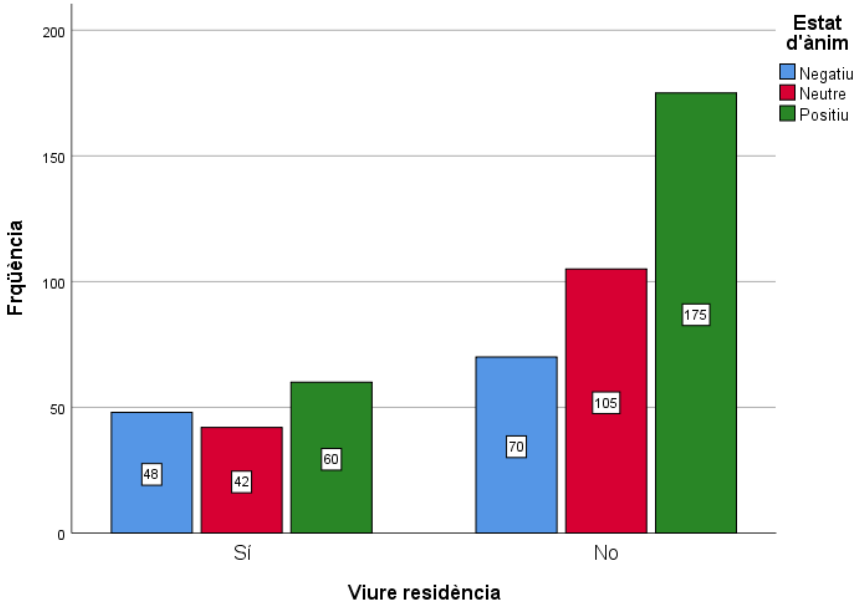


• El gràfic de barres agrupat

Exemples de gràfic de barres agrupat en què es representa la distribució conjunta de freqüències absolutes de "Estat d'ànim" i "Viure residència" i amb la posició de totes dues variables intercanviada en el gràfic. Per a diferenciar verbalment tots dos, farem referència al primer com a gràfic de barres agrupat de freqüències absolutes de la variable "Estat d'ànim" en funció de "Viure residència", mentre que al segon, com a gràfic de barres agrupat de freqüències absolutes de la

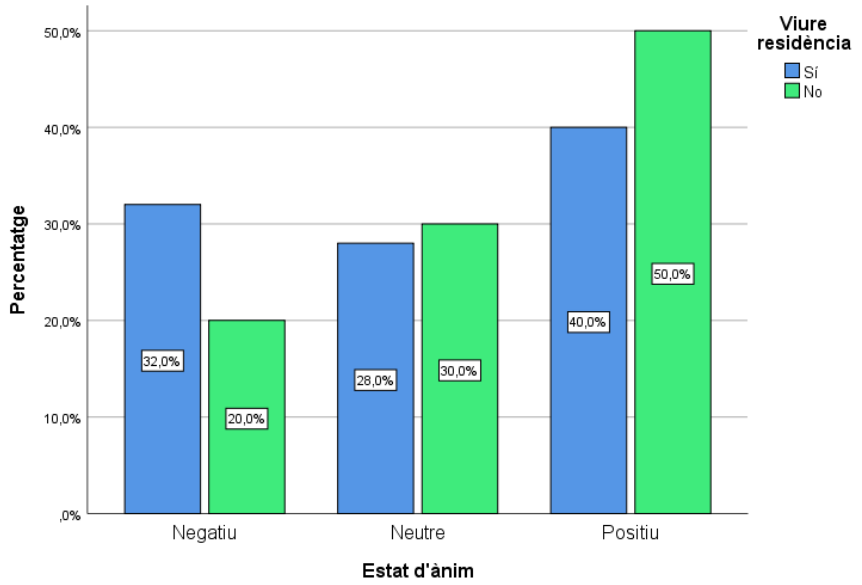
variable "Viure residència" en funció de "Estat d'ànim". En tots dos gràfics es representen freqüències absolutes, per la qual cosa les barres en tots dos han de sumar el total de la grandària de la mostra (n = 500).





• Quin dels dos gràfics és millor? –Suposem que l'objectiu de les dades recollides és esbrinar si el fet de viure o no en una residència influeix sobre l'estat d'ànim. Amb quin dels dos us resulta més fàcil valorar si existeix tal relació entre les variables? El favorit no ha de ser el mateix per a tothom. Un problema que pot ser que ja hàgem detectat en intentar contestar la pregunta anterior és que el diferent nombre de persones majors que viuen en una residència (150) i que no hi viuen (350) complica la realització d'una interpretació correcta de qualsevol dels dos gràfics. Una manera de superar aquest problema consisteix a representar les freqüències relatives condicionades o els percentatges condicionats, si bé deixem per al pròxim tema com fer-ho.

• El gràfic de barres agrupat de freqüències absolutes pot ser fàcilment obtingut amb el programa SPSS. Cal tenir en compte que, quan en aquest programa se sol·licita que es representen els percentatges (l'opció de freqüències relatives no s'ofereix), el que es representa no són els percentatges de cada casella respecte del total de casos (vegeu exemple a continuació), sinó els percentatges condicionats, que tractarem en el pròxim tema. Com a exemple, pot comprovar-se en el següent gràfic com el total de les barres d'aquest gràfic no suma 100.

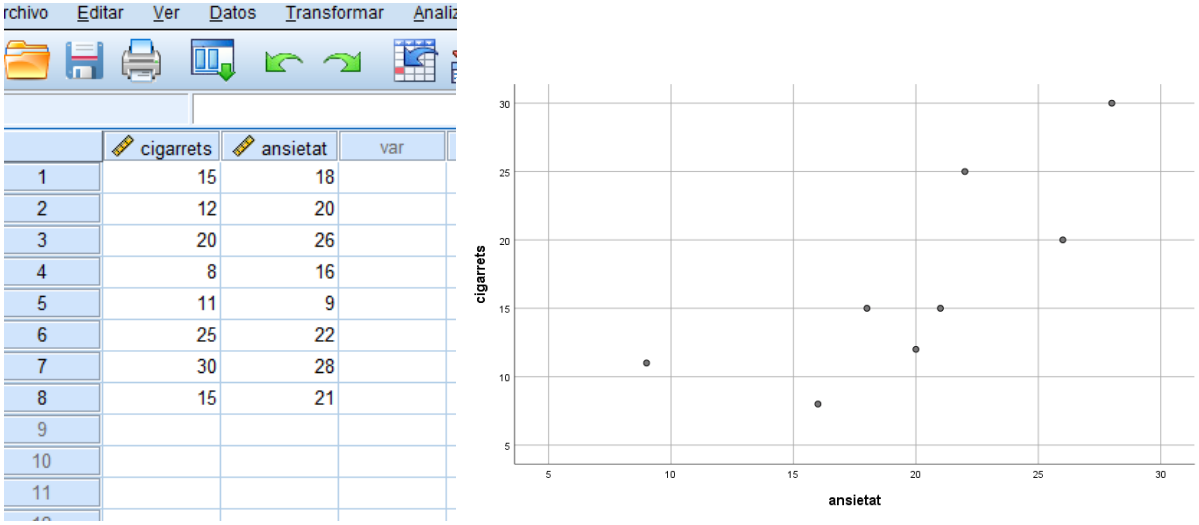


Exercici 4: A partir de les dades de les variables "Aplicació d'un programa d'intervenció per a afavorir la interacció social" i "Grau d'interacció en l'hora de l'esbarjo" (vegeu l'exercici 1), representeu gràficament la distribució conjunta de totes dues variables.

2.2. El cas de dues variables quantitatives

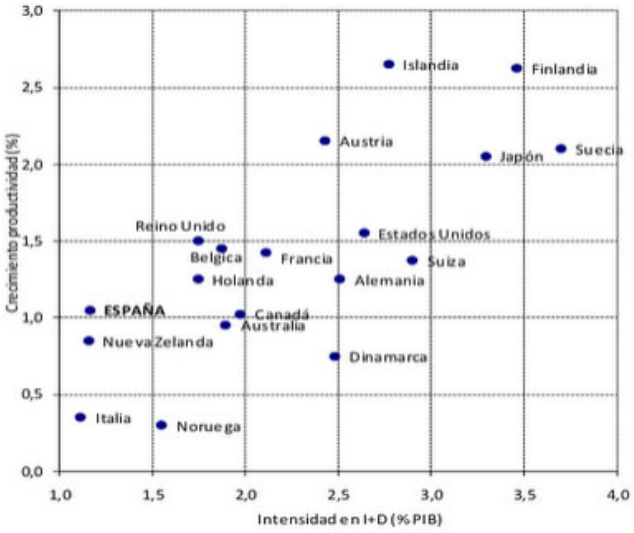
• El diagrama de dispersió

Exemple de diagrama de dispersió obtingut a partir de les dades d'una mostra de 8 fumadors en les variables "Nombre de cigarrets que es fuma al dia" i "Puntuació en un test d'ansietat [0, ..., 30]". Es mostren també les dades a partir de les quals se l'ha sigut obtingut amb el programa SPSS:



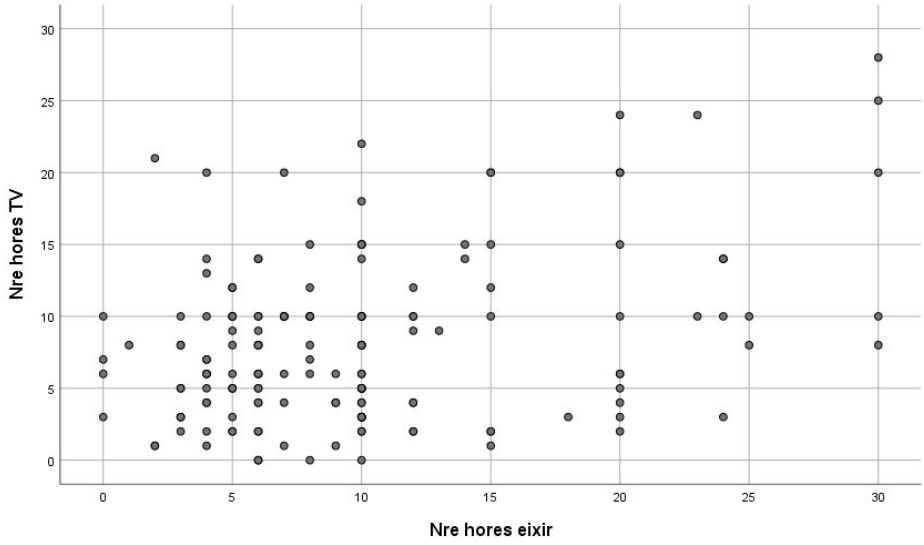
	cigarrets	ansietat	var
1	15	18	
2	12	20	
3	20	26	
4	8	16	
5	11	9	
6	25	22	
7	30	28	
8	15	21	
9			
10			
11			

A continuació es mostra un altre **exemple** a partir de les dades de diversos països en les variables: (1) inversió en I+D mesurada com a percentatge del PIB del país; (2) creixement en la productivitat expressat en percebtatges (Fonts: Eurostat i OCDE, període 2001-2007):

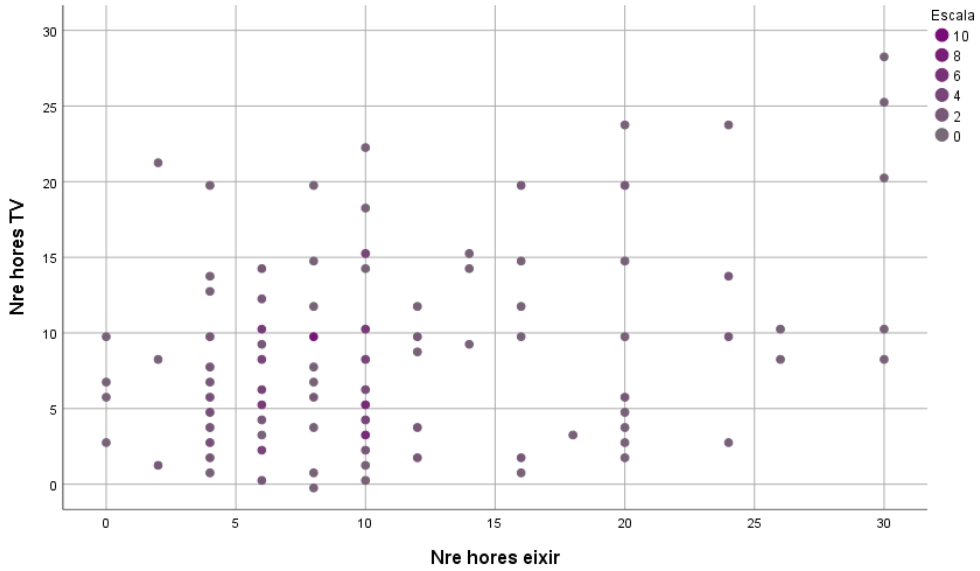


En l'anterior diagrama de dispersió, els punts apareixen etiquetats amb el nom del país corresponent, la qual cosa facilita una interpretació més detallada de la distribució conjunta de totes dues variables. Aquesta estratègia d'etiquetar el núvol de punts pot resultar interessant, encara que no ho seria si tenim un arxiu de dades amb molts casos perquè el gràfic podria ser inintel·ligible.

Un problema que es pot presentar en la representació d'un diagrama de dispersió és el de la superposició dels punts, això és, que hi haja casos amb els mateixos valors en totes dues variables, la qual cosa és freqüent en arxius de dades amb molts casos. Com a exemple, el següent diagrama de dispersió de les variables "Nombre d'hores a la setmana dedicades a eixir" i "Nombre d'hores a la setmana dedicades a veure televisió" en una mostra de 174 estudiants (exemple procedent del qüestionari de vida acadèmica):

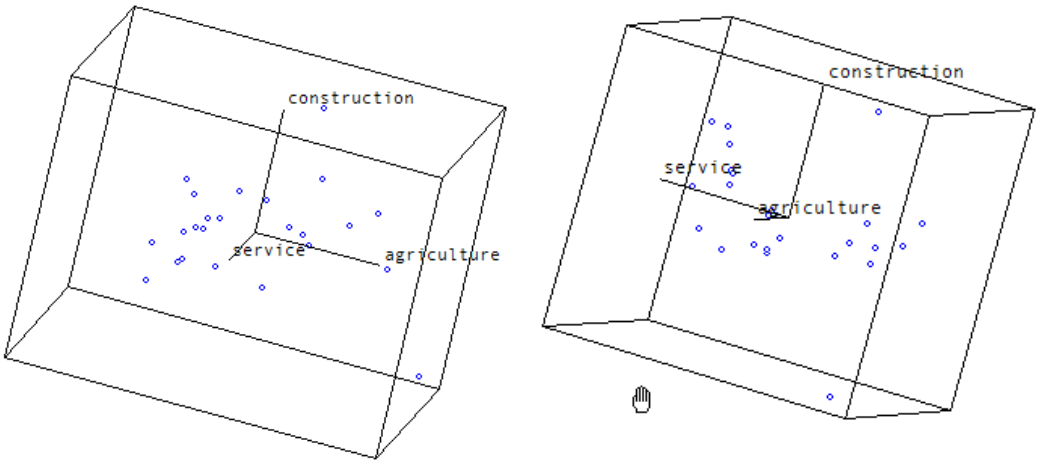


En el gràfic anterior no apareixen 174 punts perquè hi ha casos superposats, es a dir, que tenen els mateixos valors en ambdues variables. Alguns programes com SPSS permeten dimensionar els punts en funció del nombre de casos que coincideixen en la mateixa posició, la qual cosa permet obtenir una visualització més realista de la distribució conjunta de les dades. En el cas concret de l'exemple anterior, es mostra tot seguit el diagrama de dispersió amb els punts dimensionats (on posa 0 en l'escala dels punts se suposa que és 1):



• <u>El diagrama de dispersió amb 3 variables</u>:

Exemple de diagrama de dispersió amb el percentatge de població activa en tres sectors productius (agricultura, serveis i construcció) d'un conjunt de països europeus (dues instantànies d'aquestes dades obtingudes a partir de la rotació del gràfic amb el programa Vista):



La pobra visualització d'aquest tipus de diagrames de dispersió sobre el paper pot millorar-se si s'utilitza un programa que permeta la rotació del gràfic en qualsevol direcció, perquè això permet fer-se una idea més real de com és el núvol de punts tridimensional.

Exercici 5: Les següents dades procedeixen d'un estudi en què es van obtenir dades de 16 subjectes sobre el nombre d'hores d'esport que practicaven setmanalment (X) i la percepció que tenien sobre el seu estat de salut general (Y) en una escala d'1 a 10, on una major puntuació indica una percepció més positiva de la pròpia salut. Representeu gràficament la distribució conjunta de freqüències.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8
I	4	3	3	5	6	4	4	6	5	2	7	9	6	8	9	8

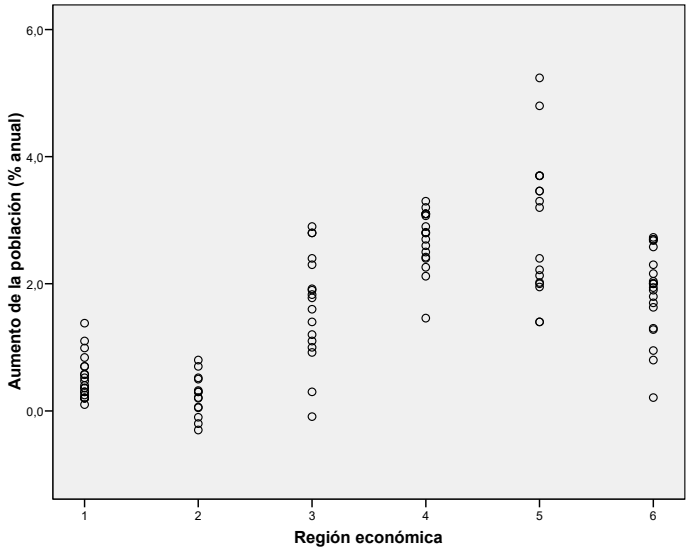
2.3. El cas d'una variable categòrica i una variable quantitativa

• El polígon de freqüències agrupat es construeix dibuixant un polígon de freqüències de la variable quantitativa per a cadascun dels subgrups definits per la variable categòrica. Aquest tipus de gràfic facilita la superposició gràfica ja que és fàcil visualitzar diferents línies en un mateix espai.

Exemple de polígon de freqüències agrupat per a la distribució de freqüències absolutes de la variable "Qualificació en una prova" [Suspens; Aprovat; Notable; Excel·lent; Matrícula d'Honor] en funció de la variable "Sexe" (Nota: encara que la variable "Qualificació" pot ser considerada com a ordinal, s'ha assumit ací el seu caràcter quantitatiu).

• El diagrama de dispersió també es pot aplicar en la representació conjunta de la distribució de freqüències absolutes d'una variable categòrica i una variable quantitativa. Aquest tipus de gràfic s'anomena en alguns textos gràfic de punts i és habitual que se situe la variable categòrica en l'eix d'abscisses i la variable quantitativa en l'eix d'ordenades.

Exemple de diagrama de dispersió (gràfic de punts) de la distribució conjunta de les variables "Regió econòmica" [1:OCDE; 2: Europa oriental; 3: Àsia/Pacífic; 4: Àfrica; 5: Orient Mitjà; 6: Amèrica llatina]" i "Percentatge anual de creixement de la població" obtinguda a partir de les dades recollides per a un total de 109 països de tot el món (N = 109):



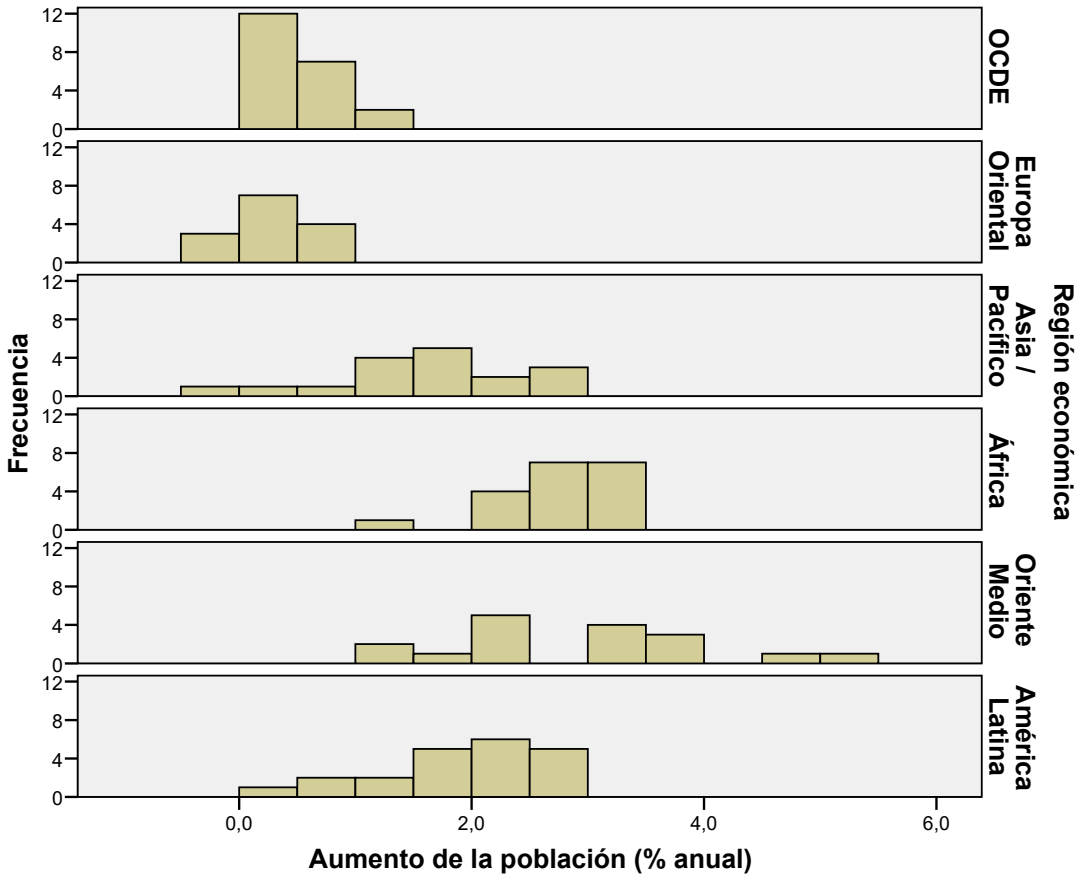
Penseu com serà la taula de dades a partir de la qual s'ha obtingut l'anterior gràfic? Quantes files i columnes deu tenir?

Aquest és un fragment de la taula de dades:

1	Acerbaján	Oriente Medio	1,4
2	Afganistán	Asia / Pacífico	2,8
3	Alemania	OCDE	,4
4	Arabia Saudí	Oriente Medio	3,2
5	Argentina	América Latina	1,3
6	Armenia	Oriente Medio	1,4
7	Australia	OCDE	1,4
8	Austria	OCDE	,2
9	Bahrein	Oriente Medio	2,4
10	Bangladesh	Asia / Pacífico	2,4
11	Barbados	América Latina	,2
12	Bélgica	OCDE	,2
13	Bielorusia	Europa Oriental	,3
14	Bolivia	América Latina	2,7
15	Bosnia	Europa Oriental	,7
16	Botswana	África	2,7
17	Brasil	América Latina	1,3
18	Bulgaria	Europa Oriental	-,2
19	Burkina Faso	África	2,8

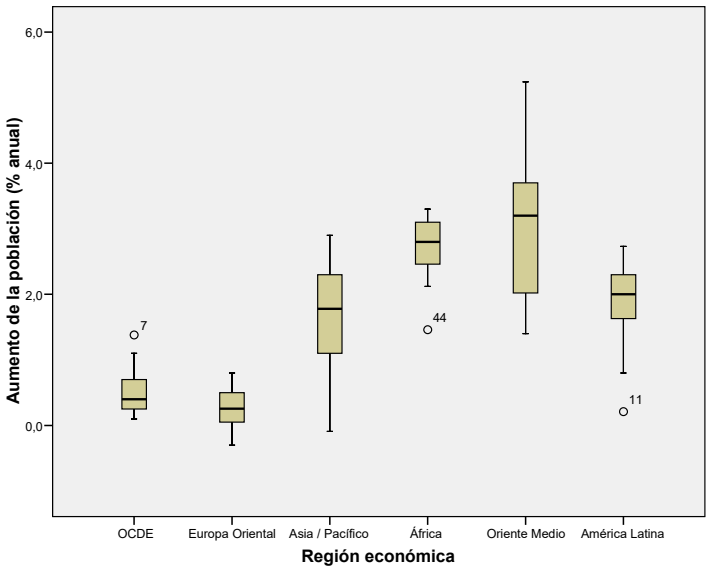
• El panell d'histogrames ofereix la visualització en forma d'histograma de la variable quantitativa agrupada en funció dels valors de la variable categòrica.

Exemple per a les variables "Regió econòmica" i "P anual de creixement de la població":



• El gràfic de caixa i bigots agrupat representa la distribució de la variable quantitativa en funció dels valors de la variable categòrica.

Exemple de gràfic de caixa i bigots de la variable "Percentatge anual de creixement de la població" en funció de la variable "Regió econòmica":



Exercici 6: Tenim les variables *X* (Aplicació d'un programa d'intervenció per a afavorir la interacció social [Sí (1), No (0)]) i *Y* (Grau d'interacció en l'hora de l'esbarjo, mesurada a partir del nombre de minuts en què s'ha participat en activitats amb altres companys). Tenim les dades d'un grup de 20 alumnes d'una classe en la qual es va avaluar l'eficàcia d'aquest programa d'intervenció. Representeu gràficament les dades obtingudes.

ID	*X*	*Y*
1	1	22
2	1	13
3	0	12
4	1	27
5	1	19
6	0	16
7	0	20
8	0	12
9	1	23
10	0	17
11	1	29
12	1	16
13	1	30
14	0	20
15	0	15
16	1	24
17	0	23
18	0	18
19	0	20
20	1	18

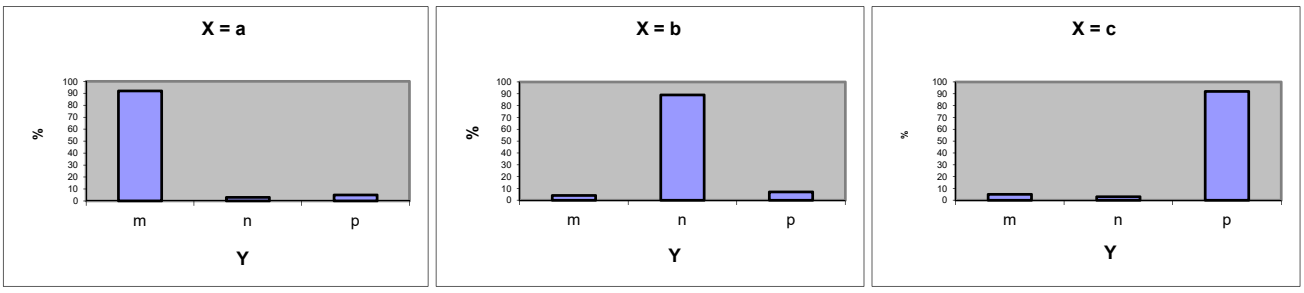
Tema 5.2 – Associació: estadístics d'associació entre variables

1. Concepte d'associació entre variables

• L'anàlisi estadística de l'associació (relació, covariància, correlació) entre variables representa una part bàsica de l'anàlisi de dades perquè una gran part de les preguntes i hipòtesis que es plantegen en els estudis impliquen analitzar la presència o no de relació entre variables.

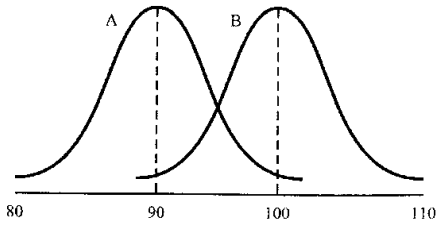
• L'existència d'associació entre dues o més variables representa la presència d'algun tipus de tendència o patró de relació entre els valors d'una variable i l'altra.

> Com a **exemple**, si tenim una variable *X* [*a, b, c*] i una variable *Y* [*m, n, p*], de manera que les dades empíriques mostren que els casos que en *X* són *a* tendeixen a ser *m* en *Y*, que els que en *X* són *b* tendeixen a ser *n* en *Y*, i que els que en *X* són *c* tendeixen a ser *p* en *Y*, això posa de manifest l'existència d'associació entre ambdues variables. Gràficament, les distribucions de freqüències (expressades en percentatges) de la variable *Y* quan *X* és igual a *a*, *b* i *c* són, respectivament:



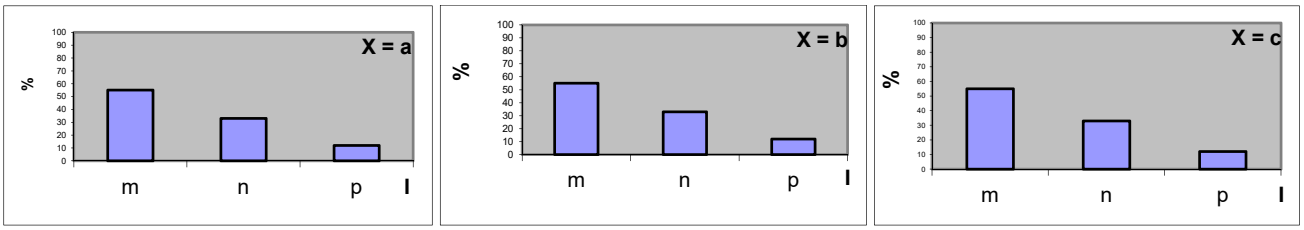
• Més formalment, Solanas *et al*. (2005) ofereixen una altra definició general d'associació entre dues variables: l'existència d'associació entre dues variables indicaria que la distribució dels valors d'una de les dues variables difereix en funció dels valors de l'altra.

Un altre **exemple** de la presència d'associació entre dues variables: la puntuació en un test d'aptitud lingüística [0 a 150] (quantitativa) i el sexe [A: Home; B: Dona] (categòrica). Tot seguit es mostren les distribucions de freqüències d'aptitud en funció del sexe. En aquest gràfic s'observa que el grup de dones (B) té puntuacions superiors al grup d'homes (A); les mitjanes aritmètiques són 100 i 90 per a dones i homes, respectivament. Així, es pot concloure que ambdues variables estan relacionades.



• Complementàriament, es parla d'independència entre variables quan no existeix aquest patró de relació entre els valors d'ambdues variables.

Seguint amb l'**exemple** anterior, *X* i *Y* són independents si la distribució de freqüències de *Y* no canvia en funció de *X*, tal com es mostra a continuació:



Per a l'**exemple** del test d'aptitud numèrica i el sexe, les dues variables són independents si les puntuacions d'aptitud numèrica són iguals per homes i dones i, per tant, ambdues distribucions se superposarien.

• L'associació entre variables no s'ha d'entendre com una qüestió de tot o res, sinó com un continu des de l'absència de relació (independència) fins a la relació total entre les variables. Aquest últim extrem representa una relació determinista, és a dir, quan a partir del valor d'un subjecte en qualsevol de les variables, es pot afirmar quin serà el seu valor en l'altra variable.

• Cal afegir que es comú utilitzar l'expressió grandària de l'efecte per a fer referència a la intensitat de la relació entre dues variables.

2. Mesurar l'associació entre dues variables

2.1. El cas de dues variables categòriques

• Què podem concloure sobre l'associació entre les dues variables de la taula de contingència, "Estat d'ànim" i "Viure en residència", a partir de les dades recollides en una mostra de 500 persones majors 70 anys? Nota: la variable "Estat d'ànim" es va mesurar mitjançant una escala amb 3 categories ordenades: negatiu (−), neutre (±) i positiu (+).

	Sí	*No*	**Total**
−	48	70	118
±	42	105	147
+	60	175	235
Total	150	350	500

• Amb l'objectiu d'avaluar si ambdues variables estan relacionades, cal observar si la distribució dels valores d'una de les variables difereix en funció dels valors de l'altra (distribucions condicionades). Si aquestes distribucions són iguals, això significa que no hi ha relació entre les variables. Per exemple, podem comparar les distribucions de freqüències d'"Estat d'ànim" entre els que *Sí* viuen en una residència (48, 42, 60) i els que *No* hi viuen (70, 105, 175). No són iguals, però podem estar segurs que hi ha relació entre ambdues variables?

• En fixar-nos en les distribucions de freqüències d'"Estat d'ànim" per cadascun dels valors de "Viure en residència" [Sí; No], hem observat que aquestes no són iguals, però també és cert que difícilment ho podran ser pel fet que hi ha més subjectes que no viuen en una residència (350) que subjectes que sí que hi viuen (150). En conclusió, no és recomanable comparar les distribucions condicionades en freqüències absolutes si el nombre de casos difereix per les distintes categories de l'altra variable.

• Per a superar aquest inconvenient, es recorre a la taula de contingència de freqüències relatives condicionades, en la qual s'anul·la l'efecte de la possible diferència de grandària dels grups. Aquest tipus de taula es pot obtenir de dues formes alternatives, bé dividint les freqüències absolutes de cada casella entre el total de casos en les seues files, bé dividint-les entre el total de casos en les seues columnes. Totes dues taules permetran arribar a la mateixa conclusió respecte a l'associació entre les dues variables.

• En la pràctica, si la relació entre les variables és asimètrica, és habitual considerar com a variable condicionant la variable explicativa (predictora, independent). Per exemple, en un estudi en què es va avaluar la influència del "Nivell d'estudis" [primaris, secundaris, superiors] sobre la "Percepció de la influència de la ciència en la societat" [negativa, indiferent, positiva], atès que el nivell d'estudis era la variable explicativa, hauríem de comparar les distribucions de freqüències relatives condicionades de "Percepció de la influència de la ciència" en funció del "Nivell d'estudis", és a dir, per a cada categoria de nivell d'estudis.

En l'**exemple** sobre "Estat d'ànim" i "Viure en residència", si assumim que la segona variable influeix sobre la primera (relació asimètrica), haurem de comparar les distribucions de freqüències relatives condicionades de "Estat d'ànim" en funció de "Viure en residència":

	Sí	*No*	**Total**
−	0,32 (48/150)	0,20 (70/350)	0,236 (118/500)
±	0,28 (42/150)	0,30 (105/350)	0,294 (147/500)
+	0,40 (60/150)	0,50 (175/350)	0,470 (235/500)
Total	1	1	1

El següent *output* mostra la taula de contingència anterior obtinguda amb SPSS (les freqüències absolutes es mostren sempre, per defecte; en canvi, és una opció que es mostren els percentatges respecte als totals de les files o columnes):

Tabla de contingencia Estado ánimo * Vivir residencia

			Vivir residencia		
			Sí	No	Total
Estado ánimo	Negativo	Recuento	48	70	118
		% dentro de Vivir residencia	32,0%	20,0%	23,6%
	Neutro	Recuento	42	105	147
		% dentro de Vivir residencia	28,0%	30,0%	29,4%
	Positivo	Recuento	60	175	235
		% dentro de Vivir residencia	40,0%	50,0%	47,0%
Total		Recuento	150	350	500
		% dentro de Vivir residencia	100,0%	100,0%	100,0%

• En la taula anterior, la comparació de les distribucions condicionals de freqüències relatives d'"Estat d'ànim" en funció de "Viure en residència" ens permetrà comprovar l'existència d'associació entre les

dues variables. En cas afirmatiu, com ho és per a l'exemple que ens ocupa, la comparació d'aquestes distribucions condicionals amb la distribució marginal o simple de la variable de resposta ens permetrà veure clarament quina és la naturalesa d'aquesta relació.

A tall d'**exemple**, si no hi haguera relació entre "Estat d'ànim" i "Viure en residència", les distribucions de freqüències relatives d'"Estat d'ànim" serien iguals per als que *Sí* que viuen en una residència i per als que *No* hi viuen. A més, ambdues serien iguals en la distribució de freqüències marginal o simple (columna Total) de la variable "Estat d'ànim":

	Sí	*No*	**Total**
−	0,236	0,236	0,236
±	0,294	0,294	0,294
+	0,470	0,470	0,470
Total	1	1	1

• La taula de contingència amb les vertaderes distribucions de freqüències relatives d'"Estat d'ànim" en funció de "Viure en residència" (vegeu la taula més avall) difereix bastant de la taula d'independència presentada més amunt i posa de manifest l'existència d'associació entre totes dues variables. Una anàlisi més exhaustiva d'aquesta relació ens permet observar, per exemple, que la proporció de subjectes que tenen un estat d'ànim negatiu entre els que viuen en una residència (0,32) és superior a la mateixa proporció per als que no hi viuen (0,20), o que la proporció de subjectes que tenen un estat d'ànim positiu entre els que no viuen en una residència (0,50) és superior a la mateixa proporció entre els que sí hi viuen (0,40).

	Sí	*No*	**Total**
−	0,32	0,20	0,236
±	0,28	0,30	0,294
+	0,40	0,50	0,470
Total	1	1	1

• Si la relació entre les variables és simètrica és indiferent quina variable siga la variable condicionant. Així, per exemple, si desitgem valorar si hi ha relació entre el lloc de residència (rural o urbà) i la branca de batxiller cursada (ciències, socials, salut o humanitats) i no considerem a priori que una de les dues variables siga la variable explicativa i l'altra la variable de resposta, podríem comparar bé les distribucions de freqüències relatives de "Lloc de residència" en funció de "Batxiller" o bé les distribucions de freqüències relatives de "Batxiller" en funció de "Lloc de residència". Les conclusions a què s'arribe seran les mateixes en ambdós casos.

Exercici 1: Analitzeu l'associació entre les dues variables dicotòmiques següents: "Participació en un programa d'intervenció escolar que pretén afavorir el rendiment acadèmic [Sí, No]" i "Resultats acadèmics a final de curs [Bons, Dolents]" a partir de les dades obtingudes en una mostra de 100 escolars d'un col·legi (Clg_1):

Clg_1	Sí	No
Bons	18	42
Dolents	12	28

En un segon col·legi (Clg_2) s'aplica aquest mateix programa d'intervenció a una mostra també de 100 estudiants i s'han obtingut les dades resumides en la següent taula de contingència. Analitzeu i interpreteu l'associació existent entre totes dues variables en aquest segon col·legi.

Clg_2	Sí	No
Bons	24	31
Dolents	16	29

Finalment, les dades recollides en un tercer col·legi (*Clg_3*) es mostren resumides en la següent taula. Analitzeu i interpreteu l'associació existent entre ambdues variables en aquest col·legi.

Clg_3	Sí	No
Bons	15	33
Dolents	42	10

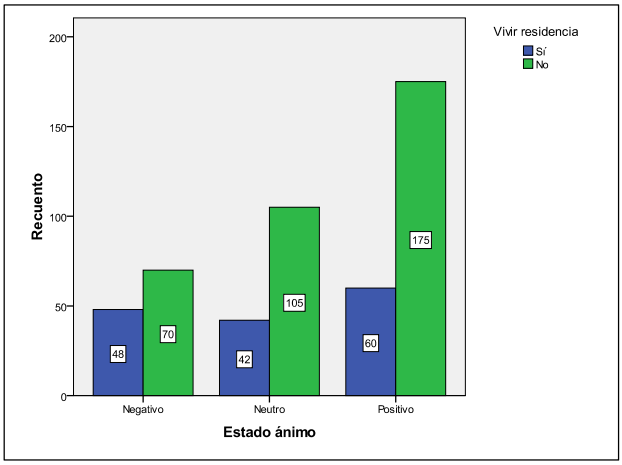
• L'anàlisi gràfica de l'associació entre 2 variables categòriques pot intuir-se a partir d'un gràfic de barres agrupat de freqüències absolutes si ens fixem en la forma de cadascuna de les distribucions condicionals (vegeu els grups de barres del mateix color en els exemples que es mostren més avall): com més similar siga la forma relativa d'aquestes, menys relació hi haurà entre les dues variables. Sens dubte, ens resultarà més fàcil visualitzar aquesta informació si el que es representa són els percentatges condicionats, perquè així s'elimina l'efecte, si escau, de la diferència de grandària dels grups. Com més similar siga la forma de les distribucions condicionals (vegeu els grups de barres del mateix color), menor serà la relació existent entre les variables. El gràfic de rectangles partits agrupat, quan és representen els percentatges condicionades, pot ser també apropiat per a avaluar l'existència d'associació entre dues variables categòriques (vegeu-ne exemples més avall).

• Quan la relació entre les variables és asimètrica, és pràctica habitual situar la variable de resposta sobre l'eix horitzontal del gràfic de barres agrupat (anomenat "eix de categories" en SPSS).

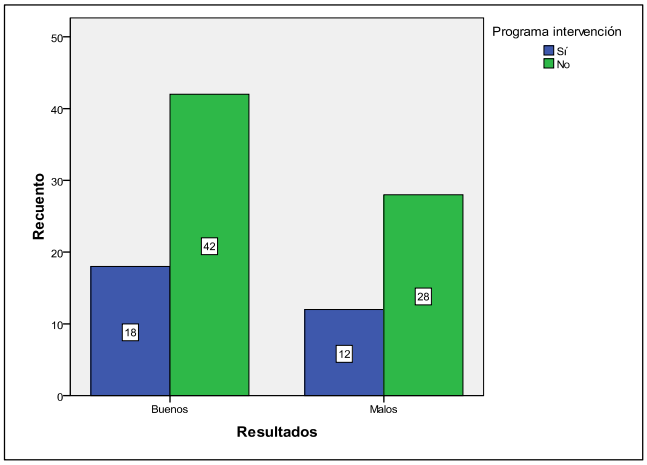
Exemples de gràfics de barres agrupats obtinguts amb SPSS per a les dades de les variables "Estat d'ànim" i "Viure en residència" (esquerra), i per a les dades de les variables "Programa d'intervenció" i "Resultats acadèmics" en el Col·legi 1 (dreta):

(1) Exemples de gràfics de barres agrupats amb freqüències absolutes

(adequat només quan els grups que es comparen són de la mateixa grandària, la qual cosa no ocorre en aquests exemples):

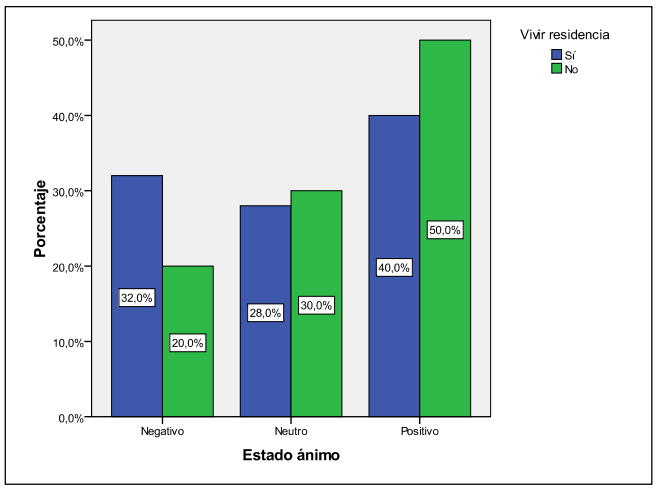


"Estat d'ànim" en funció de "Viure en residència"

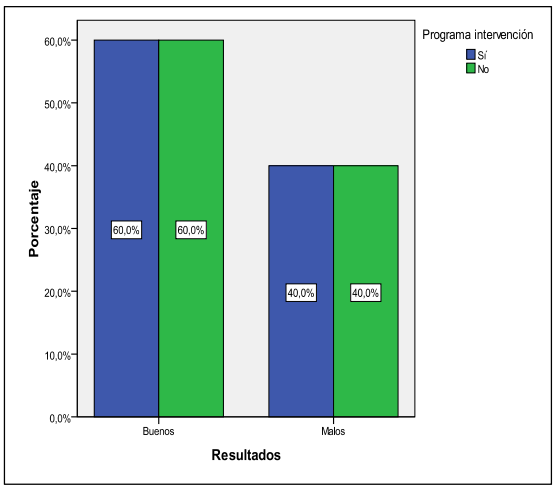


"Resultats acadèmics" en funció de "Programa d'intervenció"

(2) Exemples de gràfics de barres agrupats amb percentatges condicionats



"Estat d'ànim" en funció de "Viure en residència"



"Resultats acadèmics" en funció de "Programa d'intervenció"

Exercici 2: Feu (1) el gràfic de barres agrupat amb freqüències absolutes i (2) el gràfic de barres agrupat amb freqüències relatives condicionades per a les variables "Programa d'intervenció" i "Resultats acadèmics" en el Col·legi 2.

Exercici 3: Per a analitzar l'associació entre les variables "Motivació amb els estudis de Psicologia" i "Gaudir amb les explicacions" s'ha obtingut amb SPSS la següent taula de contingència (dades procedents de l'enquesta sobre la vida acadèmica).

Tabla de contingencia Disfrutar con las explicaciones * Motivación estudios Psicología

			Motivación estudios Psicología			
			alta	media	baja	Total
Disfrutar con las explicaciones	siempre o casi siempre	Recuento	16	8	0	24
		% dentro de Motivación estudios Psicología	???%	10,4%	,0%	???%
	algunas veces	Recuento	74	???	5	144
		% dentro de Motivación estudios Psicología	81,3%	84,4%	83,3%	82,8%
	casi nunca o nunca	Recuento	1	4	1	6
		% dentro de Motivación estudios Psicología	1,1%	???%	16,7%	3,4%
Total		Recuento	91	77	???	174
		% dentro de Motivación estudios Psicología	100,0%	100,0%	100,0%	100,0%

a) Empleneu els interrogants que apareixen en la taula de contingència.

b) Quina és la distribució marginal de freqüències absolutes de la variable "Motivació…"? I quina n'és la de "Gaudir…"?

c) A quina distribució corresponen els valores [16; 74; 1]?

Segueix…

d) A quina distribució corresponen els valores [10,4; 84,4; 5,2]?

e) Quines serien les tres distribucions condicionades de "Disfrutar…" (en percentatges condicionats) si ambdues variables foren independents?

f) Sembla haver-hi relació entre "Motivació…." i "Gaudir…"?

g) Feu un gràfic adequat per a avaluar la relació entre totes dues variables.

2.1.1 Índexs estadístics orientats a quantificar l'associació entre dues variables categòriques

• Tot seguit es presenten els índexs estadístics més utilitzats en la pràctica per a aquesta finalitat:

(1) L'índex ***khi ("xi") quadrat de Pearson*** (χ^2):

• El valor mínim de l'índex χ^2 és 0 quan les dues variables són independents i serà major que 0 si existeix relació entre les dues, més gran com més intensa siga. Ara bé, no té un límit màxim, la qual cosa suposa una dificultat a nivell interpretatiu.

• Un problema important de χ^2 és que el seu valor no depèn només de la intensitat de la relació entre les dues variables, sinó també de la grandària de la mostra (n) a partir de la qual s'obtinga, de manera que com més gran siga n, major serà també el valor de χ^2.

• No hi haurà inconvenient en la interpretació de χ^2 quan s'utilitze amb finalitat comparativa, sempre que la grandària de la mostra siga la mateixa i que les taules de contingència tinguen la mateixa dimensió (I x J). Si es compleixen les dues condicions anteriors, podrem concloure que com més alt siga el valor de χ^2, mes intensa serà la relació entre les variables corresponents.

• S'han proposat altres estadístics a fi d'avaluar la intensitat de l'associació o grandària de l'efecte entre variables categòriques, els quals no depenen de la grandària de la mostra. Tots aqueixos es basen en l'índex χ^2 i es presenten a continuació. Aquests índexs poden utilitzar-se amb finalitat comparativa, encara que les dimensions de les taules comparades i la grandària de les mostres siguen diferents.

(2) L'índex ***phi*** **de Pearson** (φ):

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

• L'índex φ pot oscil·lar entre 0 i $\sqrt{q-1}$, sent q el nombre de modalitats de la variable que en tinga menys.

• En taules de contingència de 2 × 2 oscil·la l'índex oscil·la entre 0 i 1, per la qual cosa se sol emprar en aquesta circumstància. Les normes interpretatives suggerides per Cohen per a aquest índex són: $\varphi < 0{,}3$ => nivell baix d'associació; $0{,}3 \leq \varphi < 0{,}5$ => nivell mig d'associació; $\varphi \geq 0{,}5$ => nivell alt d'associació.

(3) L'índex ***V* de Cramer**:

$$V = \sqrt{\frac{\chi^2}{n(q-1)}} \qquad (q = \min[I, J])$$

• L'índex V de Cramer oscil·la entre 0 (independència) i 1, de manera que com més pròxim a 1, més intensa l'associació entre les variables. Aquest índex es pot interpretar també seguint les normes exposades abans per l'índex φ.

Exercici 4: Obtingueu els índexs φ i V de Cramer a partir de les tres taules de contingència presentades anteriorment per als tres col·legis i, també, per a l'exemple de les variables "Estat d'ànim" i "Viure en residència". Els valors de l'índex χ^2 són: 0 (col·legi 1), 0,673 (col·legi 2), 24,97 (col·legi 3) i 8,78 ("Estat d'ànim" i "Residència").

• Tot seguit es mostren els valors d'aquests índexs per a l'exemple sobre "Estat d'ànim" i "Residència" obtinguts amb SPSS:

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	8,784[a]	2	,012
Razón de verosimilitudes	8,507	2	,014
Asociación lineal por lineal	7,788	1	,005
N de casos válidos	500		

a. 0 casillas (0,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 35,40.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,133	,012
	V de Cramer	,133	,012
N de casos válidos		500	

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

2.2. El cas d'una variable categòrica i una quantitativa

• De nou, l'anàlisi d'aquesta mena d'associació suposa comparar les distribucions d'una variable agrupada en funció dels valors que en pren l'altra. Normalment, se sol prendre com a condicionada la quantitativa i com a condicionant la categòrica, si bé les conclusions a les quals arribaríem serien les mateixes si es fera a l'inrevés. Si no hi ha diferències entre les distribucions condicionades, això indicarà que no hi ha associació entre totes dues variables.

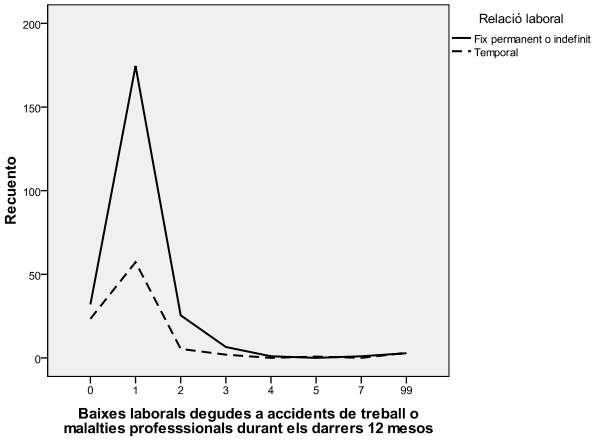
Exemple del cas en què es vulga analitzar l'associació entre les variables "Nota en un examen d'una assignatura [0 a 10]" i "Grup en el qual s'està matriculat [1 a 6]", disposant-se de les dades d'un total de 768 estudiants de 6 grups:

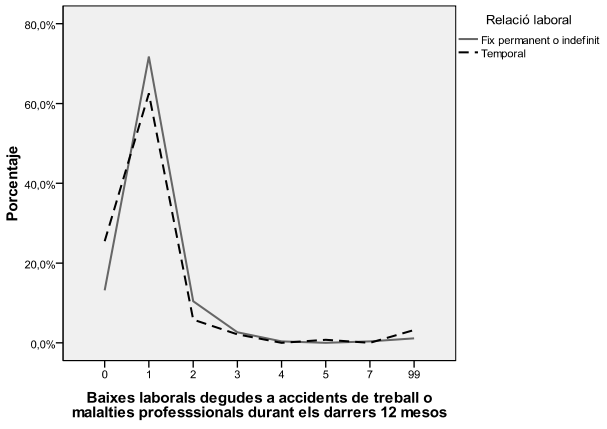
Grup 1		-----	**Grup 2**		-----	**Grup 3**		-----	**Grup 4**		-----	**Grup 5**		-----	**Grup 6**	
xi	ni		xi	ni		xi	ni		xi	ni		xi	ni		xi	ni
0	2		0	3		1,8	1		…	…		…	…		…	…
,3	3		1,3	1		2,0	7									
,5	1		1,8	1		2,3	3									
,8	3		2,0	2		2,5	2									
1,0	1		2,3	4		2,6	3									
1,3	1		2,5	2		2,8	2									
1,5	2		2,8	5		2,9	1									
1,8	2		2,9	1		3,0	2									
2,0	2		3,0	4		3,3	3									
2,3	3		3,3	3		3,5	7									
2,5	2		3,5	6		3,8	9									
2,6	1		3,8	3		3,9	1									
2,8	6		3,9	1		4,0	4									
3,0	5		4,0	2		4,1	1									
3,3	3		4,3	5		4,3	3									
3,5	5		4,5	6		4,5	9									
3,8	7		4,6	1		4,7	6									
4,0	8		4,7	8		4,8	7									
4,3	7		4,8	6		4,9	4									
4,5	5		4,9	6		5,0	3									
4,7	4		5,0	5		5,3	4									
4,8	3		5,1	1		5,5	4									
4,9	5		5,3	5		5,6	1									
5,0	3		5,4	1		5,8	1									
5,1	1		5,5	6		5,9	2									
5,3	6		5,6	1		6,0	4									
5,5	4		5,8	9		6,3	3									
5,8	7		5,9	1		6,5	6									
6,0	5		6,0	5		6,8	2									
6,1	1		6,3	2		7,0	4									
6,3	5		6,5	4		7,1	1									
6,5	3		6,8	4		7,3	2									
6,8	2		6,9	1		7,5	5									
7,0	1		7,0	2		8,0	1									
7,5	3		7,3	4		8,3	2									
7,6	1		7,5	3		8,4	1									
8,0	2		7,8	2		9,0	4									
9,0	2		8,0	2		9,3	1									
			8,1	1		9,5	1									
			8,3	1												

• Tal com ja aquest exemple evidencia, pot resultar bastant dificultós comparar les distribucions condicionades d'una variable quantitativa en funció dels valors d'una variable categòrica. Per a fer-ho, es pot recórrer a alguna de les representacions gràfiques com les que a continuació es descriuen.

(1) El polígon de freqüències agrupat, un tipus de gràfic que ja es va presentar en el capítol anterior, representa una bona opció si tenim en compte un detall important: si la grandària dels subgrups definits per la variable condicionant no és el mateix, o bastant similar, és convenient representar aquest gràfic amb proporcions o percentatges condicionats, a fi que els polígons puguen comparar-se adequadament.

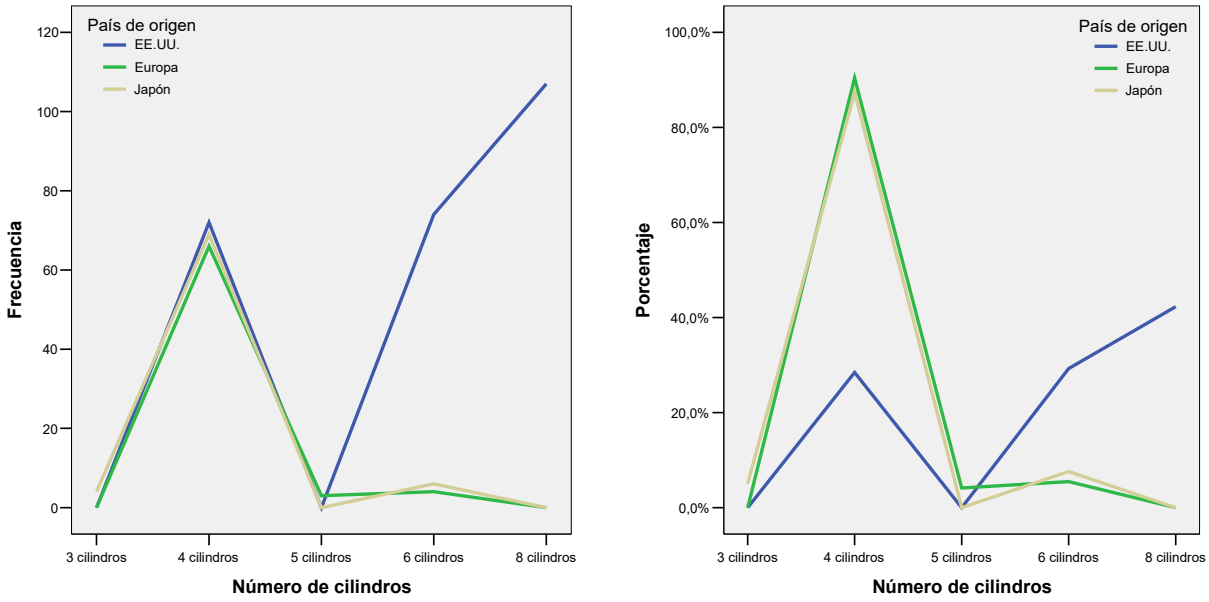
Exemple de polígon de freqüències agrupat per a la variable "Nombre de baixes laborals (durant els últims 12 mesos)" agrupada en funció de la "Relació laboral" dels treballadors [contracte fix; contracte temporal]. Cal recordar que quan la grandària dels grups és desigual, no s'han de representar les freqüències absolutes sinó freqüències relatives o percentatges condicionats, és a dir, dividint la freqüència absoluta per la grandària de cadascun dels grups. Vegeu en aquest exemple que el gràfic de l'esquerra (amb freqüències absolutes) pot resultar enganyós en donar la sensació que totes dues distribucions són bastant diferents, però, aquest efecte és resultat del fet que el nombre de treballadors fixos és bastant superior al de treballadors temporals. En el gràfic de la dreta, on es representen les distribucions de percentatges condicionats, es pot comprovar que totes dues distribucions són, en realitat, bastant similars, cosa que posa de manifest la absència de relació entre totes dues variables.



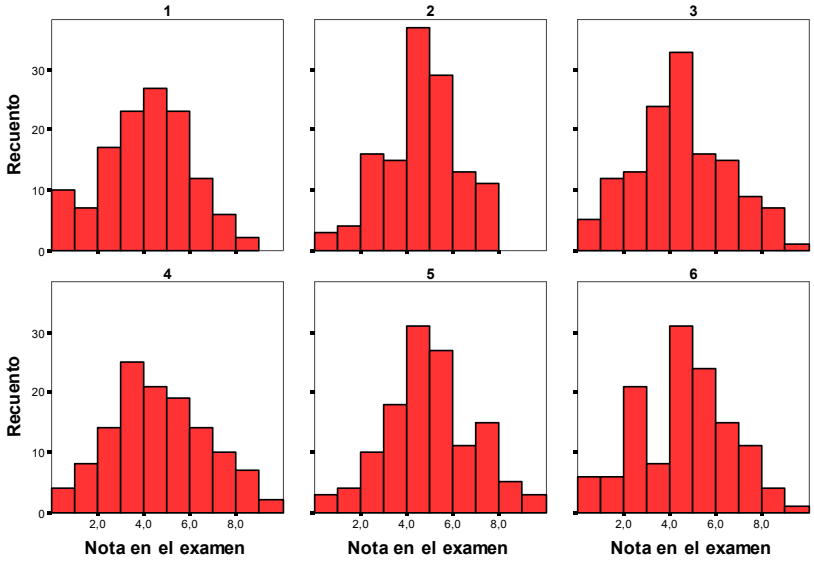


Un altre **exemple** en el qual s'aprecia aquest fet és el següent, amb les dades d'un estudi que es va fer als EUA sobre les característiques dels diferents models de cotxes existents en el mercat. En concret, a continuació es mostra la informació corresponent a la distribució conjunta de freqüències de les variables "Nombre de cilindres" i "País d'origen" per a una mostra de 405 vehicles, així com els corresponents polígons de freqüències agrupats obtinguts tant amb freqüències absolutes com amb percentatges condicionats:

		País de origen			
		EE.UU.	Europa	Japón	Total
Número de cilindros	3 cilindros			4	4
	4 cilindros	72	66	69	207
	5 cilindros		3		3
	6 cilindros	74	4	6	84
	8 cilindros	107			107
Total		253	73	79	405

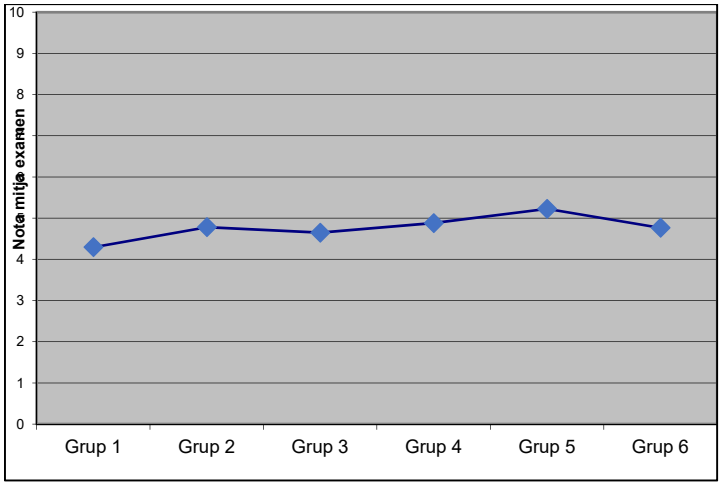


(2) El panell d'histogrames mostra un histograma de la variable quantitativa per a cadascun dels subgrups definits per la variable categòrica. A continuació es mostra un **exemple** d'aquest tipus de representació per a les dades de les variables "Nota en un examen d'una assignatura [0 a 10]" i "Grup en el qual s'està matriculat [1 a 6]" presentats al principi d'aquesta secció ($n = 768$). En aquest exemple no s'ha optat per representar les freqüències relatives o els percentatges condicionats perquè els sis grups eren molt similars en la seua grandària.



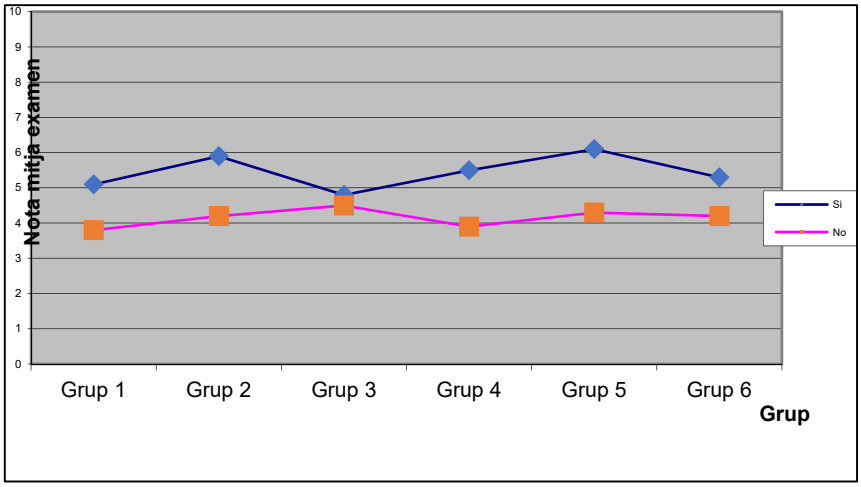
• Mentre que el polígon de freqüències agrupat i el panell d'histogrames representen la distribució de freqüències de la variable quantitativa per a cada modalitat de la variable categòrica, el que representen els gràfics que apareixen a continuació són determinats estadístics que resumeixen les característiques d'eixes distribucions de freqüències condicionals.

(3) El gràfic de mitjanes: **exemple** per a la variable "Nota en un examen d'una assignatura" agrupada en funció de la variable "Grup en el qual s'està matriculat [1 a 6]":

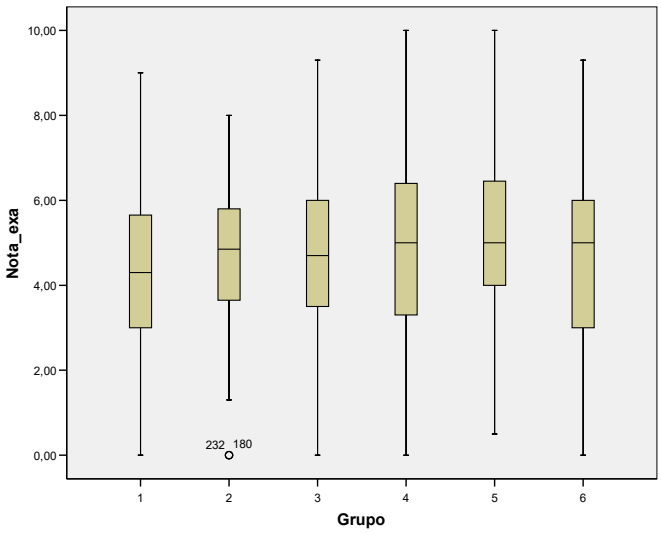


• L'agregació de les dades originals en forma de mitjanes fa factible incloure en aquesta representació gràfica la informació d'una variable categòrica addicional, la qual cosa ens permetrà presentar la informació corresponent a 3 variables.

Exemple de gràfic de mitjanes de la variable "Nota en un examen d'una assignatura" agrupada en funció de les variables "Grup" [1 a 6] i "Assistència regular a les classes" [Sí, No]:

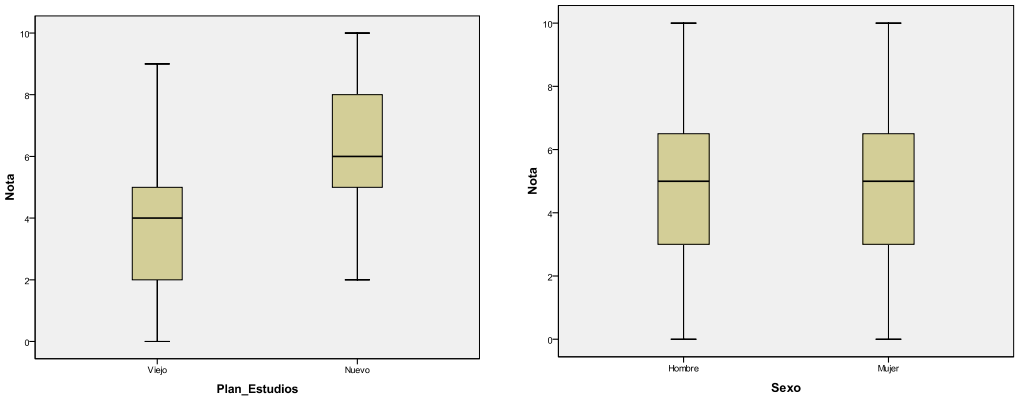


(4) El gràfic de caixa i bigots agrupat pot ser especialment apropiat si la variable categòrica té un nombre ampli de modalitats, perquè resulta fàcil encaixar nombroses caixes en el mateix espai gràfic. A continuació es mostra un **exemple** per a les mateixes variables representades en el panell d'histogrames i el gràfic de mitjanes presentats anteriorment.



• Les quatre representacions gràfiques presentades ens permetran comparar el grau de coincidència de les distribucions condicionades. En general, com més gran siga aqueixa coincidència, menor serà la intensitat de l'associació entre les dues variables i, viceversa, com més gran siga la discrepància, major serà la grandària de l'efecte de la relació. Així, en els exemples anteriors en què apareixien representades les variables "Grup" i "Nota" s'observa bastant coincidència entre les 6 distribucions condicionals, la qual cosa posa de manifest una baixa relació entre totes dues variables.

Exemple de diferent intensitat en l'associació entre dues variables (en el gràfic de l'esquerra la relació entre la "Nota" i el "Pla d'estudis" i en el gràfic de la dreta la relació entre la "Nota" i el "Sexe"). La conclusió quan es comparen els dos gràfics és que hi ha una relació més baixa entre les variables "Nota" i "Sexe" (major coincidència de les distribucions) que entre les variables "Nota" i "Pla d'estudis" (menor coincidència).



2.2.1 Índexs estadístics orientats a quantificar l'associació entre una variable categòrica i una variable quantitativa

• Per tal de captar les diferències existents entre les distribucions condicionades de la variable quantitativa per a cadascun dels valors de la variable categòrica, la majoria dels índexs estadístics que han sigut proposats s'han centrat en comparar un aspecte específic d'aqueixes distribucions: la seua tendència central i, més comunament, les seues mitjanes aritmètiques.

• Els estadístics que a continuació es presenten estan basats en les diferències entre les mitjanes en els subgrups definits per la variable categòrica.

(1) L'índex d'associació ***d* de Cohen** és apropiat quan es tinga una variable quantitativa *I* i una variable categòrica *X* dicotòmica [a, b], i es calcula com:

$$d = \frac{\left|\overline{Y}_a - \overline{Y}_b\right|}{s_Y}$$

on el numerador és la diferència entre les mitjanes de la variable quantitativa en ambdós grups i el denominador és la desviació típica o estàndard de la variable quantitativa per al conjunt de les dades. L'índex *d* de Cohen és una diferència de mitjanes tipificada, per la qual cosa el seu valor pot oscil·lar entre 0 (si les variables són independents) i un valor que, encara que no té a priori límit màxim, com és el cas de les puntuacions típiques (*z*), és infreqüent que adopte valors per damunt de 2. Cohen (1992) va suggerir les següents normes interpretatives, encara que el mateix autor va remarcar que s'han d'utilitzar només en el cas que no es tinga cap criteri substantiu que permeta realitzar la interpretació: valors absoluts de *d* al voltant de 0,2 indicarien una intensitat de l'associació (grandària de l'efecte) baixa; al voltant de 0,5, mitjana; i al voltant de 0,8 o superior, alta.

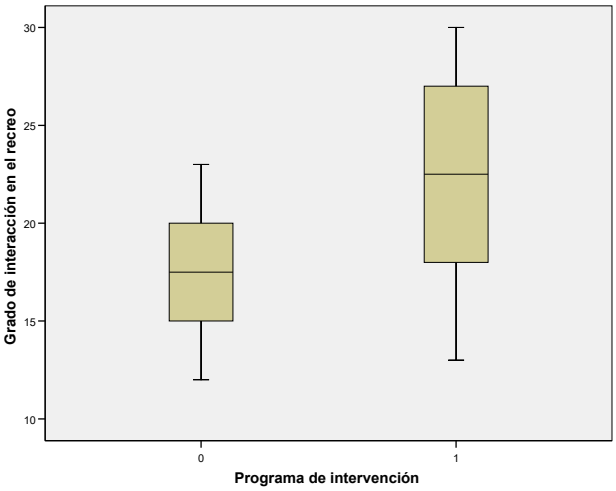
(2) L'índex ***f* de Cohen** permet analitzar la relació entre una variable quantitativa (*I*) i una categòrica (*X*) en cas que aquesta última tinga més de dos valors possibles (*k* valors). Es fonamenta en el càlcul de la dispersió de les mitjanes dels diferents subgrups definits pels *k* valors de la variable *X*:

$$f = \frac{s_{\overline{Y}}}{s_Y}, \qquad \text{on } s_{\overline{Y}} = \sqrt{\frac{\sum_{i=1}^{k} n_i \cdot (\overline{Y}_i - \overline{Y})^2}{n}}$$

• Si les mitjanes dels subgrups són iguals o molt pròximes, la desviació típica de les mitjanes serà igual o pràcticament igual a 0, la qual cosa indica l'absència d'associació entre ambdues variables. El valor de la *f* de Cohen serà sempre major o igual a 0, més gran com més intensa siga l'associació entre les variables.

Exercici 5: Siguen les variables *X* ("Aplicació d'un programa d'intervenció per a afavorir la interacció social" [Sí (1), No (0)]) i *Y* ("Grau d'interacció en l'hora d'esbarjo", mesurat pel nombre de minuts en què s'ha participat en activitats amb altres alumnes). Tenim dades recollides en un grup de 20 alumnes d'una classe en la qual es va avaluar l'eficàcia del citat programa d'intervenció. Calculeu i interpreteu l'índex *d* de Cohen (la desviació estàndard de la variable *Y* és igual a 5,09).

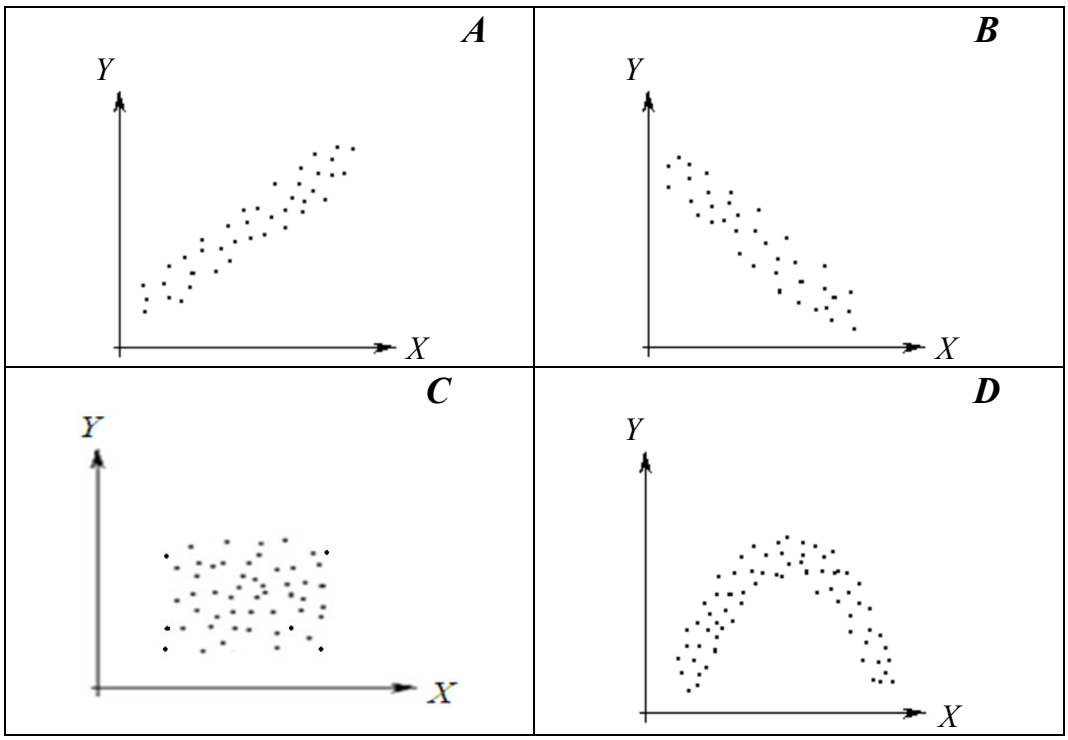
ÍD	X	Y
1	1	22
2	1	13
3	0	12
4	1	27
5	1	19
6	0	16
7	0	20
8	0	12
9	1	23
10	0	17
11	1	29
12	1	16
13	1	30
14	0	20
15	0	15
16	1	24
17	0	23
18	0	18
19	0	20
20	1	18



2.3. El cas de dues variables quantitatives

• Igual que en els casos anteriors, l'existència de correlació o associació entre dues variables quantitatives ve determinada per la presència de diferències en les distribucions condicionades d'una variable per als diferents valors de l'altra.

• No obstant això, com que el nombre de distribucions condicionals que es poden arribar a obtenir en aquest cas és molt ampli, el més habitual és analitzar l'associació directament sobre un diagrama de dispersió i observar la disposició del núvol de punts que representa la distribució conjunta de totes dues variables. Així, què podríem dir sobre l'associació entre les variables en els quatre diagrames de dispersió que es mostren a continuació?



• Un aspecte rellevant de l'anàlisi de la correlació entre dues variables quantitatives és que la presència d'aquesta es pot plantejar d'acord amb diferents models o patrons d'associació: per exemple, en forma de línia recta, tal com en els exemples *A (*relació lineal directa o positiva) i *B* (relació lineal inversa o negativa) de la figura superior, o en forma curvilínia tal com en *D* (relació parabòlica o quadràtica). Així, la manera d'avaluar la intensitat de la correlació és analitzar l'ajust del núvol de punts al model d'associació que es considere que representa més adequadament la distribució conjunta de totes dues variables.

2.3.1. Índexs estadístics orientats a quantificar l'associació entre dues variables quantitatives

• En la quantificació de l'associació entre 2 variables quantitatives ens limitarem al supòsit que un model de relació lineal representa adequadament l'associació entre ambdues variables. Cal ressaltar que, amb freqüència, s'obvia en els textos estadístics que la relació que s'analitza és en realitat una relació de tipus lineal. Els índexs més utilitzats en la pràctica estadística per tal d'analitzar la intensitat o grandària de l'efecte de la relació lineal entre dues variables quantitatives són els tres següents:

(1) La covariància (S_{XY} *o* σ_{XY}):

$$S_{XY} = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n}$$

• El numerador d'aquesta expressió es coneix en la literatura estadística com a suma de productes creuats (SP_{XY}), per la qual cosa l'anterior formula es pot expressar així: $s_{XY} = SP_{XY} / n$

• La covariància pot tenir valors tant positius com negatius. A l'hora d'interpretar-ho, un major valor de la covariància, en valor absolut, indicarà una relació lineal més intensa entre les dues variables. Un valor positiu indica una relació lineal directa, un de negatiu una relació lineal inversa, i si el valor és igual o molt pròxim a 0, la inexistència de relació lineal entre les dues variables.

(2) El coeficient de correlació producte-moment de Pearson (R_{XY})

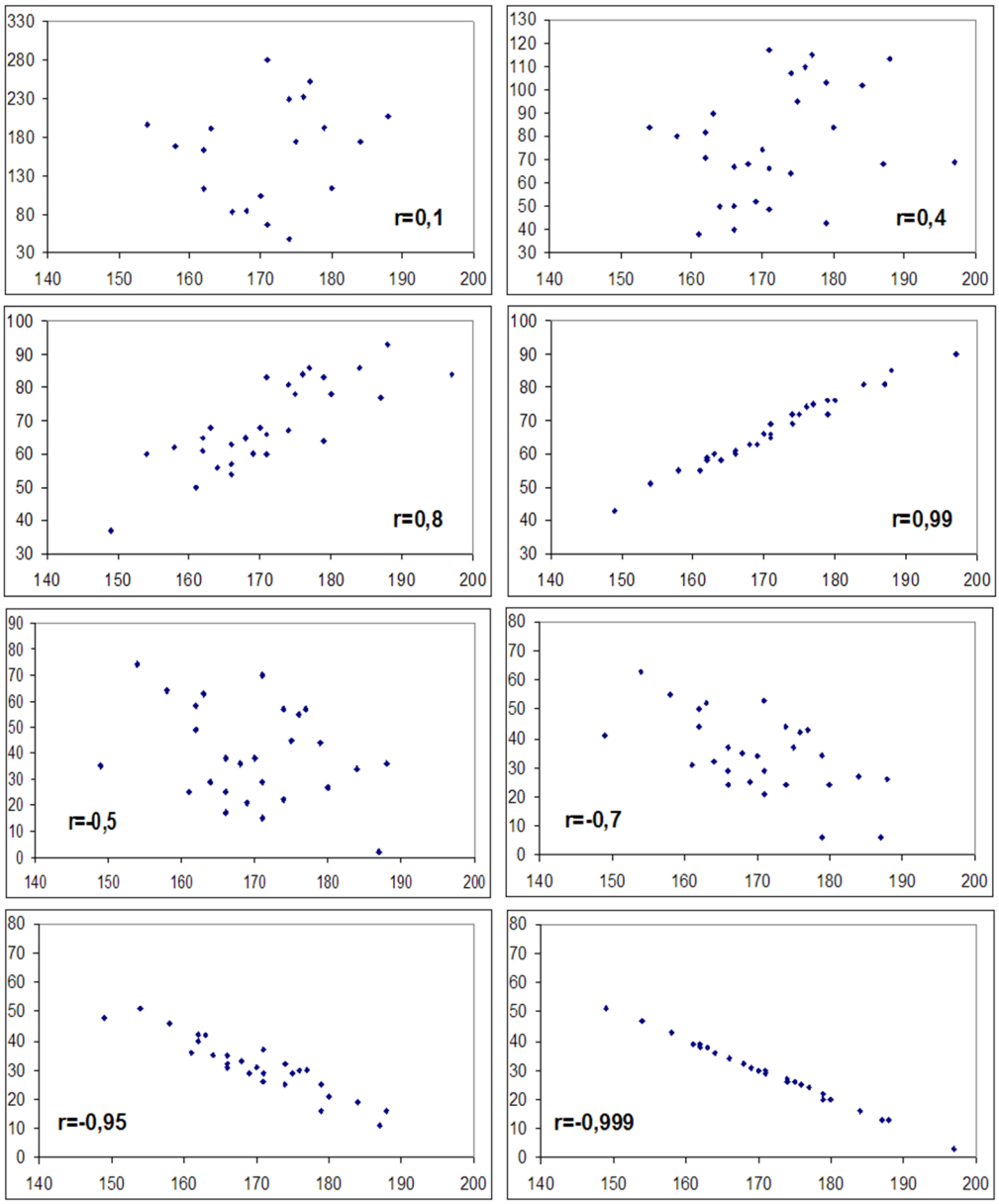
• Els inconvenients de la covariància –d'una banda, no té valors màxim i mínim i, d'altra banda, depèn de les unitats de mesura de les variables– es resolen estandarditzant aquest índex dividint el seu valor pel producte de les desviacions típiques de totes dues variables. S'obté així el conegut com a coeficient de correlació producte-moment de Pearson:

$$R_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

• El coeficient de correlació de Pearson s'interpreta de manera anàloga a la covariància, però, com que oscil·la entre -1 i 1, la interpretació d'aquest és més intuïtiva, alhora que facilita l'establiment de comparacions entre els coeficients obtinguts per a conjunts de dades diferents.

• En cas que la desviació típica d'una de les dues variables fora igual a 0, la fórmula de R_{XY} resultaria en una indeterminació; ara bé, això ocorrerà si tots els valors d'aqueixa variable foren iguals (cas en el qual tampoc es pot parlar pròpiament d'una variable).

Exemples del valor de R_{XY} obtingut per a diferents conjunts de dades (Barón-López, 2005):



• La matriu de correlacions constitueix un tipus de representació en forma de taula que permet mostrar l'associació existent entre un conjunt de variables per parelles. Les variables es presenten en les files i les columnes de la taula, i cada casella de la taula mostra el valor de la correlació entre la variable de la fila i la columna corresponents. Cal tenir en compte:

(1) Com que es tracta d'una matriu simètrica, alguns paquets estadístics només presenten una de les dues meitats de la matriu.

(2) En la diagonal de la matriu apareix el valor "1", atès que aquestes cel·les representen la correlació d'una variable amb si mateixa.

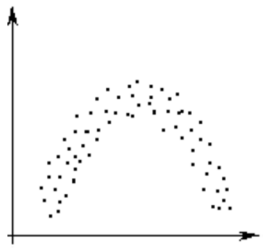
(3) Una matriu de correlacions podria construir-se amb variables de qualsevol tipus; no obstant això, en la literatura apareix només per a variables quantitatives.

Exemple de matriu de correlacions a partir del rendiment d'un grup de xiquets en 5 matèries: música (A), matemàtiques (B), llenguatge (C), esport (D) i ciències naturals (E).

	A	*B*	*C*	*D*	*E*
A	1				
B	0,23	1			
C	0,36	0,24	1		
D	−0,45	−0,34	−0,29	1	
E	0,07	0,38	0,17	0,13	1

• Alguns comentaris respecte a la interpretació del valor de R_{XY} :

(a) Un valor de R_{XY} (i el mateix per a la covariància) nul o pròxim a 0 indica que no existeix relació lineal entre totes dues variables, la qual cosa no significa que no puga existir algun altre tipus de patró de relació entre elles. (=> important primer visualitzar gràficament la relació).



=> $R_{XY} \approx 0$

(b) La intensitat de la correlació entre dues variables pot ser valorada seguint diferents esquemes interpretatius: per exemple, alguns autors consideren que un valor absolut de R_{XY} superior a 0,5 ha de ser ja considerat com a alt. Al contrari, altres autors critiquen aquesta pràctica i defensen que, per a valorar un coeficient de correlació, s'ha de tenir en compte el context i la informació ja existent relativa a la relació entre aqueixes dues variables.

(c) La presència de correlació entre dues variables no implica que existisca una relació de causalitat entre ambdues, per molt alt que siga l'índex d'associació. Tal interpretació podria ser encertada en alguns casos però, en molts altres, pot representar un error greu. L'existència de correlació és condició necessària, però no suficient, per a establir una relació de causa-efecte entre dues variables. S'han de satisfer altres condicions que estan associades als dissenys d'investigació experimental. El tema dels tipus de disseny d'investigació ja va ser introduït en el primer capítol d'aquest temari, si bé es farà un tractament més en profunditat d'aquest tema en l'assignatura de Dissenys d'Investigació.
(Aquest comentari es fa extensiu a tots els índexs d'associació tractats en aquest tema).

Exemple: Si s'observa una relació positiva entre l'alçada i el sou seria un error afirmar sense més que com més alta siga una persona, major serà el seu sou. En realitat, hi ha una variable de confusió, el sexe –atès que el sexe està relacionat tant amb l'alçada com amb el sou. En conseqüència, si el sexe no es té en compte en l'anàlisi s'obtindrà una correlació espúria entre l'alçada i el sou.

(3) El coeficient de determinació (R^2_{XY}):

• El coeficient de determinació és el quadrat del coeficient de correlació de Pearson i, per tant, oscil·la entre 0 (independència entre les variables) i 1 (relació lineal perfecta).

• Aquest índex, a part d'utilitzar-se en el context de la regressió lineal que es tractarà en una tema posterior, és també el més apropiat per a comparar la relació lineal existent entre dues parelles (o més) de variables (o, també, en una única variable mesurada en dos moments temporals o en dos grups de subjectes diferents). D'altra banda, és inadequat per raons teòriques inherents al coeficient de correlació de Pearson dir, per exemple, que la intensitat de l'associació entre *X* i *Y* és el doble que entre *M* i *N* si s'han obtingut per a totes dues parelles de variables un $R_{XY} = 0{,}8$ i un $R_{XY} = 0{,}4$, respectivament. No obstant això, sí que és possible tal interpretació a partir dels coeficients de determinació, per exemple, si foren $R^2_{AB} = 0{,}32$ i $R^2_{CD} = 0{,}16$.

Exercici 6: Es va calcular el coeficient de correlació entre les puntuacions en dos tests *X* i *Y* en dues mostres de subjectes pertanyents a dos països *A* i *B*. Per a la mostra *A* es va obtenir un $R_{XY} = 0{,}3$ mentre que per a la mostra *B*, un $R_{XY} = 0{,}6$. Què es pot dir en termes comparatius sobre l'associació entre *X* i *Y* en tots dos països?

Exercici 7: A partir de les següents dades, presentades en el tema anterior, procedents d'un grup de 16 subjectes sobre el nombre d'hores d'esport que practicaven setmanalment (*X*) i la percepció que tenien sobre el seu estat de salut general (*Y*) en una escala d'1 a 10, avalueu l'associació entre les dues variables tant gràficament (el diagrama de dispersió ja es va realitzar en el tema anterior) com analíticament a través dels índexs S_{XY}, R_{XY} i R^2_{XY}. (Es recomana l'obtenció amb un paquet estadístic, per exemple, SPSS.)

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8
I	4	3	3	5	6	4	4	6	5	2	7	9	6	8	9	8

Exercici 8: S'han obtingut amb SPSS els següents resultats en l'anàlisi de la relació entre les variables "Nombre d'anys d'escolarització" i "Puntuació prestigi professional (escala de 0 a 100)". Calculeu-ne el valor del coeficient de correlació de Pearson entre totes dues variables.

Correlaciones

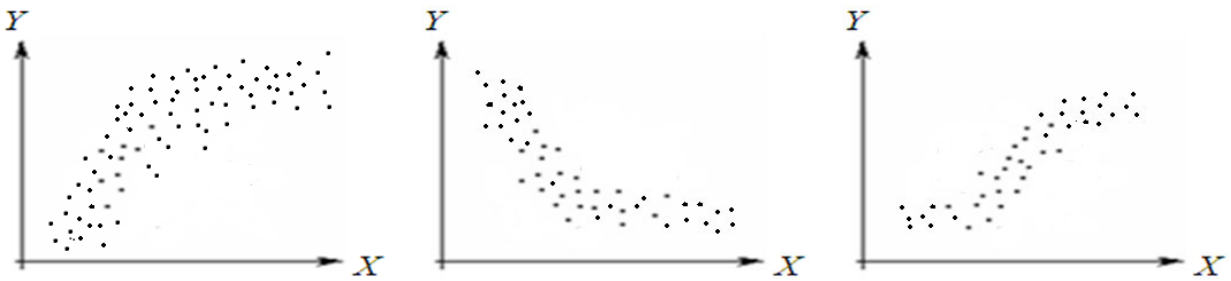
		Número de años de escolarización	Puntuación de prestigio profesional (1980)
Número de años de escolarización	Correlación de Pearson	1	¿?
	Sig. (bilateral)		,000
	Suma de cuadrados y productos cruzados	13436,719	28442,286
	Covarianza	8,904	20,115
	N	1510	1415
Puntuación de prestigio profesional (1980)	Correlación de Pearson	¿?	1
	Sig. (bilateral)	,000	
	Suma de cuadrados y productos cruzados	28442,286	241965,769
	Covarianza	20,115	170,759
	N	1415	1418

Estadísticos descriptivos

	Media	Desviación típica	N
Número de años de escolarización	12,88	2,984	1510
Puntuación de prestigio profesional (1980)	42,93	13,067	1418

(4) El coeficient de correlació de Spearman (R_s)

• Quan tinguem dues variables quantitatives la relació de les quals no es lineal –encara que sí monòtona, ja siga creixent o decreixent (vegeu-ne exemples gràfics a continuació)– és més adequat aplicar el coeficient de correlació de Spearman, el qual es basa a reconvertir els valors originals de les variables en valors d'ordre (al valor més baix de cada variable se li assigna un 1, al següent un 2, i així successivament).



• A continuació es mostra la fórmula per a calcular R_S, on D representa la diferència, per a cada subjecte, entre el seu valor d'ordre en una variable i en una altra. En qualsevol cas, no incidirem ací en la seua obtenció, perquè pot ser calculat fàcilment amb un paquet estadístic (e. g., SPSS). La interpretació de R_S és exactament la mateixa que la del coeficient de correlació de Pearson.

$$R_S = 1 - \frac{6 \cdot \sum D^2}{N \cdot \left(N^2 - 1\right)}$$

• L'obtenció del coeficient de correlació de Spearman resulta també recomanable en dues situacions addicionals: (1) quan tinguem variables quantitatives amb valors anòmals en una d'aquestes o en ambdues; (2) quan per a alguna de les variables, o per a ambdues, no es tinga clar que la seua escala de mesura siga quantitativa i es preferisca que siguen considerades com a variables ordinals.

> **Exemple** d'obtenció del R_S amb SPSS per a 6 parelles de variables resultants de combinar per parelles tres variables que representen, respectivament, la valoració de les relacions amb els companys, amb els professors i amb el personal d'administració i serveis (PAS), feta per una mostra de 174 estudiants universitaris del Grau de Psicologia. L'escala de valoració en les tres variables oscil·lava entre 0 (Gens satisfactòria) i 10 (Molt satisfactòria). L'analista va decidir no considerar l'escala de mesura d'aquestes variables com a quantitatives i, en conseqüència, va decidir aplicar el coeficient de correlació de Spearman a fi de valorar l'associació entre aqueixes variables.

Correlaciones

			Relación con compañeros	Relación con profesores	Relación con PAS
Rho de Spearman	Relación con compañeros	Coeficiente de correlación	1,000	,192	,070
		Sig. (bilateral)	.	,011	,358
		N	174	174	174
	Relación con profesores	Coeficiente de correlación	,192	1,000	,401
		Sig. (bilateral)	,011	.	,000
		N	174	174	174
	Relación con PAS	Coeficiente de correlación	,070	,401	1,000
		Sig. (bilateral)	,358	,000	.
		N	174	174	174

Referències

Barón-López, J. (2005). *Bioestadística: métodos y aplicaciones*. Apunts i material disponibles en http://www.bioestadistica.uma.es/baron/apuntes/

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Solanas, A., Salafranca, L., Fauquet, J. i Núñez, M. I. (2005). *Estadística descriptiva en Ciencias del Comportamiento*. Madrid: Thompson.

Tema 6 – Regressió: l'anàlisi de regressió lineal

1. Conceptes bàsics sobre l'anàlisi de regressió

• Models predictius o de regressió: la representació de la relació entre dues (o més) variables a través d'un model formal suposa poder comptar amb una expressió logicomatemàtica (i. e., una equació) que, a banda de resumir com és aqueixa relació, permetrà realitzar prediccions dels valors que prendrà una de les variables –la que s'assumisca com a variable de resposta– a partir dels valors que prenguen les altres variables.

• Pel que fa al paper que juguen les variables en el model, mentre que en l'anàlisi de la relació entre dues variables no s'assumia un rol específic per a aquestes (**rol simètric** de les variables –era el mateix el coeficient de correlació de la variable *A* amb la variable *B*, que el de *B* amb *A*), l'aplicació d'un model de regressió suposa que una de les dues variables adopta el paper de variable explicativa i l'altra el de variable de resposta i és, per tant, que es diu que les variables adopten un **rol asimètric**– no és el mateix el model de regressió de *B* sobre *A*, que el de *A* sobre *B*.

• En la literatura estadística s'han plantejat diferents tipus de models predictius que han donat resposta a les diferents característiques de les variables que hi poden aparèixer implicades, ja siga la seua escala de mesura, la forma de la seua distribució... El més conegut és el model de regressió lineal (variable de resposta quantitativa), si bé altres opcions a tenir en compte són el model de regressió logística (variable de resposta categòrica) o el model de Poisson (variable de resposta quantitativa amb distribució molt asimètrica), entre altres.

• El model de regressió lineal és el més utilitzat per a predir els valors d'una variable quantitativa a partir dels valors d'una altra variable explicativa també quantitativa (model de regressió lineal simple). Una generalització d'aquest model, el de regressió lineal múltiple, permet considerar més d'una variable explicativa quantitativa en el model.

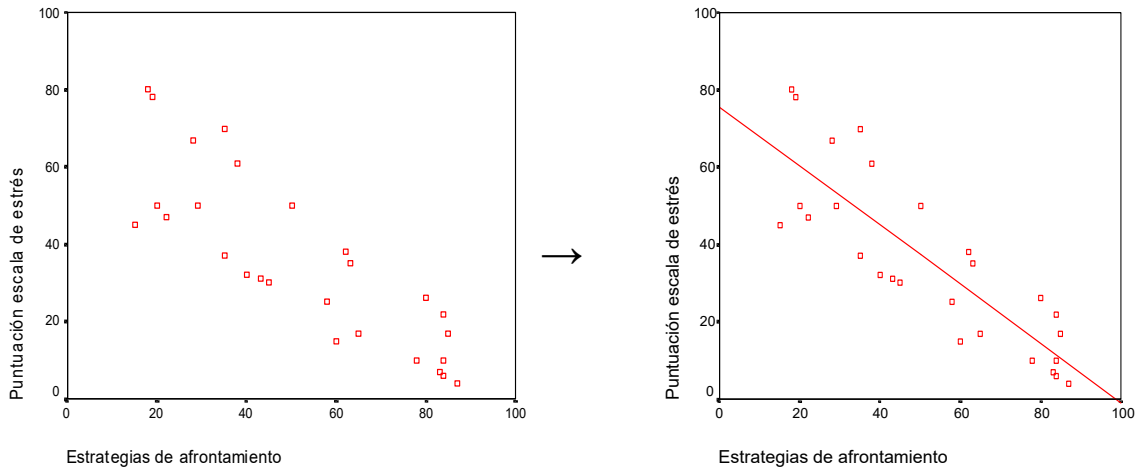
• En concret, segons el model de regressió lineal simple, la distribució conjunta de dues variables, una de les quals considerada com a variable explicativa (X) i l'altra com a variable de resposta (Y), la relació de la qual siga més o menys lineal (vegeu el diagrama de dispersió) pot ser representada (modelada) per l'equació d'una línia recta:

$$\hat{Y} = B_0 + B_1 \cdot X$$

Exemple d'aplicació d'un model de regressió lineal simple a fi de modelar la distribució conjunta de les variables "Estratègies d'afrontament" i "Estrès" (vegeu el diagrama corresponent de dispersió en el gràfic de baix a l'esquerra). En aquest exemple concret, el model de regressió lineal simple es concreta en l'ajustament a les dades de la següent equació (també coneguda com a recta de regressió):

$$\hat{Y} = 75,4 + (-0,76) \cdot X$$

La manera com han sigut obtinguts els coeficients d'aquesta equació (B_0 i B_1) serà tractada més endavant. En el gràfic de baix a la dreta s'ha dibuixat la recta corresponent a l'anterior equació.



Un important avantatge de conèixer la recta de regressió d'"Estrès" sobre "Afrontament" és que podem predir quina serà la puntuació en "Estrès" a partir d'un valor qualsevol d'Afrontament"; així, per exemple, per a una puntuació de 50 en Afrontament, la puntuació predita d'"Estrès" serà de 37,4 $(= 75,4 - 0,76 \cdot 50)$.

• Els dos coeficients de l'equació del model de regressió lineal simple, B_0 i B_1, són coneguts com la constant i el pendent del model, respectivament. En conjunt reben el nom de coeficients de l'equació de regressió. Si l'equació de la recta de regressió és obtinguda a partir d'una mostra, que no una població, els coeficients de l'equació de regressió que obtinguem seran estadístics, no paràmetres, i l'equació s'expressa simbòlicament com:

$$\hat{Y} = B_0 + B_1 \cdot X_1$$

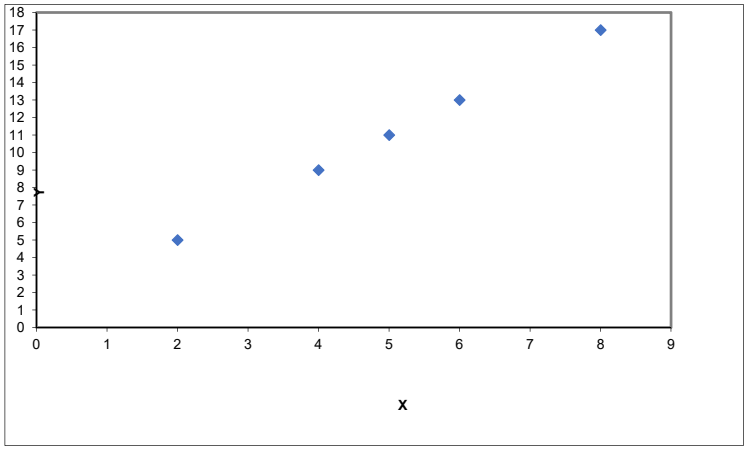
Cal assenyalar també que l'equació de la recta de regressió apareix expressada en alguns llibres de text així:

$$\hat{Y} = A + B \cdot X$$

• Una vegada que es coneguen els valors de B_0 i B_1 del model de regressió lineal simple, aquest pot ser utilitzat com a model predictiu, això és, per a realitzar prediccions dels valors que prendrà la variable de resposta per a determinats valors de la variable explicativa. Només caldrà substituir en l'equació de regressió el valor concret de X que es vulga (X_i). I s'obtindrà així el valor predit per a Y per a aquells casos que en la variable X prenguen el valor X_i. Aquest valor és conegut de manera genèrica com a puntuació predita, i es representa simbòlicament com Y_i' o $\hat{Y}_i$.

Exercici 1: A partir de la distribució conjunta de les variables quantitatives X i Y, dibuixeu en el diagrama corresponent de dispersió la recta de regressió que millor s'ajusta al núvol de punts. Encara que no s'haja vist com obtenir els valors de B_0 i B_1 de la recta de regressió, intenteu deduir de manera intuïtiva aquests valors. Si heu plantejat l'equació, utilitzeu-la per obtenir els valors predits en Y per a diferents valors de X (per exemple, per a $X_i = 3$, per a $X_i = 6$, per a $X_i = 9$…). També podeu realitzar aquestes prediccions utilitzant el gràfic inferior a partir de la recta de regressió que heu dibuixat.

X	Y
2	5
4	9
5	11
6	13
8	17



• Relacions deterministes vs. probabilístiques i error de predicció: l'anterior exemple representa el cas d'una relació determinista o perfecta entre X i Y –si es calcula R_{XY}, serà igual a 1. En conseqüència, els valors predits $\hat{Y}$ a partir de X segons el model de regressió coincidiran exactament amb els valors observats en Y i no hi haurà cap error de predicció. No obstant això, aquesta situació és inusual en l'àmbit de les ciències socials i les ciències de la salut, on quasi sempre ens trobem amb relacions entre variables no perfectes ($R_{XY} \neq 1$ o $R_{XY} \neq -1$). En aquests casos, quan s'utilitza l'equació de la recta de regressió per a predir quin serà el valor en Y a partir d'un determinat valor X_i, és més que probable que se cometa un error en la predicció realitzada. Aquest error rep el nom d'error de predicció o residual (E_i) i queda definit, per tant, com la diferència entre el vertader valor d'un subjecte en la variable I (Y_i) i el seu valor predit segons l'equació de regressió ($\hat{Y}_i$):

$$E_i = Y_i - \hat{Y}_i$$

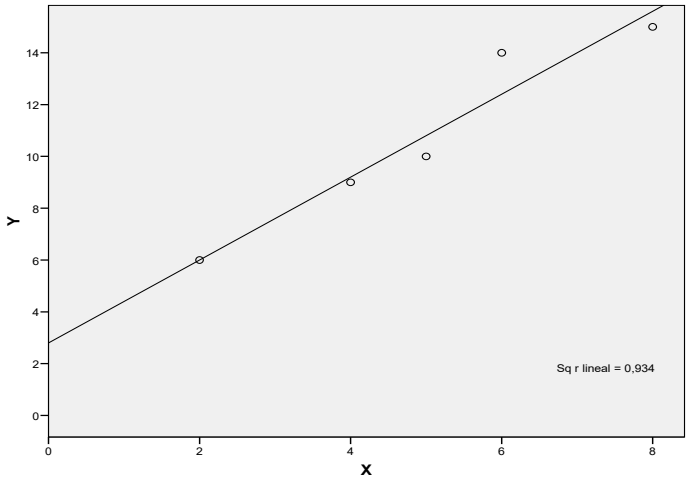
De l'expressió anterior es dedueix que la puntuació observada d'un subjecte en Y es pot obtenir sumant la puntuació predita, l'error de predicció o residual per a aquesta puntuació, això és:

$$Y_i = \hat{Y}_i + E_i$$

Exemple dels conceptes presentats per a dues variables X i Y ($N = 5$), en què el model de regressió lineal estimat per a la distribució conjunta de totes dues variables és el següent:

$$\hat{Y} = 2{,}8 + 1{,}6 \cdot X$$

ID	*X*	*Y*
1	2	6
2	4	9
3	5	10
4	6	14
5	8	15

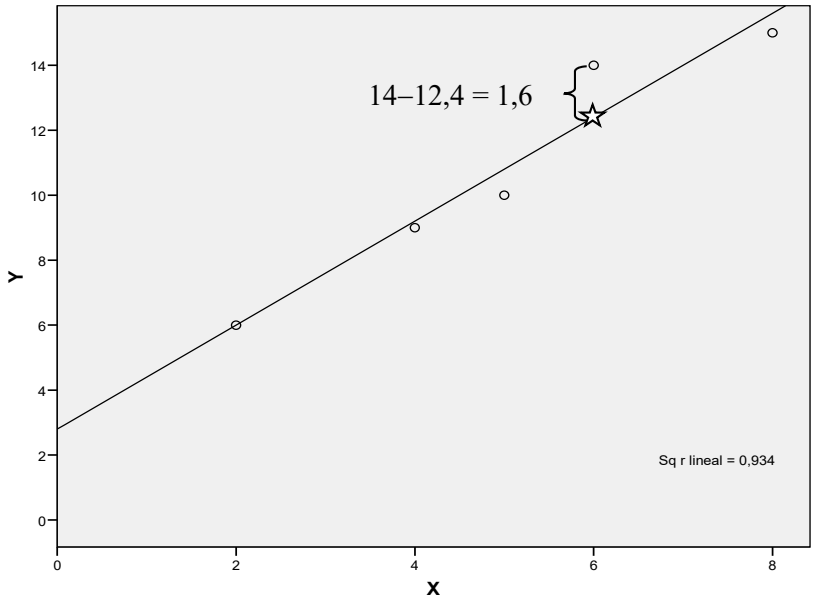


Si emprem l'equació de regressió ajustada a les dades, quin error cometem en predir Y a partir de X per a cadascun dels 5 casos? Per exemple, per al cas 4 en la taula ($X_4 = 6$), el valor predit és 12,4 ($\hat{Y}_4 = 2{,}8 + 1{,}6 \cdot 6 = 12{,}4$) i, en conseqüència, el seu error de predicció o residual és 1,6 ($E_4 = 14 - 12{,}4$), això és, la diferència entre el seu vertader valor en la variable Y ($Y_4 = 14$) i el seu valor predit segons l'equació de regressió ($\hat{Y}_4 = 12{,}4$). Així, per a tots els casos tenim:

ID	*X*	*Y*	$\hat{Y}$	*E*
1	2	6	6,0	0
2	4	9	9,2	-0,2
3	5	10	10,8	-0,8
4	6	14	12,4	1,6
5	8	15	15,6	-0,6

Cal avançar ja que la columna dels errors de predicció constitueix un element d'informació clau per a tractar el concepte de bondat d'ajustament d'un model de regressió, la qual cosa s'abordarà en una secció posterior.

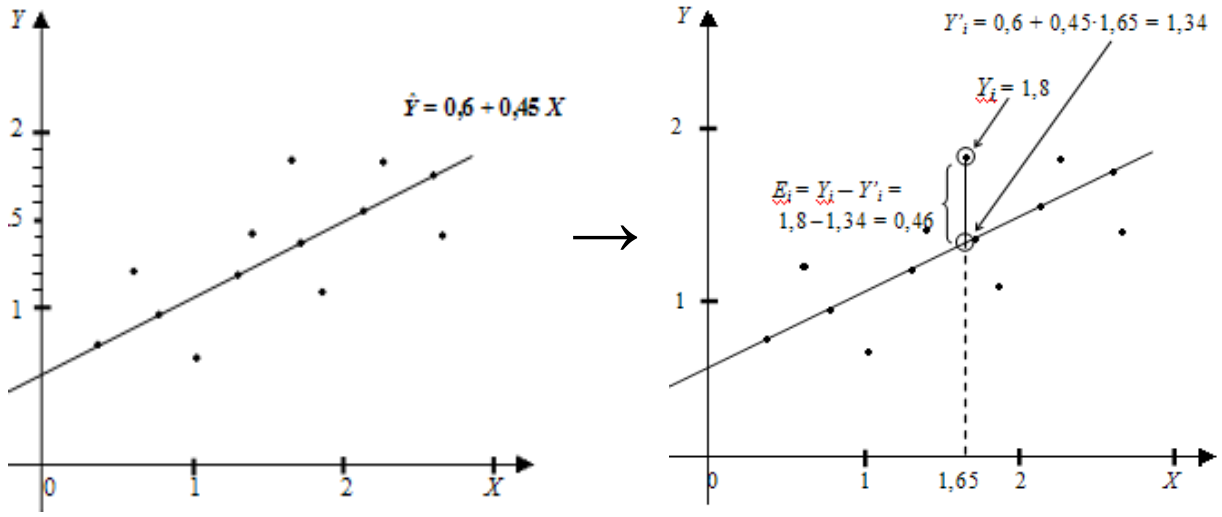
• Gràficament, el residual corresponent a qualsevol cas (punts en el diagrama de dispersió) és representat per la distància vertical del punt corresponent a la recta de regressió, tal com es mostra a baix per al cas 4° del exemple anterior.



Un altre **exemple** (Losilla i cols., 2005) per al cas de dues variables *X* e *Y* el diagrama de dispersió de les quals es mostra a continuació (gràfic de l'esquerra). El model de regressió lineal obtingut per a la distribució conjunta de totes dues variables ve definit per la següent equació:

$$\hat{Y} = 0{,}6 + 0{,}45 \cdot X$$

En el gràfic de la dreta es mostra l'error de predicció, d'acord amb el model de regressió lineal ajustat, per al cas amb puntuacions en *X* i *Y* iguals a 1,65 i 1,8, respectivament. Com pot observar-se, la puntuació predita en *Y* per als subjectes que en *X* tenen un valor de 1,65 és igual a 1,34, per la qual cosa l'error de predicció resultant és igual a 0,46.

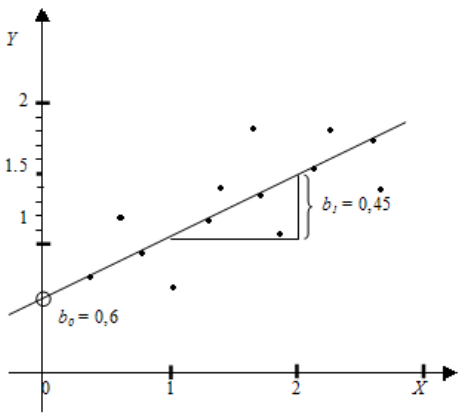


• Interpretació de B_0 i B_1

– La constant de l'equació de la recta de regressió (B_0) representa el valor predit en Y quan la variable X és igual a 0. Es tracta d'un valor que no comporta especial interés a nivell interpretatiu.

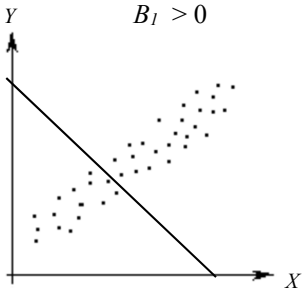
– El valor del pendent (B_1) representa el canvi esperat en la variable Y per cada unitat d'increment en la variable X. En aquest sentit, B_1 representa un indicador de la rellevància de l'efecte que els canvis en X tenen sobre Y. Cal assenyalar que, com que representa l'increment en $\hat{Y}$ per cada increment de X en una unitat, el valor del pendent estarà expressat en les mateixes unitats que la variable de resposta Y.

Exemple per al cas de 2 variables, X i Y, en què l'equació de regressió de Y sobre X és la següent: $\hat{Y} = 0{,}6 + 0{,}45 \cdot X$. Tal com pot observar-se en el diagrama de dispersió, quan s'augmenta en una unitat el valor de X, es produeix un augment de 0,45 unitats en el valor de Y predit per la recta de regressió.

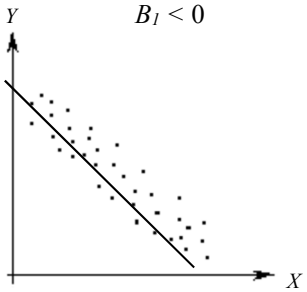


• Valors que pot tenir B_1 : Pot tenir valors tant positius com negatius, depenent del sentit (directe o invers) de la relació entre les variables. El seu valor, en valor absolut, serà més alt com més gran siga la relació lineal entre les variables.

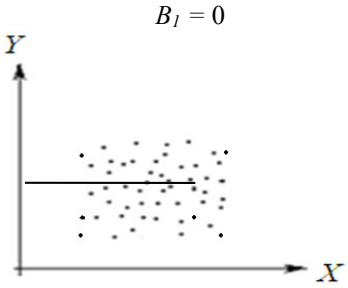
A continuació es mostren quatre **exemples** que mostren el vincle directe entre el valor de B_1 i el tipus de relació existent entre les variables. En l'exemple 1 la relació entre X e Y és directa, per la qual cosa el valor del pendent serà de signe positiu. En l'exemple 2 la relació entre X e Y és inversa, per tant, B_1 serà menor de 0. Les figures 3 i 4 evidencien la no existència de relació lineal entre les variables i, en aquest cas, el valor de B_1 serà nul o pràcticament nul.



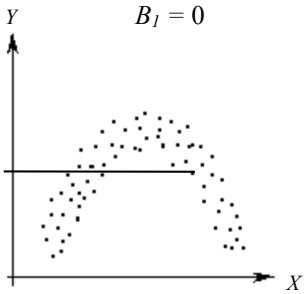
Exemple 1. Relació lineal directa



Exemple 2. Relació lineal inversa



Exemple 3. Absència de relació



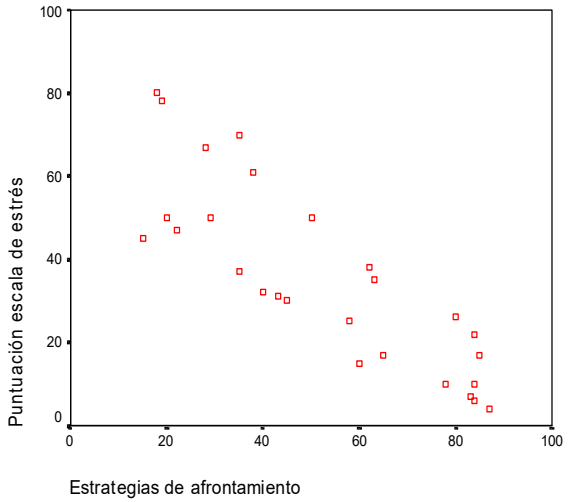
Exemple 4. Relació no lineal

• A continuació es presenten les dades d'un estudi l'objectiu del qual va ser investigar l'efecte del nivell d'estratègies d'afrontament (X) dels subjectes sobre el seu nivell d'estrès (Y). Aquestes dades seran utilitzades en els següents apartats per a il·lustrar: (1) com obtenir el valor dels dos coeficients del model de regressió lineal –la qual cosa es coneix com l'estimació o *identificació del model*; (2) com utilitzar el model de regressió obtingut per a realitzar prediccions en "Estrès" a partir del valor d'"Afrontament" dels subjectes; i (3) com valorar la qualitat d'aquestes prediccions –la qual cosa es coneix com l'anàlisi de la *bondat d'ajustament o capacitat predictiva del model*.

Les dades corresponents a aquest estudi es mostren en la taula inferior: en concret, les puntuacions recollides a partir d'una mostra de 27 subjectes en una escala observacional d'"Estrès" i en un test orientat a avaluar la utilització d'estratègies d'"Afrontament". El rang de puntuacions en ambdues

variables podia oscil·lar entre 0 i 100, on puntuacions més altes indiquen major estrès i major capacitat d'utilització de mecanismes d'afrontament. El diagrama de dispersió permet visualitzar la distribució conjunta de les puntuacions en totes dues variables.

Caso	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Estrés	61	26	32	22	38	80	17	10	47	15	50	25	50	30	78	10	35	31	4	6	7	17	37	45	50	67	70
Afronta	38	80	40	84	62	18	65	78	22	60	50	58	20	45	19	84	63	43	87	84	83	85	35	15	29	28	35



2. Estimació de la recta de regressió lineal

• La identificació o estimació d'un model de regressió suposa obtenir els coeficients que el caracteritzen. En el cas del model de regressió lineal simple, B_0 i B_1.

• Això suposa aplicar un procediment de càlcul (mètode d'estimació) que permeta, a partir de les dades disponibles, obtenir els coeficients de l'equació d'una línia recta que represente òptimament la distribució conjunta de les variables modelades. Ara bé, quina és la línia recta que representa òptimament el núvol de punts d'un diagrama de dispersió?

• En principi, un criteri natural de bondat d'ajustament suposa considerar l'equació de regressió que produïsca un menor error en les prediccions. Aquest aspecte es concreta en el fet que la millor recta siga aquella per a la qual la suma dels quadrats dels errors (*SCE*) tinga un valor més pròxim a 0. Així, per a aquest mètode, conegut com a mètode dels mínims quadrats ordinaris, la millor recta de regressió, d'entre totes les possibles que es poden ajustar a la distribució conjunta de 2 variables, serà aquella per a la qual la *SCE* siga mínima:

$$\textit{Millor model de regressió} \rightarrow \min(SCE) = \min\left(\sum E_i^2\right) = \min\left(\sum (Y_i - \hat{Y}_i)^2\right)$$

• Després de realitzar les derivacions matemàtiques pertinents, sobre les quals no s'entrarà ací, les fórmules d'obtenció dels coeficients del model de regressió lineal simple ($\hat{Y} = B_0 + B_1 \cdot X$) que fan que la *SCE* siga mínima (criteri del mètode dels mínims quadrats ordinaris) són les següents:

$$B_1 = R_{XY} \cdot \frac{S_Y}{S_X} \qquad B_0 = \overline{Y} - B_1 \cdot \overline{X}$$

Com pot observar-se, l'obtenció de B_0 implica haver calculat prèviament B_1.

Exercici 2:

a) Obtingueu el valor dels coeficients B_0 i B_1 per a l'exemple sobre les variables "Afrontament" i "Estrès" (vegeu-ne l'enunciat més amunt, p. 7-8), tenint en compte la següent informació sobre aquestes variables: R_{xy}= –0,847; S_X = 24,80; S_Y = 22,37; $\overline{X}$ = 52,22 i $\overline{Y}$ =35,56

b) Escriviu l'equació de la recta de regressió.

c) Quina predicció d'estrès faríem per al subjecte núm. 8, el qual té una puntuació de 78 en l'escala d'afrontament (X_i = 78)? Quin seria l'error de predicció (E_i) per a aquest subjecte?

d) Interpreteu els coeficients de la recta de regressió.

e) Dibuixeu (de manera aproximada) la recta de regressió sobre el diagrama de dispersió de les variables presentat anteriorment.

f) A continuació es mostren els *outputs* obtinguts amb el programa SPSS de l'anàlisi de regressió per a aquest exemple. Identifiqueu-hi els resultats obtinguts anteriorment.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.847[a]	.717	.705	12.14

a. Variables predictoras: (Constante), Estrategias de afrontamiento

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	75.425	5.532		13.634	.000
	Estrategias de afrontamiento	-.763	.096	-.847	-7.951	.000

a. Variable dependiente: Puntuación escala de estrés

3. Bondat d'ajustament

• La bondat d'ajustament d'un model de regressió es refereix al grau en què aquest és convenient com a model que representa la distribució conjunta de les variables implicades. Tal com hem vist, en ajustar un model de regressió lineal simple a la distribució conjunta de dues variables obtindrem la millor recta de regressió d'entre totes las possibles que es poden ajustar a aqueixa distribució; ara bé, això no significa que siga bona. Així, pot ocórrer que la distribució conjunta de dues variables siga difícil de modelar a causa de la inexistència de relació entre les variables (vegeu, per exemple, el cas de la Figura 1 a continuació) o bé que el model de regressió lineal no siga el més adequat per a aqueix propòsit (vegeu, per exemple, el cas de la Figura 2).

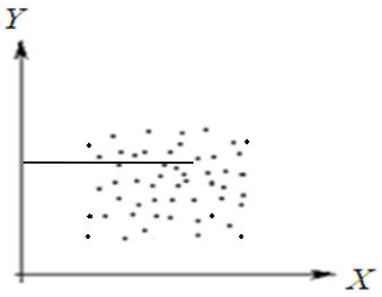


Figura 1: Absència de relació

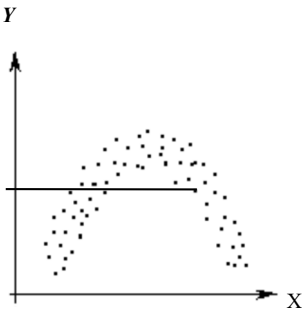
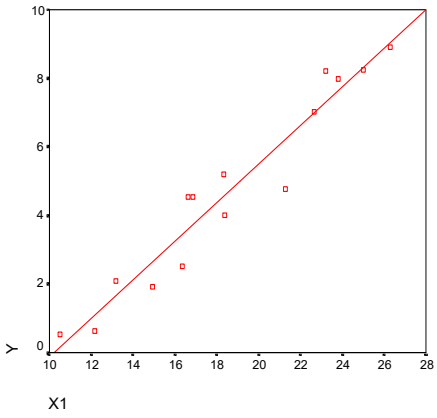
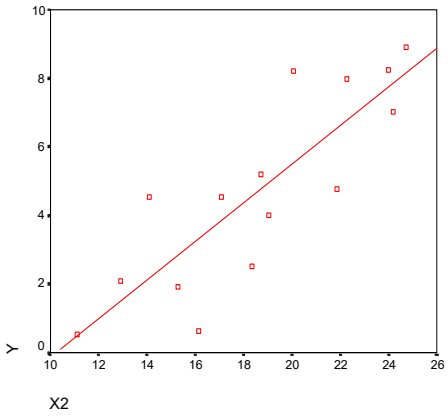


Figura 2: Relació no lineal

Exemple: la relació entre les dues parelles de variables *(X1, Y1)* i *(X2, Y2)* que apareix representada en els dos següents diagrames de dispersió (Losilla i cols., 2005) és descrita, *casualment*, per la mateixa equació de regressió lineal ($\hat{Y} = -5{,}74 + (0{,}56 \cdot X)$). Tanmateix, tal com es pot intuir a nivell visual, la bondat d'ajustament de l'equació de la figura de l'esquerra serà millor que la de la figura de la dreta.



Model 1: $\hat{Y} = -5{,}74 + (0{,}56 \cdot X)$



Model 2: $\hat{Y} = -5{,}74 + (0{,}56 \cdot X)$

• Existeixen diferents aproximacions en l'avaluació de la bondat d'ajustament d'un model a la realitat que aqueix model pretén representar. Una primera elemental consisteix a comparar les puntuacions predites pel model de regressió ($\hat{Y}_i$) amb les puntuacions reals a partir de les quals ha sigut estimat (Y_i). L'índex més utilitzat en aquesta aproximació és, precisament, el conegut com la suma de quadrats dels errors de predicció (o residuals) ($SCE\ o\ SC_{Y \cdot X}$), al com ja es va al·ludir en l'apartat anterior:

$$SCE\ (o\ SC_{Y \cdot X}) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

• La suma de quadrats dels errors o residuals pot oscil·lar entre 0 i qualsevol valor positiu. Si aquest sumatori és igual a 0, el model de regressió s'ajusta perfectament a les dades; com més gran siga el seu valor, això significarà que més errònies són les prediccions de l'equació de regressió i, per tant, pitjor la seua bondat com a model predictiu. Com a conseqüència d'aquesta absència d'un valor màxim, aquest índex pot resultar difícil d'interpretar en la pràctica.

• Un altre índex que supera el problema interpretatiu de la *SCE* ha sigut proposat prenent com a punt de referència una relació bàsica que es dona quan s'ajusta un model de regressió lineal. És la que es coneix com igualtat de la descomposició de la variància de Y, la qual es deriva del axioma que estableix que la puntuació observada en la variable de resposta és igual a la predita segons el model de regressió més l'error de predicció comés:

$$Y_i = \hat{Y}_i + E_i .$$

A partir de l'anterior igualtat es pot derivar algebraicament aquesta altra expressada en forma de variàncies:

$$S_Y^2 = S_{\hat{Y}}^2 + S_{Y \cdot X}^2$$

> Així, la variància en les puntuacions de la variable de resposta (Y) és igual a la variància explicada pel model de regressió (variància de les puntuacions predites) més la variància no explicada pel model de regressió (variància dels errors o residuals).

• Com a conseqüència de la igualtat de descomposició de la variància, es pot plantejar de manera immediata un índex de bondat d'ajustament del model de regressió com a raó de la variància explicada pel model de regressió ($S_{\hat{Y}}^2$) respecte a la variància total (S_Y^2) $\rightarrow$ $\frac{S_{\hat{Y}}^2}{S_Y^2}$

Com més gran siga la variància explicada respecte a la variància total, més pròxim a 1 serà el valor d'aquest índex, posant de manifest una major capacitat predictiva de la variable explicativa respecte a la variable de resposta. Aquest índex, conegut en la literatura estadística com a coeficient de determinació, s'expressa simbòlicament com R^2_{XY}, perquè també pot ser obtingut elevant al quadrat el coeficient de correlació de Pearson entre la variable predictora i la variable de resposta.

$$R^2_{XY} = \frac{S^2_{\hat{Y}}}{S^2_Y}$$

• El coeficient de determinació (R^2_{XY}) representa la proporció de variància de *Y* explicada per les variables implicades en el model de regressió ajustat a les dades (*X* en el model de regressió lineal simple). Atès que es una raó, aquest coeficient oscil·larà sempre entre 0 i 1, de manera que com més pròxim siga R^2_{XY} a 1, indicarà millor bondat d'ajustament del model de regressió a la distribució conjunta de les variables. Si R^2_{XY} és igual a 1, la bondat d'ajustament serà perfecta.

• El següent quadre resumeix els principals conceptes exposats en aquest apartat:

Variància de *Y* = **Variància de *Y* explicada por *X*** (variància de las puntuacions predites) + **Variància de *Y* no explicada por *X*** (variància dels errors o residuals)

$$S^2_Y = S^2_{\hat{Y}} + S^2_{Y \cdot X}$$

$\frac{S^2_{\hat{Y}}}{S^2_Y}$ = Proporció de la variància de *Y* que es explicada per *X* = Coeficient de determinació = R^2_{XY}

$\frac{S^2_{Y \bullet X}}{S^2_Y}$ = Proporció de la variància de *Y* que NO es explicada per *X*

Exemple de càlcul de les variàncies associades a la igualtat de descomposició de la variància en una anàlisi de regressió d'una variable *Y* sobre una variable *X*:

X	*I*
4	2
8	11
11	9
2	3
5	10

Sabent que: $\overline{X} = 6$; $S_X = 3{,}16$; $\overline{Y} = 7$; $S_I = 3{,}74$; $R_{XY} = 0{,}69$

i que l'equació de la recta de Y sobre X és: $\hat{Y} = 2{,}08 + 0{,}82 \cdot X$

vegem com s'obtenen els valors predits ($\hat{Y}_i$) i els residuals (E_i) per a cada cas:

X	I	$\hat{Y}$	E_i	E_i^2	$(\hat{Y}_i - \overline{Y})^2$
4	2	5,36	-3,36	11,29	2,69
8	11	8,64	2,36	5,57	2,69
11	9	11,1	-2,1	4,41	16,81
2	3	3,72	-0,72	0,52	10,76
5	10	6,18	3,82	14,59	0,67
				$S^2_{Y \cdot X} = 36{,}4/5 = 7{,}28$	$S^2_{\hat{Y}} = 33{,}62/5 = 6{,}72$

A partir dels residuals i els valors predits es podrien obtenir les seues variàncies respectives que, en aquest cas són iguals a 7,28 i 6,72, respectivament:

- Variància dels errors (o residuals) → $S^2_{Y \cdot X} = 7{,}28$
- Variància de les puntuacions predites → $S^2_{\hat{Y}} = 6{,}72$

Descomposició de la variància de Y ($S^2_Y = 3{,}74^2 = 14$):

$$14 = 6{,}72 + 7{,}28$$
$$\downarrow \quad \downarrow \quad \downarrow$$
$$S^2_Y = S^2_{\hat{Y}} + S^2_{Y \cdot X}$$

Coeficient de determinació (proporció de la variància de Y explicada per X):

$R^2 = 6{,}72/14 = 0{,}48$

Si s'eleva al quadrat el coeficient de correlació entre X e Y ($= 0{,}69^2 = 0{,}48$) obtindrem també el mateix valor ja calculat per al coeficient de determinació.

Exercici 3: Tenim dues variables Q i T de les quals sabem que la variància de T és 10 (S^2_T=10) i que en l'anàlisi de regressió de T sobre Q, la variància dels errors és 8 ($S^2_{Y \cdot X}$= 8). A partir d'aquesta informació, obtingueu el coeficient de correlació de Pearson entre *Q* i *T*.

Exercici 4: En una mostra de 10 alumnes d'ensenyament secundari s'han mesurades dues variables: (1) rendiment en el curs, quantificat com la mitjana de les qualificacions de les assignatures del curs (*Y*); (2) la mitjana d'hores d'estudi setmanal durant el curs, obtinguda a partir del informe dels estudiants (*X*). Les dades recollides (*N* = 10) són els que es mostren a continuació:

X	*Y*
5	3
12	6
7	4
9	5
15	9
10	6
12	6
8	5
18	9
14	7

Sabent que $\overline{X} = 11$, $\overline{Y} = 6$, $S_X = 3{,}77$, $S_I = 1{,}84$ i que $R_{XY} = 0{,}964$, obtingueu: (a) l'equació del model de regressió lineal de *Y* sobre *X*; (b) els valors predits per l'equació de regressió per a cada subjecte ($\hat{Y}_i$); (c) els errors de predicció o residuals per a cada subjecte (Y_i); (d) sabent que la variància dels errors ($S^2_{Y \cdot X}$) és igual a 0,239 i que la variància de les puntuacions predites ($S^2_{\hat{Y}}$) és igual a 3,16, comproveu que es compleix la igualtat de la descomposició de la variància; (e) obtingueu el coeficient de determinació [de dues formes: (e.1) a partir de les variàncies; (e.2) a partir del coeficient de correlació entre *X* e *Y*]; (f) interpreteu els coeficients de la recta de regressió obtinguts (B_0 i B_1); (g) estimeu quina serà la puntuació mitjana obtinguda a final de curs per un estudiant que estudia 16 hores a la setmana de mitjana.

Exercici 5: A continuació es mostra l'*output* de l'anàlisi de regressió realitzat amb SPSS per a les dades de l'exercici anterior. Identifiqueu-hi els resultats obtinguts anteriorment.

Resum del model

Model	R	R quadrat	R quadrat corregida	Error típ. de l'estimació
1	.964(a)	.930	.921	.546

ANOVA

Model		Suma de quadrats	gl	Mitjana quadràtica	F	Sig.
1	Regressió	31.613	1	31.613	105.935	.000(a)
	Residual	2.387	8	.298		
	Total	34.000	9			

Coeficients

Model		Coeficients no estandarditzats		Coeficients estandarditzats	T	Sig.	Interval de confiança per a *B* al 95%	
		B	Error típ.	Beta			Límit inferior	Límit superior
1	(Constant)	.810	.533		1.52	.167	-.419	2.039
	Hores estudi	.472	.046	.964	10.29	.000	.366	.578

Exercici 6: En un exemple introduït anteriorment amb les variables d'"Afrontament" i "Estrès", sabem que $R_{XY} = -0{,}847$ i que $S_Y = 22{,}37$. Preguntes: (a) quin és el valor del coeficient de determinació?, com s'interpreta aquest valor?; (b) quin és el valor de la variància de Y explicada pel model de regressió (en aquest cas, per la variable "Afrontament")?; (c) quin és el valor de la variància dels residuals?

4. La regressió lineal múltiple

• Una generalització del model de regressió lineal simple és el model de regressió lineal múltiple, que permet considerar més d'una variable explicativa –en principi, quantitatives– en el model de regressió. La formulació del model és:

$$\hat{Y} = B_0 + B_1 \cdot X_1 + B_2 \cdot X_2 + ... + B_j \cdot X_j$$

• La importància relativa de cada variable explicativa és avaluada a través del seu coeficient b_j corresponent. Cal assenyalar que perquè aquests coeficients siguen comparables, ja que la seua magnitud depèn de la unitat de mesura de la variable X a la qual multipliquen, s'ha de comparar el seu valor estandarditzat (els denominats coeficients tipificats o Beta).

Exemple d'aplicació d'un model de regressió lineal múltiple per a explicar la "Satisfacció amb la carrera" en una mostra d'estudiants universitaris.

Satisfacció carrera' = $B_0 + B_1 \cdot$ Hores estudi + $B_2 \cdot$ Relació professors + $B_3 \cdot$ Nota accés

A continuació es mostren els resultats obtinguts amb SPSS en ajustar el model de regressió anterior a un conjunt de dades. Quin percentatge de la variància de "Satisfacció amb la carrera" és explicat a partir d'aquest model? Quina de les variables explicatives és més rellevant?

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,362[a]	,131	,116	1,385

a. Variables predictoras: (Constante), Nota media de acceso, Horas de estudio, Relación con profesores

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	6,386	1,305		4,894	,000	3,809	8,963
	Horas de estudio	,031	,017	,135	1,823	,070	-,003	,066
	Relación con profesores	,238	,058	,303	4,090	,000	,123	,353
	Nota media de acceso	-,186	,193	-,070	-,960	,339	-,568	,196

a. Variable dependiente: Satisfacción con la carrera

5. Descripció estadística de la relació entre dues variables: taula resum

• En els temes 5 i 6 s'han presentat diversos procediments estadístics, tant numèrics com gràfics, adequats per a descriure la relació entre dues variables. En la taula següent es resumeix aquesta informació en funció de l'escala de mesura de les variables implicades.

	Gràfics	Índexs numèrics
Categòrica – Categòrica	Gràfic de barres agrupat Gràfic de barres apilades agrupat	*khi-quadrat* de Pearson *phi* de Pearson *V* de Cramer
Categòrica – Quantitativa	Gràfic de caixa i bigots agrupat Polígon de freqüències agrupat Panell d'histogrames Gràfic de mitjanes.	Diferència de mitjanes *d* de Cohen *f* de Cohen
Quantitativa – Quantitativa	Diagrama de dispersió	Covariància Coef. de correlació de Pearson Coef. de determinació Equació de regressió lineal Coef. de correlació de Spearman (si relació no lineal)

Referències

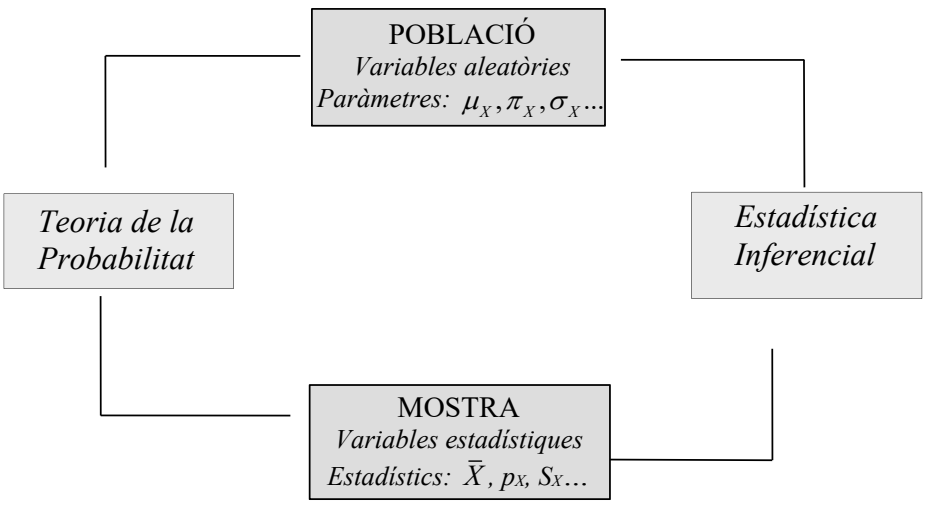
Losilla, J. M., Navarro, B., Palmer, A., Rodrigo, M. F. i Ato, M. (2005). *Del contraste de hipótesis al modelado estadístico.* Documenta Universitaria. [www.edicionsapeticio.com]

Tema 7 – Ús de la probabilitat en la investigació en ciències de la salut

1. Teoria de la Probabilitat

2. Variables aleatòries

• La importància de la Teoria de la Probabilitat en l'àmbit de l'estadística es deriva del fet que constitueix un dels pilars teòrics de l'estadística inferencial. Així, tal com s'il·lustra en la figura que es mostra a continuació, quan coneixem en la població les característiques (tendència central, dispersió, associació...) d'una o més variables, la Teoria de la Probabilitat ens permet fer prediccions de les característiques que aqueixes variables tindran en una mostra de subjectes extreta a l'atzar d'aqueixa població. En sentit invers, l'estadística inferencial –a partir del coneixement desenvolupat per la Teoria de la Probabilitat en aqueix camí de la població en la mostra- ha establit les bases per a abordar el camí oposat: inferir, a partir de les dades en una o més variables d'una mostra, com seran les característiques (tendència central, dispersió, associació...) d'aqueixes variables en la població a la qual aqueixa mostra representa.



1. Teoria de la Probabilitat

• Davant un esdeveniment de resultat incert, el camp de coneixements de la Teoria de la Probabilitat ha dirigit els seus esforços a determinar el grau en què pot ocórrer qualsevol dels resultats possibles [successos] que es poden derivar de la realització d'aquest esdeveniment incert [experiment aleatori].

> **Exemples** d'esdeveniment incert: el llançament d'una moneda (successos possibles: cara; creu); el meu estat de salut durant el mes vinent (successos possibles: bo; dolent; regular); la pràctica religiosa d'un estudiant de la Universitat triat a l'atzar (successos possibles: cap; catòlica; protestant; etc.); el *CI* d'aqueix mateix estudiant (successos possibles: que siga igual a 120; que siga igual a 85; que siga...). En aquest últim cas, els successos es podrien expressar, no en forma *de successos elementals*, sinó de *successos compostos*, per exemple: que siga menor de 110; que siga major o igual a 110; que estiga entre 89 i 120, etc.

• L'esforç de la Teoria de la Probabilitat per determinar el grau en què pot ocórrer un qualsevol dels successos associats a un determinat experiment aleatori s'ha concretat en l'assignació d'un valor numèric que reflectisca el grau en què és previsible l'ocurrència d'aqueix succés. A aquest valor numèric se'l coneix com a probabilitat (P) i pot, per convenció, oscil·lar entre 0 i 1 (0: probabilitat nul·la; 1: probabilitat segura). Així, per a un succés i d'un experiment aleatori X, l'anterior propietat s'expressa com:

$$0 \leq P(X_i) \leq 1$$

Una altra propietat important de les probabilitats és que, per als diferents (k) successos elementals associats a un experiment aleatori, la suma de les seues probabilitats serà igual a 1:

$$\sum_{i=1}^{n} P(X_i) = 1$$

• A continuació es descriuran 3 aproximacions en l'estimació de les probabilitats associades als resultats possibles d'un esdeveniment incert o experiment aleatori. Atès que normalment aquests enfocaments el que permeten obtenir són estimacions, no els vertaders valors de probabilitat, farem referència a aquests valors estimats amb el símbol *P'*, mentre que per al vertader valor de probabilitat es reserva el símbol *P*.

(1) Aproximació subjectiva: suposa estimar la probabilitat d'un succés en funció del grau de confiança personal que es té sobre l'ocurrència d'aquest, ja vinga aqueixa confiança determinada per la nostra experiència vital, per les nostres conviccions personals o creences o per qualsevol altra font sobre la qual es base el coneixement que tenim del nostre entorn. Es tracta del procediment més utilitzat des de sempre en la pràctica per a estimar probabilitats, especialment, quan no es tenen certes nocions sobre

altres aproximacions al càlcul de probabilitats o bé quan aplicar aquestes resulte poc operatiu. Per exemple, quan trac el cap a la finestra abans d'eixir de casa i veig el cel, realitze una estimació de la probabilitat que ploga durant el dia i, com a conseqüència d'aqueixa estimació, decidisc quina roba posar-me o si agafar un paraigua o no. En realitat, fem aquest tipus d'estimacions subjectives de probabilitat en múltiples situacions, encara que no sempre de forma conscient, i són un element determinant de les decisions que finalment prenem.

> **Exemples** d'estimació subjectiva de probabilitat: (a) per a estimar la probabilitat que en llançar dos daus el resultat en tots dos siga un sis, moltes persones realitzarien una estimació subjectiva de la mateixa perquè, encara que existeixen altres aproximacions més precises per a realitzar aqueixa estimació, la seua aplicació és desconeguda per a molts; (b) també les persones fan estimacions subjectives de la probabilitat que els toque el 'primer premi' en un sorteig de loteria -en general, molt a l'alça-; (c) també és habitual realitzar estimacions subjectives de la probabilitat respecte al resultat d'un partit, per exemple, que guanye el València CF en el partit del pròxim cap de setmana.

(2) Aproximació clàssica (o a priori): consisteix a estimar la probabilitat d'un succés (X_i) com la raó entre els resultats favorables a aqueix succés i el nombre total de resultats possibles que es poden donar en la realització de l'experiment aleatori.

$$P'(X_i) = \text{nombre de resultats favorables / nombre de resultats possibles}$$

> **Exemple**: quina és la probabilitat que en llançar un dau isca un 5?
>
> $$P'(X_i = 5) = \frac{1}{6} = 0{,}167$$

Exercici 1: quina és la probabilitat que en llançar un dau isca un 3?; i que isca un número parell?; i que, en llançar dos daus, la suma dels punts siga igual a 7?; i que en la loteria de Nadal em toque el primer premi?

Cal subratllar que, per a calcular una probabilitat, l'aplicació de l'aproximació clàssica assumeix el conegut com a principi d'indiferència, això és, que la probabilitat d'ocurrència de tots els successos és la mateixa. Si es compleix aquest supòsit en la realització d'un determinat experiment aleatori, llavors podrem dir que les estimacions realitzades d'acord amb aquesta aproximació seran els vertaders valors de probabilitat. No obstant això, el compliment d'aquest principi que assumeix que els successos són equiprobables resulta difícil d'acceptar en moltes situacions en la pràctica. Per exemple, si apliquem l'aproximació clàssica per tal d'estimar la probabilitat que un/a estudiant triat a l'atzar de la Universitat

siga viudo/a, aquesta seria igual a ¼ , un resultat poc creïble però que ha vingut motivat per realitzar l'estimació en un cas en què no es compleix el principi d'indiferència. Per experiència, bé sabem que els 4 successos possibles [solter/a, casat/a, separat/a i viudo/a] no són en absolut equiprobables en aquesta població.

En alguns casos sí que es pot assumir el compliment d'aquest principi –per exemple, en jocs d'atzar–, però en molts altres casos es pot tenir seriosos dubtes sobre la satisfacció d'aquest, la qual cosa qüestionaria l'aplicació d'aquesta aproximació per al càlcul de la probabilitat.

(3) Aproximació basada en la freqüència relativa de l'esdeveniment (o a posteriori): si tenim un succés X_i associat a la realització d'un determinat experiment aleatori, l'estimació de la probabilitat de X_i es fonamenta en la repetició d'una gran quantitat de vegades l'experiment aleatori en les mateixes condicions, per a així obtenir la raó entre el nombre de vegades que ha ocorregut aqueix succés (n_i) i el nombre de repeticions de l'experiment (n):

$$P'(X_i) = \frac{n_i}{n}$$

Si ens fixem en la fórmula anterior, l'estimació de la probabilitat d'un succés es correspon amb la fórmula de la freqüència relativa o proporció (p_i) que vam veure en construir una distribució de freqüències:

$$P'(X_i) = p_i$$

Exercici 2: Com s'estimaria la probabilitat, d'acord amb aquesta aproximació, que isca un 3 en el llançament d'un dau? I de la resta de successos plantejats en l'exercici 1?

D'acord amb l'aproximació basada en la freqüència relativa, com més gran siga el nombre de repeticions de l'experiment aleatori, més pròxim serà el valor de probabilitat estimat ($P'(X_i)$) al vertader valor de probabilitat $P(X_i)$. En notació matemàtica:

$$P(X_i) = \lim_{n\to\infty} \frac{n_i}{n}$$

> **Exemple**: Si llancem una moneda 10 vegades per tal d'estimar la probabilitat que isca cara, la probabilitat estimada podria ser, per exemple, P'(cara) = 0,6 si isqueren 6 cares i 4 creus en els 10 llançaments. No obstant això, a mesura que augmenta el nombre de llançaments (idealment, fins a infinit) aquesta estimació s'anirà acostant a la probabilitat vertadera. Se suposa que aqueix valor serà igual a 0,5 en infinits llançaments de la moneda.

En realitat, qualsevol de les variables vistes en els exemples dels temes precedents pot contemplar-se com la repetició d'un experiment aleatori concret, perquè consisteix en la mesura d'un atribut determinat en múltiples ocasions –tantes com diferents subjectes siguen mesurats–, sense tenir certesa a priori de quins seran els valors resultats (successos) d'aquestes mesures. D'acord amb aquesta l'aproximació al càlcul de la probabilitat d'un esdeveniment, les freqüències relatives que s'obtinguen en calcular la distribució de freqüències d'eixa variable seran estimacions de les probabilitats corresponents.

> **Exemple**: Volem estimar la probabilitat d'estar casat (X_i = casat) dels estudiants de la UVEG i disposem d'una mostra de 500 estudiants d'aquesta universitat:
> - Experiment aleatori: obtenir informació de l'estat civil d'un estudiant de la UVEG.
> - Repeticions de l'experiment aleatori: es pregunta a 500 estudiants (n = 500).
> - Nombre d'ocurrències del succés d'interés: nombre d'estudiants casats en aqueixa mostra, suposem que en són 60 (n_{casat} = 60).
> - $P'(X$= casat) = 60/500 = 0,12. Aquesta estimació s'aproximarà més a la probabilitat vertadera com més gran siga el nombre de repeticions, en aquest cas, com més gran siga la mostra. Estes probabilitats es consideraran els vertaders valors de probabilitat (i no estimacions), en el cas en què es compte amb dades per a tots els elements de la població (i no d'una mostra).

Exercici 3 (adaptat a partir d'exemple de Baró-López, 2005): S'ha repetit en 1000 ocasions l'experiment de triar a una dona de la població espanyola de dones entre 45 i 55 anys, i se n'han obtingut dades de les variables "Nivell de massa òssia" [*NO*: Normal; *ON*: Osteopènia; *OR*: Osteoporosi (segons classificació de l'OMS)] i "Haver passat la menopausa" [*N*: No; *S*: Sí]. Les dades obtingudes es mostren resumides en la següent taula de contingència:

		Menopausa		
		No (*N*)	Sí (*S*)	Total
Classificació OMS	Normal (*NO*)	189	280	469
	Osteopènia (*ON*)	108	359	467
	Osteoporosi (*OR*)	6	58	64
Total		303	697	1000

A partir de les dades recollides, contesteu les següents qüestions (entre claudàtors apareix l'equivalent de la qüestió, expressada de manera simbòlica):

a) Quina és la probabilitat (estimada) que una dona (extreta a l'atzar de la població espanyola de dones entre 45 i 55 anys) tinga osteoporosi? [$P'(OR)$]

b) I que no haja passat la menopausa? [$P'(N)$]

c) I que tinga osteopènia o osteoporosi? [$P'(ON \cup OR)$]

d) I que no haja passat la menopausa i tinga osteoporosi? [$P'(N \cap OR)$]

e) I que tinga osteoporosi si sabem que no ha passat la menopausa? [$P(OR|N)$]

f) Si sabem que una dona té osteopènia, quina és la probabilitat estimada que haja passat la menopausa? [$P'(S \mid ON)$]

En aquest exercici es planteja l'aplicació pràctica d'algun dels teoremes bàsics de la probabilitat (més detalls sobre els mateixos en, per exemple, Botella i cols. (2001, tema 12):

- probabilitat de la intersecció de dos successos: $P(A \cap B)$.
- probabilitat de la unió de dos successos: $P(A \cup B)$.
- probabilitat condicional: $P(B \mid A)$ o $P(A \mid B)$

2. Variables aleatòries

• A diferència del concepte de variable estadística, el concepte de variable aleatòria suposa tenir informació de la probabilitat associada a cadascuna de les modalitats de la variable, la qual cosa implica tenir dades de tota la població perquè, en un altre cas, el que tindríem serien freqüències relatives, això és, estimacions de les probabilitats, i no les probabilitats verdaderes.

Exemple: Si mesurem la variable "Estat civil" en tota la població d'estudiants de la *UVEG* (N = 45000) i obtenim que 350 són viudos/es, llavors la freqüència relativa corresponent a la modalitat 'ser viudo/a ' ($p_{viudo/a}$ = 350/45000 = 0,008) serà precisament la probabilitat de 'ser viudo' ($P_{viudo/a}$) en la població d'estudiants de la *UVEG*, i no una estimació d'aquesta. Anàlogament, si obtenim les probabilitats associades a les altres modalitats (successos) de la variable "Estat civil", tindrem la distribució de probabilitat associada a aquesta variable aleatòria en eixa població. Siga, per exemple, la següent:

X_i	$P(X_i)$
solter/a	0,884
casat/ada	0,105
separat/ada	0,009
viudo/a	0,002
	1,00

• La distribució de probabilitat d'una variable aleatòria –de manera anàloga a la distribució de freqüències d'una variable estadística– consisteix en la correspondència entre els diferents valors que pren la variable i les probabilitats associades a aqueixos valors.

• Aquesta correspondència entre les modalitats d'una variable i les seues probabilitats (i. e., distribució de probabilitat) s'anomena funció de probabilitat, en cas de tractar-se d'una variable aleatòria discreta (variables categòriques, ordinals o quantitatives discretes), i funció de densitat de probabilitat, en cas que siga contínua (variables quantitatives contínues).

Exemple: funció de probabilitat corresponent a la variable "Nombre de contractes laborals en els 2 últims anys" per a la població de persones en edat laboral de la comarca del Camp de Morvedre:

X_i	$P(X_i)$
0	0,08
1	0,31
2	0,35
3	0,18
4	0,07
5	0,01
	1

I la funció de distribució és la correspondència entre cada valor de la variable i la probabilitat que es done un valor com aqueix o inferior (probabilitat acumulada (P_a)), concepte anàleg al de freqüència relativa acumulada. Aquest concepte no és aplicable si la variable és categòrica, perquè no és aplicable el concepte de probabilitat acumulada en aquest tipus de variable.

Exemple: funció de distribució per a la variable "Nombre de contractes laborals…" obtinguda a partir de la funció de probabilitat anterior:

X_i	$P_{a\,(Xi)}$
0	0,08
1	0,39
2	0,74
3	0,92
4	0,99
5	1

• La distribució de probabilitat d'una variable no sol ser coneguda, atès que, amb freqüència, no és viable recollir dades de tota la població d'interès per a una determinada variable. Una aproximació a aquesta distribució (basada en la freqüència relativa) consisteix a estimar les probabilitats corresponents a partir de les dades recollides per a una mostra de la població. Una altra via d'aproximació a la distribució de probabilitat d'una variable consisteix a assumir, a partir de raons substantives o de l'experiència pràctica acumulada, que la variable té una distribució que segueix algun

model teòric de característiques conegudes (p. ex. distribució normal, distribució binomial, distribució *t* de Student…).

• Quan s'obté un índex qualsevol –per exemple, la mitjana– a partir de la distribució de probabilitat d'una variable, al valor resultant se l'anomena paràmetre, mentre que si s'obtingués a partir d'una distribució de freqüències, se l'anomenaria estadístic.

• Se solen utilitzar lletres gregues minúscules per a representar als paràmetres. Així, per exemple, donada una variable *X*, s'utilitza μ_X per a representar la mitjana, σ_X per a la desviació típica, π_X per a la proporció... És el mateix per al cas dels índexs estadístics bivariats, per exemple, σ_{XY} per a la covariància, ρ_{XY} per al coeficient de correlació de Pearson, β_0 per a la constant de l'equació de regressió, β_1 per al pendent... Hi ha algun cas especial, sent el més notable el de la mitjana aritmètica que, com a paràmetre, apareix també representada com a *E(X)* i denominada, de manera alternativa, *valor esperat* o, també, *esperança matemàtica*. Finalment, alguns índexs no tenen el privilegi de gaudir d'una doble assignació simbòlica segons siguen paràmetres o estadístics: per exemple, la mediana (*Md*) o el coeficient de variació (*CV*), entre altres.

• L'aplicació sobre la distribució de probabilitat d'una variable aleatòria dels índexs de tendència central, dispersió, etc. implica algunes adaptacions en les fórmules presentades en els temes previs. A títol il·lustratiu, es mostren a continuació les fórmules de la mitjana aritmètica (o valor esperat o esperança matemàtica) i la variància per al cas en què la variable (*X*) siga ordinal o quantitativa discreta:

$$E(X) = \mu_X = \sum X_i \cdot P(X_i)$$
$$\sigma_X^2 = \sum (X_i - \mu_X)^2 \cdot P(X_i)$$

Si les variables són quantitatives contínues, les fórmules es compliquen bastant més perquè intervé el càlcul integral en la seua aplicació. Podeu consultar aquestes fórmules en qualsevol llibre d'estadística avançada.

Exercici 4: A partir de l'exemple presentat abans de la distribució de probabilitat de la variable "Nombre de contractes laborals en els 2 últims anys":

a) Quina és la probabilitat que una persona seleccionada a l'atzar de la població anterior haja tingut 3 contractes en els 2 últims anys?
b) I que haja tingut més de 2 contractes?
c) Obtingueu la mediana i la moda de la variable.
d) Obtingueu el valor esperat (mitjana aritmètica) de la variable.

Exercici 5: Tenim dades de dues variables, "Nombre d'accidents laborals en l'últim any" (X) i "Tipus de contracte" (Y) [Fix; Temporal], en la població de treballadors del sector de la construcció de Gandia ($n = 1000$). La distribució conjunta de freqüències absolutes de totes dues variables aleatòries es mostra en la següent taula de contingència:

Y_i \ X_i	0	1	2	3	
Fix	250	90	50	40	430
Temporal	150	160	160	100	570
	400	250	210	140	1000

a) Obtingueu-ne la distribució de probabilitat (funció de probabilitat) de la variable X [$P(X_i)$].
b) Obtingueu-ne la funció de distribució de X [$P_a(X_i)$].
c) Obtingueu-ne la distribució de probabilitat conjunta de totes dues variables [$P(X_i, Y_j)$].
d) Quina és la probabilitat que un treballador (extret a l'atzar d'aquesta població)…
    1. haja tingut 1 o més accidents? [$P\ (X \geq 1)$]
    2. tinga contracte fix i haja tingut 0 accidents? [P (Fix $\cap$ 0)]
    3. haja tingut 2 o 3 accidents? [$P(2 \cup 3)$]
    4. tinga un contracte fix? [P (Fix)]
    5. haja tingut 0 accidents, si sabem que té un contracte fix? [$P\ (0 \mid Fix)$]
e) Obtingueu la moda, la mediana i l'esperança matemàtica de la variable X.

• La següent taula resumeix alguns dels conceptes plantejats fins ara, diferenciats en funció que facen referència a una mostra o a una població:

MOSTRA	*POBLACIÓ*
1. Variable estadística	1. Variable aleatòria
2. Freqüència relativa [p_i]	2. Probabilitat [$P(X_i)$]
3. Distribució de freqüències relatives	3. Distribució de probabilitat → Dos tipus: Funció de probabilitat Funció de densitat de probabilitat
4. Freqüència relativa acumulada [$p_{a\,]}$.	4. Probabilitat acumulada [$P_{a\,(X\,i)}$]
5. Distribució de freqüències relatives acumulades	5. Funció de distribució
6. Estadístic	6. Paràmetre

Referències

Barón-López, J. (2005). *Bioestadística: métodos y aplicaciones.* Anotacions i material disponibles en http://www.bioestadistica.uma.es/baron/apuntes/

Botella, J., León, O. G., San Martín, R. i Barriopedro, M. I. (2001). *Análisis de datos en psicología I: teoría y ejercicios.* Madrid: Pirámide.

Tema 8 – Principals models teòrics de distribució de probabilitat

1. **La distribució binomial**
2. **La distribució o corba normal**
3. **Les distribucions *khi-quadrat*, *t* i *F***

Principals models teòrics de distribució de probabilitat

• El coneixement acumulat en les ciències de la salut ha permès evidenciar com algunes variables d'interès en aquest camp es distribueixen d'una manera característica, això és, tenen una distribució de probabilitat particular que es repeteix al llarg del temps i per a diferents mostres. Per a algunes d'aquestes distribucions de probabilitat s'han plantejat els models teòrics que les representen matemàticament i que, per tant, permeten obtenir fàcilment, a partir d'una funció matemàtica, quina serà la probabilitat (o probabilitat acumulada) associada a un valor qualsevol de la variable.

• Dos dels models més rellevants en el context de les ciències de la salut són el de la distribució binomial, per a variables categòriques, i el de la distribució normal, per a variables quantitatives. En els següents apartats es descriuen les característiques d'aquests dos models teòrics de distribució de probabilitat i es mostra la seua aplicació en la pràctica.

• Altres models teòrics de distribució de probabilitat com la distribució *t* de Student, la distribució *khi-quadrat* i la distribució *F* de Snedecor són també especialment importants en el camp de l'estadística, pel fet que la 'distribució mostral' d'alguns estadístics s'ajusta a aquests models teòrics de distribució de probabilitat. La distribució mostral d'un estadístic és un concepte clau en l'estadística inferencial que serà introduït en un capítol posterior.

1. La distribució binomial

• Comencem amb un cas pràctic: suposem que es tria a l'atzar una mostra de sis persones per a formar part d'un jurat popular que ha de jutjar una persona immigrant i sabem que, en la població de la qual s'ha extret aqueixa mostra, un 30 % de les persones són racistes. A partir d'aquestes dades, quina és la probabilitat que la meitat o més dels membres del tribunal siguen racistes?

• La resposta a la pregunta anterior es pot resoldre fàcilment si assumim que la distribució de probabilitat de la variable que ens ocupa és la distribució binomial. A continuació veurem més formalment les condicions per a poder assumir que la distribució de probabilitat d'una variable s'ajusta al model teòric de la distribució binomial:

→ 1a condició: que es tinga una variable categòrica dicotòmica de la qual es conega la seua distribució de probabilitat en la població d'interès.

Siga la variable dicotòmica X [X_1; X_2]: el primer que s'ha de decidir és quina de les dues modalitats d'aquesta és la que ens interessa, això és, quina és la que es correspon amb allò que ens interessa conèixer (suposem que és X_1). La probabilitat associada a aqueixa modalitat s'expressa simbòlicament com π (i, complementàriament, a la de X_2 com $1-\pi$):

X_i	$P(X_i)$
X_1	π
X_2	$1-\pi$
	1

Exemple: Variable "Ser racista" [Sí; No] => modalitat d'interès per a contestar a la pregunta plantejada en l'exemple: 'Sí' => $P(X_1) = \pi = 0{,}30$

X_i	$P(X_i)$
Sí	0,3
No	0,7
	1

Cal recordar que en la pràctica de l'anàlisi de dades és bastant habitual tenir dades de variables dicotòmiques (o dicotomitzades): per exemple, variables en què s'han recollit dades del tipus correcte/incorrecte, a favor/en contra, d'acord/en desacord, bé/malament, sí/no, curat/no curat, tractament/no tractament, etc. Convé destacar també que és bastant freqüent en la literatura de la distribució binomial considerar a aqueixes dues modalitats, de manera genèrica, amb els termes *Èxit* i *Fracàs*, de manera que $P(\text{Èxit}) = \pi$ i $P(\text{Fracàs}) = 1-\pi$. Aquesta pràctica pot a vegades generar certa confusió, atès que el significat de les paraules *èxit* i *fracàs* no encaixa amb el

significat de les dues modalitats de moltes variables categòriques –com és el cas de la variable d'aquest exemple.

→ 2a condició: Es realitza *n* vegades l'experiment aleatori representat per la variable aleatòria en qüestió, per exemple, mesurar una variable *X* en una mostra de *n* casos –cada vegada que es mesura la variable és un experiment aleatori, perquè no sabem a priori el resultat. S'ha de satisfer la condició que π es mantinga constant al llarg de la realització d'aqueixos experiments.

Exemple: En el cas del jurat popular, es realitzaria 6 vegades (n = 6) l'experiment aleatori consistent a triar un membre del tribunal, perquè són sis els membres del tribunal triats a l'atzar de la població. Sabem que la probabilitat de ser racista en la població és de 0,30 (π = 0,30) i és raonable assumir que aqueixa probabilitat es mantindrà constant al llarg del procés d'elecció dels sis membres. Podria no ser raonable tal assumpció si, per exemple, es dilatara molt en el temps l'elecció de cadascun dels membres del tribunal, perquè al llarg d'aqueix temps podria canviar la probabilitat de ser racista en la població.

• Si es compleixen les dues condicions anteriors, la variable aleatòria "Nombre d'experiments (casos) en els quals es verifica la condició X_I" es distribueix segons el model teòric de la distribució binomial. Si la distribució de probabilitat d'una variable *X* s'ajusta al model binomial s'expressa simbòlicament com: $X \rightarrow B\ (n;\pi)$.

Exemple: atès que es compleixen les dues condicions anteriors, podem afirmar que la variable "Nombre de membres del tribunal que són racistes" (X) es distribueix segons la distribució binomial amb n = 6 i π = 0,30

$$X \rightarrow B\ (6;\ 0{,}30)$$

• La distribució de probabilitat (o funció de probabilitat) d'una variable binomial *X* ve definida per la funció matemàtica següent, on X_i pot oscil·lar entre 0 i n:

$$P(X_i) = \frac{n!}{X_i!(n-X_i)!} \cdot \pi^{X_i} \cdot (1-\pi)^{n-X_i}$$

Si, per **exemple**, volem conèixer la probabilitat que dos membres del tribunal siguen racistes:

$$P(2) = \frac{6!}{2!(6-2)!} \cdot 0{,}30^2 \cdot (1-0{,}30)^{6-2} = \frac{720}{2 \cdot 24} \cdot 0{,}30^2 \cdot 0{,}70^4 = 0{,}324$$

Si substituïm, en la fórmula del model binomial, els diferents valors que pot tenir X en el nostre exemple (des de 0 persones racistes fins a 6 persones racistes), obtindríem la distribució de probabilitat completa de la variable "Nombre de membres del tribunal que són racistes" (X):

X_i	$P(X_i)$
0	0,118
1	0,303
2	0,324
3	0,185
4	0,060
5	0,010
6	0,001
	1

• Un altre procediment alternatiu, que no requereix fer el càlcul anterior, per a obtenir els valors de la distribució de probabilitat d'una variable que segueix el model binomial és acudir a la taula de la distribució de probabilitat d'aquest model, la qual es pot trobar en *l'apèndix de taules estadístiques* que sol aparèixet en la part final de molts llibres d'estadística. En aquesta taula es poden trobar tabulades les distribucions de probabilitat d'una variable binomial per a diferents valors de π i de n (vegeu la Taula 1 al final del tema).

Exemple: Buscant en la taula podríem obtenir fàcilment, per exemple, la distribució de probabilitat d'una variable X que es distribuïsca segons el model binomial amb paràmetres $n = 4$ i $\pi = 0{,}50$ $[X \rightarrow B\ (4;0{,}50)]$

X_i	$P(X_i)$
0	0,062
1	0,250
2	0,375
3	0,250
4	0,062

Un exemple de variable que es distribueix segons aquesta distribució de probabilitat és el "Nombre de cares en llançar una moneda a l'aire 4 vegades".

• Una taula més àmplia de la distribució binomial pot trobar-se en les pàgines finals de molts llibres d'estadística o bé,podem recórrer a l'obtenció informatitzada del valor exacte de probabilitat que ens interesse. Per exemple, en el programa MS Excel® cal introduir en una casella qualsevol la següent expressió amb els valors entre parèntesis que ens interesse:

```
=DISTR.BINOM(nombre_èxit;assajos;prob_èxit;0)
```

Per exemple , si escrivim en una casella `=DISTR.BINOM(2;6;0,30;0)` i premem la tecla *Intro*, automàticament obtindrem el valor de probabilitat que anteriorment vam calcular amb la fórmula del model binomial (P = 0,324). Si el "0" que apareix en últim lloc en el parèntesi el canviem per un "1", obtindrem la probabilitat acumulada, això és, la probabilitat d'obtenir el valor que ens interessa o un valor inferior.

Exercici 1: Per a la variable "Nombre de membres del tribunal que són racistes" presentada abans, obtingueu les següents probabilitats: (a) que 4 membres del tribunal siguen racistes; (b) que, com a màxim, 2 en siguen; (c) que més de la meitat en siguen; (d) obtingueu l'esperança matemàtica d'aquesta variable aleatòria.

Exercici 2: Sabent que en la població espanyola la proporció de dones és de 0,60: (a) quina és la probabilitat que en seleccionar una mostra aleatòria de 7 persones d'aqueixa població, cap d'elles siga dona? (b) i quina es la probabilitat que totes siguen dones?; (c) obtingueu la distribució de probabilitat corresponent a la variable "Nombre de dones en extraure a l'atzar una mostra de 7 persones de la població espanyola" (X); (d) obtingueu la mitjana (o valor esperat), la mediana i la moda de la variable aleatòria X; (e) representeu gràficament la funció de probabilitat de X (f) i també la funció de distribució de X; (g) obtingueu la distribució de probabilitat de la variable aleatòria complementaria, això és, la de la variable "Nombre d'homes en extraure a l'atzar una mostra de 7 persones de la població espanyola".

• És una propietat de qualsevol variable que es distribueix segons la distribució binomial que el seu valor esperat (mitjana) i la seua variància es poden obtenir segons les següents fórmules:

$$\mu_X = n \cdot \pi$$
$$\sigma^2_X = n \cdot \pi \cdot (1-\pi)$$

Exercici 3: Obtingueu amb les anteriors fórmules l'esperança matemàtica i la variància de la variable "Nombre de membres del tribunal que són racistes".

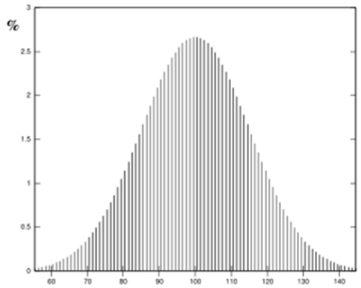
Exercici 4: Suposant que es contesta completament a l'atzar un examen de 10 preguntes de vertader/fals i que es corregeix puntuant amb un 1 els encerts i amb un 0 els errors, obtingueu la probabilitat que es traga un 5 en l'examen. I quina és la probabilitat de traure un 10? I la de traure un 5 o més? Obtingueu el valor esperat de la variable "puntuació en l'examen" i interpreteu-la.

2. La distribució o corba normal

• Es tracta d'un model teòric de distribució de probabilitat per a variables aleatòries quantitatives que es caracteritza, gràficament, per tenir una forma similar a la d'una campana. Per això, i per haver sigut estudiada inicialment pel matemàtic Karl Gauss, s'anomena també corba o campana de Gauss.

• La importància d'aquesta distribució resideix en el fet que diverses variables, com els caràcters fisiològics i morfològics d'individus —alçada, pes o longevitat—, atributs sociològics, psicològics i, en general, variables que venen determinades per molts factors, es distribueixen segons el model de la corba normal.

• A continuació es mostra la representació gràfica d'una variable aleatòria que es distribueix segons el model teòric de la distribució normal (més usualment dit, "que es distribueix normalment").



Com pot observar-se, algunes de les característiques distintives d'aquesta distribució són:

(1) és unimodal;

(2) és simètrica, situant-se l'eix de simetria sobre el valor de la mitjana (mediana, moda) de la distribució de la variable;

(3) és asimptòtica per totes dues cues de la distribució;

(4) fa correspondre valors de probabilitat alts per als valors centrals de la variable, mentre que aqueixes probabilitats disminueixen de manera progressiva quan ens allunyem del centre de la distribució, més acceleradament en la zona central i menys en els extrems.

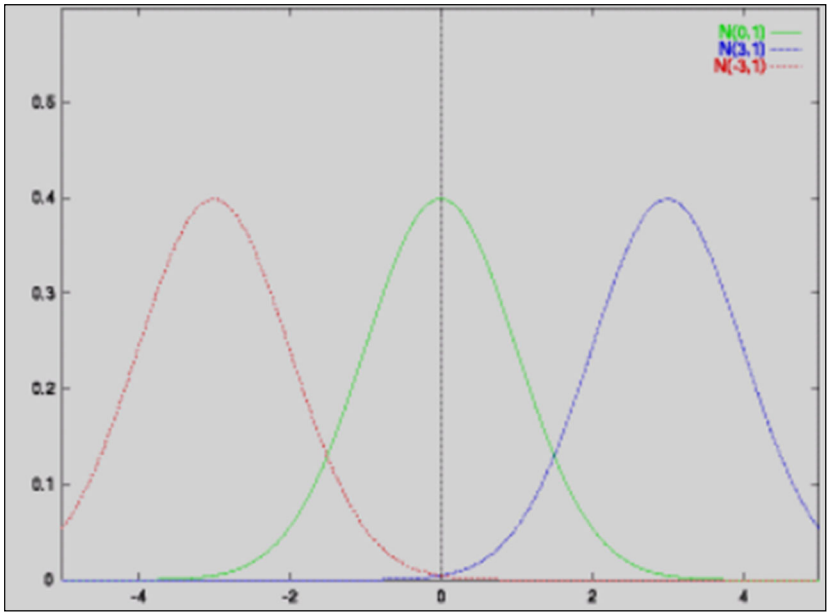
• La distribució de probabilitat (funció de densitat de probabilitat) de la corba normal ve definida matemàticament per la següent funció matemàtica, originalment plantejada per Abraham de Moivre en 1733:

$$P(X_i) = \frac{1}{2{,}507 \cdot \sigma_X} \cdot e^{-0{,}5 \cdot \frac{(X_i - \mu_X)^2}{\sigma_X^2}}$$

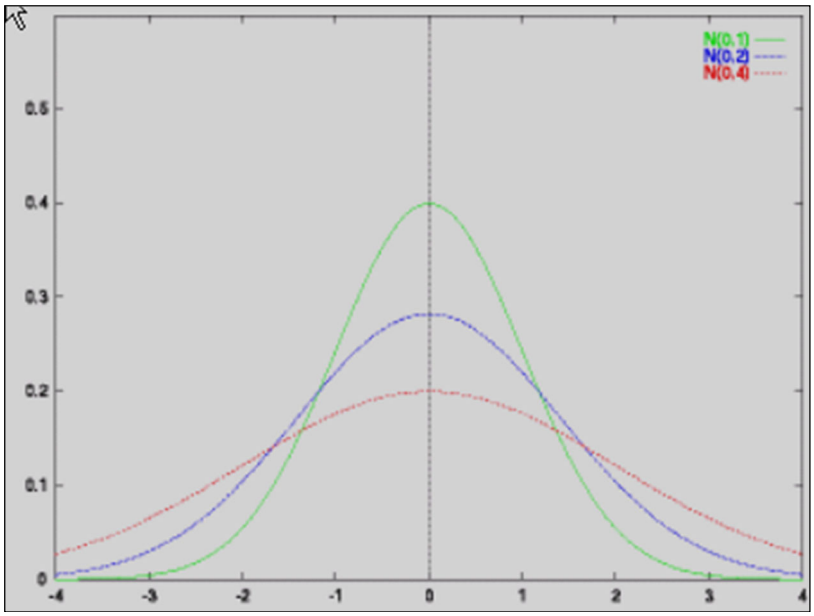
on X_i és un valor concret de la variable X, e és una constant matemàtica (=2,72), i μ i σ són dos paràmetres de la funció que es corresponen, precisament, amb la mitjana i la desviació típica de X.

• Tenint en compte la fórmula presentada, la distribució normal pot adoptar diferents formes, tantes com diferents valors de μ i σ es consideren (és a dir, infinites). Tots aquests models integren la coneguda com a <u>família de la distribució normal</u> i per a representar simbòlicament cadascun dels membres d'aqueixa família s'utilitza l'expressió $N(\mu; \sigma)$. A continuació es mostra la representació gràfica de diversos models de la família de la distribució normal per diferents valors de μ i σ:

- Exemples de models de distribució normal amb diferent valor de μ però el mateix valor de σ:



- Exemples de models de distribució normal amb el mateix valor de μ però diferent valor de σ:



• Entre els models de la família de la distribució normal, el més rellevant en la pràctica és el denominat com distribució normal estàndard o unitària, això és, el model de la família de la distribució normal que té $\mu = 0$ i $\sigma = 1$ [$X \rightarrow N(0; 1)$]. Així, és comú que en els llibres d'estadística s'incloga en un apèndix final de taules estadístiques la corresponent a la corba normal estàndard. Encara que hi ha variacions en la forma en què es presenta aquesta taula en els llibres, és habitual que hi puguem consultar, per a un rang de valors entre -3 i 3, quin és el valor de probabilitat acumulada corresponent al valor que vulguem. En síntesi, es tracta d'una representació tabular de la funció de distribució de la corba normal amb mitjana 0 i desviació típica igual a 1 (vegeu la Taula II al final d'aquest tema). Així, si es té una variable X que es distribueix segons la distribució normal estàndard [$X \rightarrow N(0; 1)$], en aquesta taula podem consultar per a diferents valors de X, quin és el valor de probabilitat acumulada [$P_a(X)$] corresponent.

• Si assumim que una determinada variable es distribueix segons el model de la distribució normal, de manera immediata es pot donar resposta fàcilment a diferents tipus de preguntes com, per exemple, quin és el percentatge acumulat o percentil corresponent a una determinada puntuació, el nombre de subjectes que és d'esperar que tinguen un valor inferior o igual a aquest, o superior a aquest, o entre dos valors determinats, etc. (vegeu les preguntes de l'exercici 5).

• De manera anàloga al que es va dir per a la distribució binomial, una taula completa de les probabilitats acumulades corresponents a la distribució normal estàndard pot trobar-se en l'apèndix de taules de la majoria dels llibres d'estadística; d'altra banda, no resulta difícil trobar aplicacions informàtiques que ens faciliten l'obtenció del valor que té una probabilitat acumulada determinada. En el cas del programa MS Excel®, cal introduir en una casella qualsevol la següent expressió amb el valor de X que ens interesse:

=DISTR.NORM.ESTAND(*X*)

Per exemple, si escrivim en una casella **=DISTR.NORM.ESTAND(*1,5*)** i premem la tecla *Intro*, automàticament obtindrem el valor de probabilitat acumulada corresponent a una puntuació de 1,5 ($P_{a\,(1,5)}$ = 0,933).

Si el que es desitja és seguir el camí invers, això és, a partir d'un valor de probabilitat acumulada, obtenir la puntuació X a la qual correspon aqueixa probabilitat acumulada, podem utilitzar la següent funció de MS Excel®:

=DISTR.NORM.ESTAND.INV(*probabilitat_acumulada*)

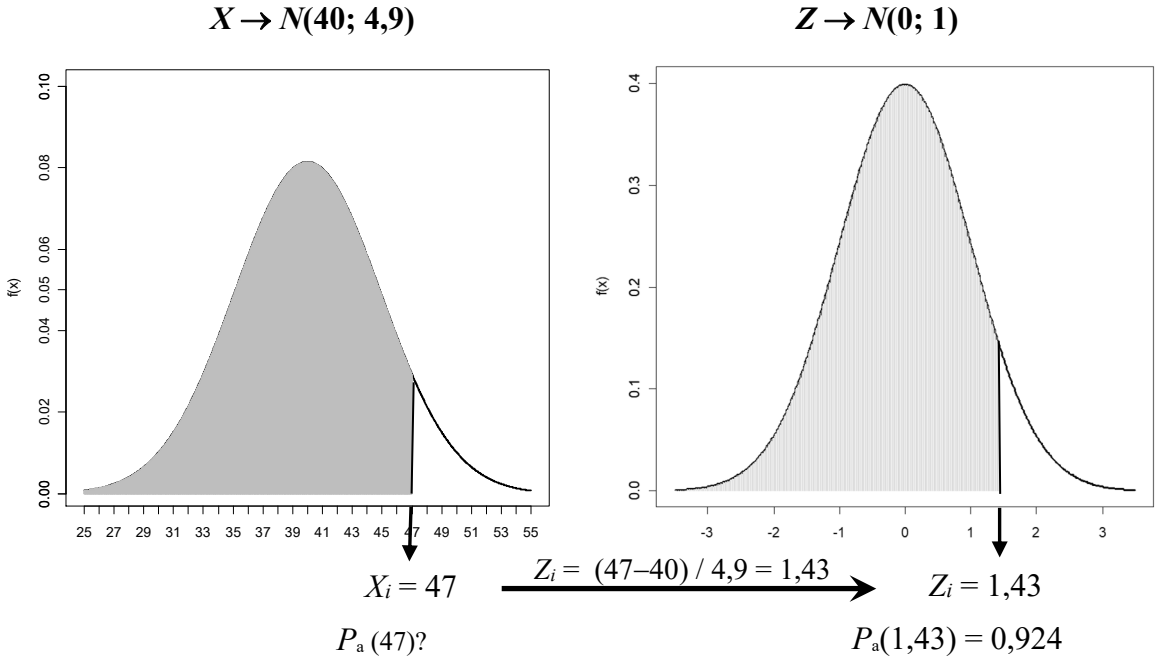
Per exemple , si escrivim en una casella `=DISTR.NORM.ESTAND.INV(0,933)` i premem la tecla *Intro*, automàticament obtindrem el valor de la variable a la qual correspon aqueixa probabilitat acumulada de 0,933 (X (0,933) = 1,5).

Exercici 5: Contesteu les següents preguntes relatives a una variable quantitativa X que per a un determinat grup de subjectes es distribueix segons $N(0;1)$ (per a cadascuna de les preguntes es mostra entre claudàtors l'expressió simbòlica d'aquesta): (a) quina és la probabilitat acumulada corresponent a un valor de X igual a 1,18 [P_a (1,18)]?; (b) quin percentatge de subjectes tindran una puntuació inferior o igual a 1,18?; (c) sabent que el grup de subjectes era de 1000 persones (n =1000), quantes tindran una puntuació inferior o igual a 1,18?; (d) quina és la probabilitat que, en extraure un subjecte a l'atzar, aquest tinga una puntuació inferior o igual a 1 [P_a (1)]? (e) i que siga superior a 1 [1–P_a (1)]? (f) i que estiga entre 1 i 2 [$P(1 \leq X \leq 2)$] (g) i que estiga entre la mitjana de la distribució i 1 [$P(0 \leq X \leq 1)$]?; (h) a quin valor de la variable X li correspon una probabilitat acumulada de 0,75 [$P_a(X) = 0,75$] (això és, el 75 % dels subjectes obtenen una puntuació inferior o igual a aqueix valor en la variable => Q_3)?; (i) quin valor de la variable X serà superat només pel 25 % dels subjectes? (j) quin valor correspon al percentil 25 [$P_a(X) = 0,25$]?; (k) quina és la probabilitat que, en extraure un subjecte a l'atzar, aquest tinga una puntuació inferior o igual a -1 [P_a (-1)]? (l) i quina és que siga superior a -1 [1–P_a (-1)]?

• Una conseqüència pràctica derivada de l'aplicació del model teòric de la distribució normal i, en general, de qualsevol model teòric de distribució de probabilitat és que, si sabem o podem assumir que una variable es distribueix segons un model teòric, llavors es poden obtenir les probabilitats associades a qualsevol valor d'aqueixa variable i, en conseqüència, la corresponent distribució de probabilitat. Serà suficient per a això aplicar la fórmula matemàtica del model de probabilitat corresponent o, més senzill, recórrer a una taula estadística d'aqueix model i consultar-hi els valors que ens interessen.

• Ara bé, com aprofitar la taula de la corba normal unitària si tinc una variable, encara que es puga assumir que es distribueix normalment, la mitjana i la desviació típica de la qual no són precisament 0 i 1? La resposta és transformar els valors de la variable a puntuacions típiques (Z), amb la qual cosa la variable es continuarà distribuint segons la corba normal, si bé tindrà mitjana 0 i desviació típica 1, de manera que es farà factible la utilització de la taula de la corba normal unitària.

Exemple: Siga una variable X que es distribueix segons el model de la corba normal amb mitjana igual a 40 i variància igual a 24 i suposem que desitgem saber quina és la probabilitat d'obtenir un valor inferior o igual a 47 en aqueixa variable [$P(X \leq 47)$].



Seguint amb **l'exemple** anterior on $X \rightarrow N(40; 4,9)$:

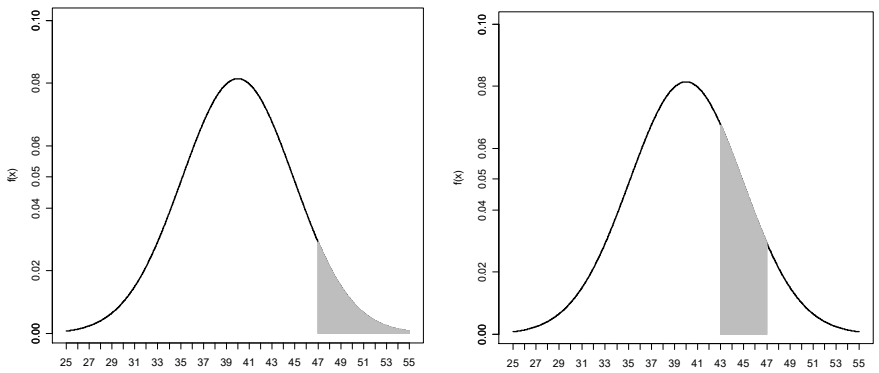
-Quina seria la probabilitat d'obtenir un valor superior a 47 en aquesta variable? $P_a(X > 47)$.

Aquesta probabilitat correspon a l'àrea ombrejada en la figura (a) inferior i seria igual a (1-0,924) = 0,076.

-Quina seria la probabilitat d'obtenir un valor entre 43 i 47 en aquesta variable? $P(43 \leq X \leq 47)$.

Aquesta probabilitat correspon a l'àrea ombrejada en la figura (b) inferior i s'obté com la diferència entre les probabilitats acumulades per a les puntuacions Z corresponents a les puntuacions 43 i 47 (Z=0,61 i Z=1,43, respectivament). Així, restant la probabilitat acumulada major menys la menor s'obté: $P(43 \leq X \leq 47) = 0,924 – 0,729 = 0,195$

(a) (b)



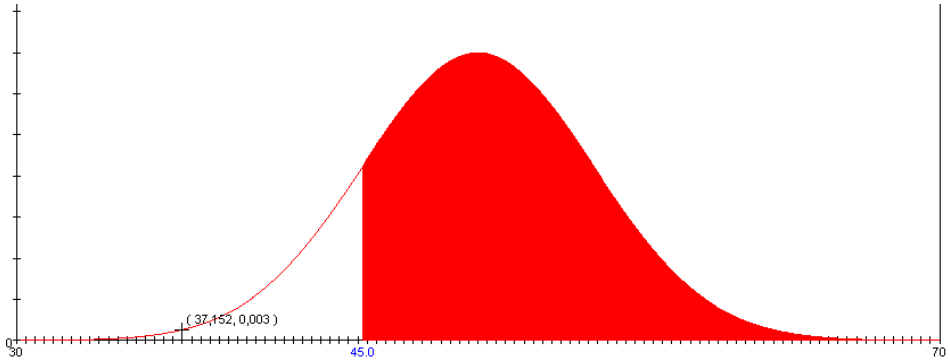
Exercici 6. En la població espanyola de nounats la variable "Alçada en nàixer" es distribueix normalment amb mitjana = 50 cm i desviació típica igual a 5 cm. En alguns del gràfics apareixen les respostes a les preguntes plantejades. Responegueu les preguntes sense mirar aquestes respostes.

Contesteu les següents preguntes, utilitzant les taules de la distribució normal estandarditzada (0,1) o les funcions de MS Excel® d'aquesta distribució:

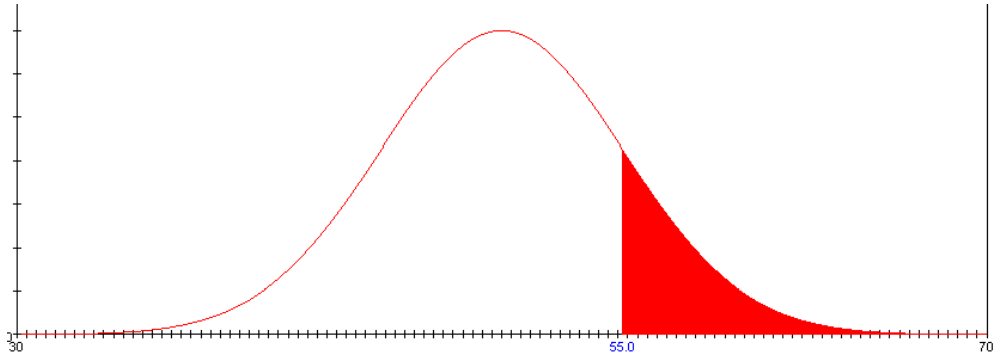
`=DISTR.NORM.ESTAND(X)`: la funció retorna la probabilitat acumulada per al valor *Z* especificat entre parèntesi.

`=DISTR.NORM.ESTAND.INV(probabilitat acumulada)`: la funció retorna el valor *Z* al qual li correspon la probabilitat acumulada especificada entre parèntesi.

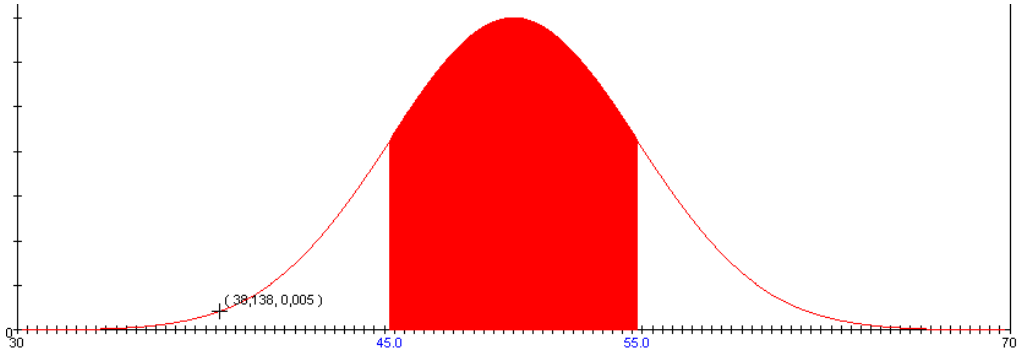
a) Quina és la probabilitat que un xiquet en nàixer tinga una alçada inferior a 45 cm (àrea blanca)? I que aquesta siga superior a 45 cm (àrea roja)?



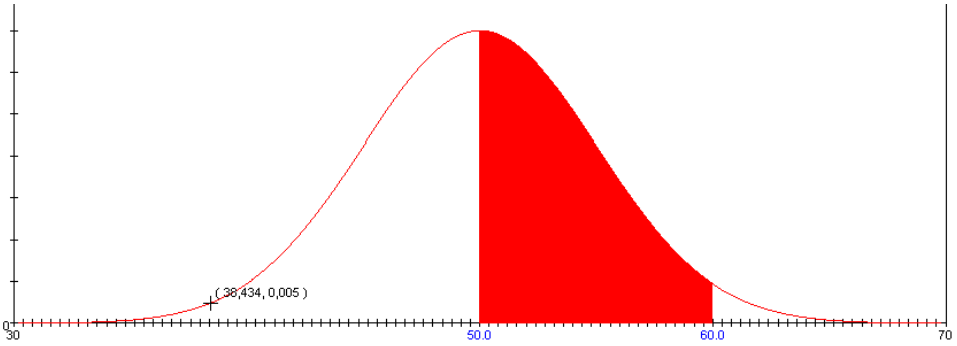
b) Quin percentatge de xiquets tenen una alçada en nàixer superior a 55 cm (àrea roja)?



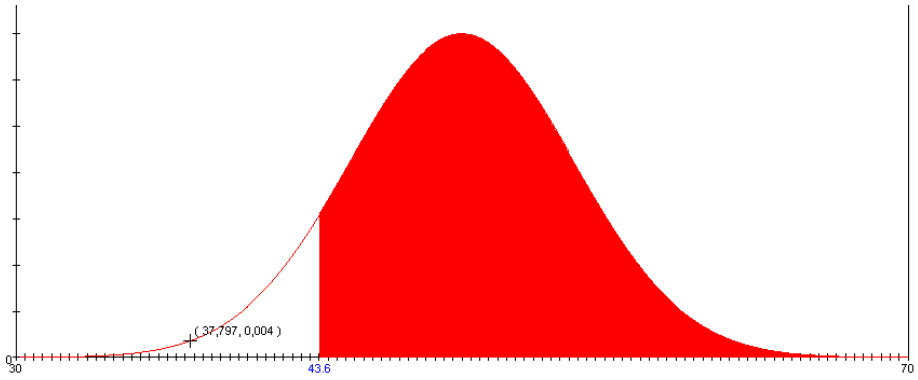
c) Quina és la probabilitat que un xiquet tinga una alçada entre 45 i 55 cm? (àrea roja)



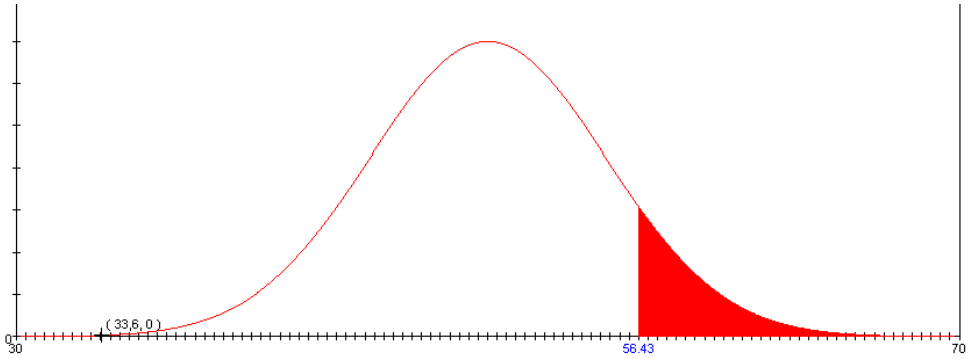
d) Quina és la probabilitat que un xiquet tinga una alçada entre 50 i 60 cm? I que siga superior a 60 cm?



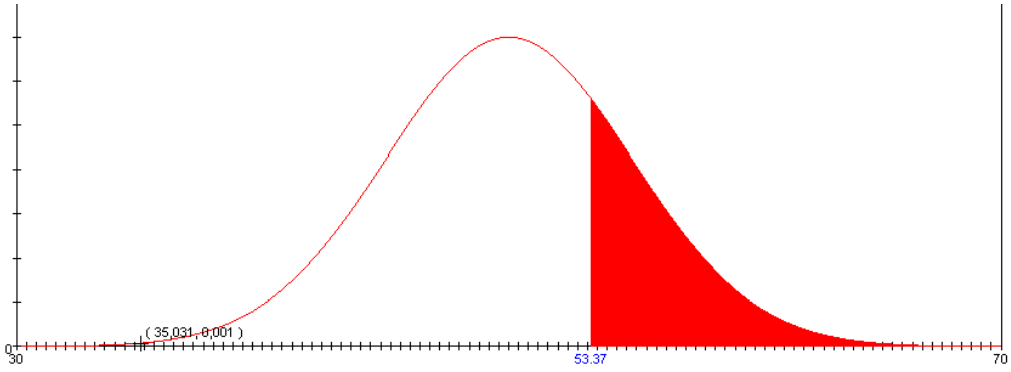
e) El 10 % dels xiquets amb una alçada menor en nàixer tenen una alçada inferior a ___ cm (àrea blanca).



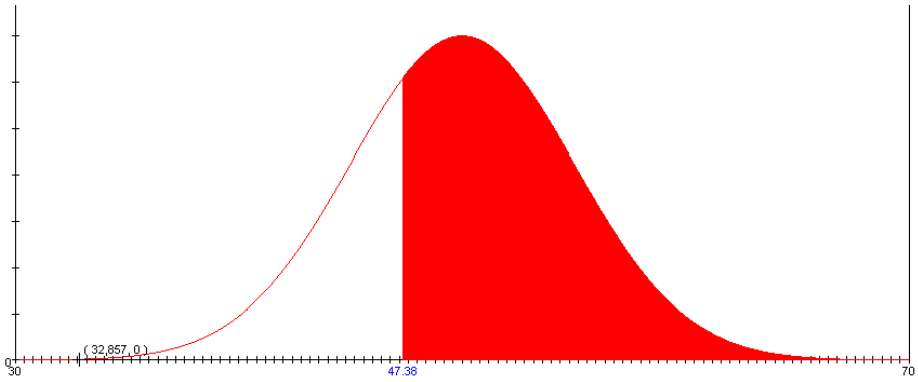
f) El 10 % dels xiquets amb més alçada en nàixer tenen una alçada superior a ___ cm (àrea roja).



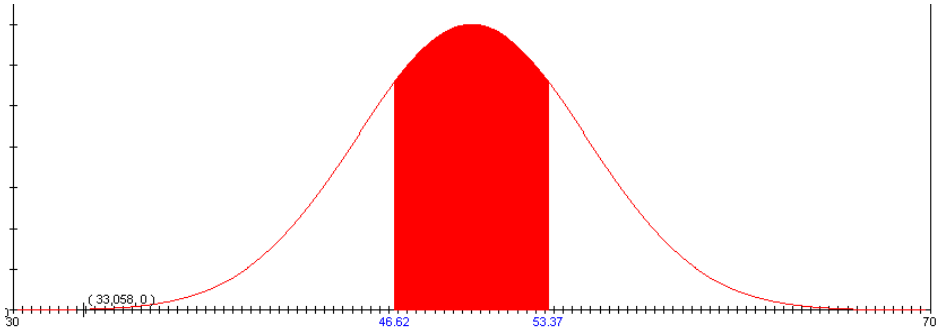
g) El 25 % dels xiquets amb més alçada en nàixer tenen una alçada superior a ___ cm (àrea roja).



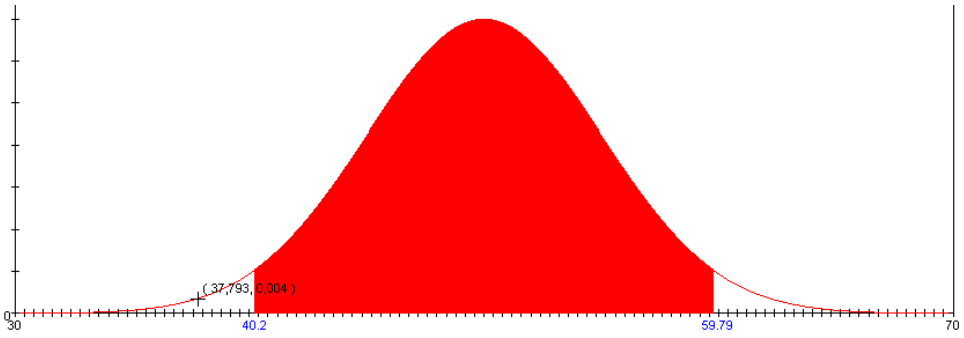
h) A quin valor correspon el percentil 30 de la distribució de l'alçada?



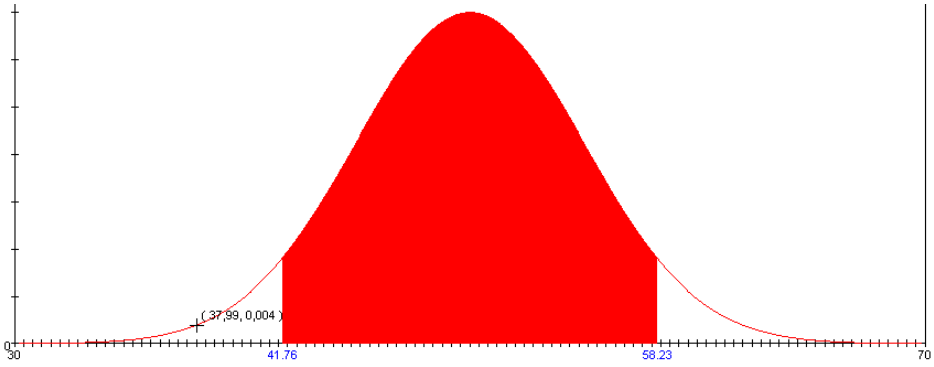
i) El 50 % central dels xiquets tenen una alçada entre ____ i ____ cm.



j) Entre quines puntuacions típiques (z) es troben el 95 % dels xiquets? A quina alçada corresponen aqueixos valors?



k) Entre quines puntuacions típiques (z) es troben el 90 % dels xiquets? A quina alçada corresponen aqueixos valors?



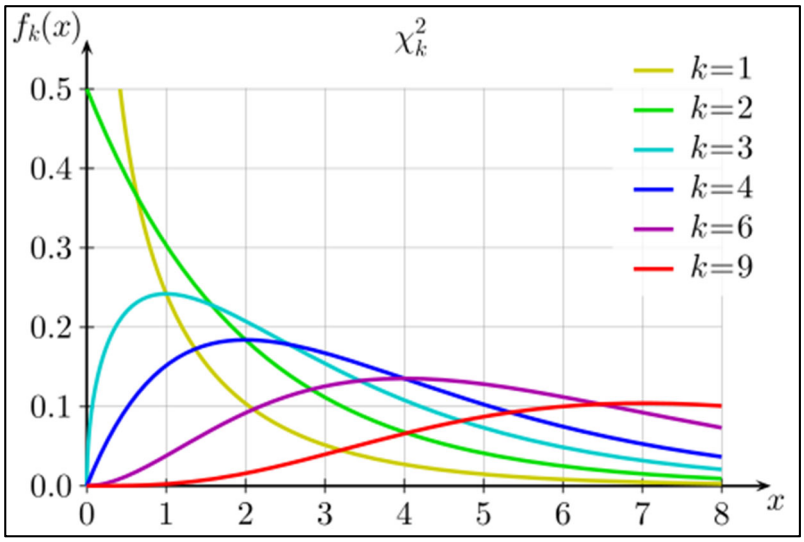
Exercici 7. La variable "Pes" en la població de dones europees és N(65;10).
(a) Quin percentatge de dones pesa menys de 50 quilos? (b) Quina és la probabilitat que una dona d'aquesta població pese més de 70 quilos? (c) Quina és la probabilitat que pese entre 60 i 70 quilos?, (d) El 5 % de les dones amb major pes pesen més de ___ quilos. (e) El 5 % de les dones amb menor pes pesen menys de ___ quilos. (f) Quin és el Q1 d'aquesta distribució? (g) El 95 % central de les dones pesen entre ___ i ___ quilos. (h) El 90 % central de les dones pesen entre ___ i ___ quilos.

Exercici 8: Suposem que un conegut ens diu que ha obtingut en un test d'intel·ligència una puntuació *CI* igual a 95. Assumint que les puntuacions en aquest test d'intel·ligència es distribueixen normalment amb mitjana igual a 100 i desviació típica igual a 15, què li podem dir sobre la seua puntuació?, més concretament: (a) quin percentatge de subjectes és d'esperar que obtinguen un valor inferior o igual a 95?, o (b) quin percentatge de subjectes és d'esperar que obtinguen un valor superior a 95? Suposem també que ens pregunta (c) quina puntuació *CI* caldria traure en el test d'intel·ligència per a estar en el 30 % inferior? (puntuació *CI* que deixa el 30 % de subjectes per davall); (d) i per a estar en el 10 % superior? (puntuació *CI* que és superada només pel 10 % dels subjectes); (e) entre quins valors de *CI* es troben el 50 % central dels subjectes?

3. Les distribucions *khi-quadrat*, *t* i *F*

3.1. La distribució χ^2 (*khi-quadrat*) constitueix, en realitat, una família de distribucions de probabilitat. Cadascun dels membres d'aquesta família ve determinat per un paràmetre *k* que fa referència al nombre de graus de llibertat de la distribució. És usual fer referència a qualsevol membre d'aquesta família de distribucions amb l'expressió χ^2_k, on *k* expressa el nombre de graus de llibertat. Per exemple, χ^2_{18} representa la distribució χ^2 amb 18 graus de llibertat.

• En la figura inferior apareix representada gràficament la funció de densitat de probabilitat de les distribucions χ^2 amb 1, 2, 3, 4, 6 i 9 graus de llibertat. Com pot observar-se, la família de distribucions χ^2 és asimètrica positiva, si bé, a mesura que *k* és major, la distribució tendeix a ser més simètrica. Per a valors de *k* superiors a 50, la distribució χ^2 es pot considerar igual a la distribució normal.



• Els llibres d'estadística solen incloure taules amb informació sobre la distribució de probabilitat acumulada (funció de distribució) de diversos dels membres de la família de distribucions χ^2. Podem obtenir els valors corresponents a la probabilitat acumulada per a qualsevol valor X que desitgem amb el programa MS Excel® introduint en una casella qualsevol la següent expressió amb els valors de χ^2 i els graus de llibertat (k) que ens interesse:

`=1-DISTR.CHI(`***`χ²;k`***`)`

Inversament, si el que es desitja és obtenir, a partir d'un valor de probabilitat acumulada, el valor X de la distribució χ^2 amb k graus de llibertat al qual correspon aqueixa probabilitat acumulada, podem utilitzar la següent funció de MS Excel®:

`=INV.CHICUAD(`***`probabilitat_acumulada;k`***`)`

Per exemple, si escrivim en una casella `=1-DISTR.CHI(4,17;9)` i premem la tecla *Intro*, automàticament obtindrem el valor de probabilitat acumulada corresponent a una puntuació de 4,17 en la distribució χ^2 amb 9 graus de llibertat ($P_a\ (4,17)$ = 0,100).

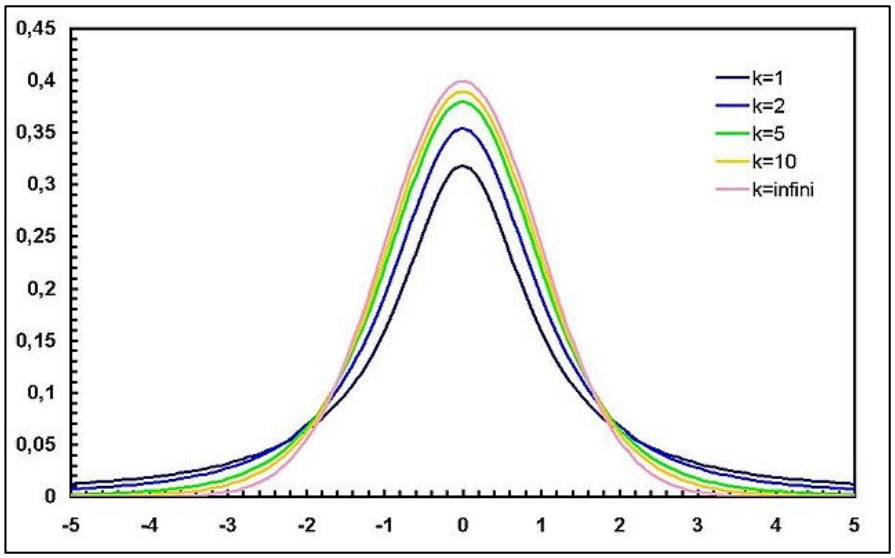
Si el que es desitja és seguir el camí invers, això és, a partir d'un valor de probabilitat acumulada (per exemple, 0,100), obtenir la puntuació X a la qual correspon aqueixa probabilitat acumulada en una determinada distribució (per exemple, la distribució χ^2 amb 9 graus de llibertat), cal escriure en una casella de MS Excel® la següent funció:

`=INV.CHICUAD(0,100;9)`

Exercici 9: Siga X una variable la distribució de la qual és χ^2 amb 7 graus de llibertat. Obtingueu: a) el valor la probabilitat acumulada del qual és igual a 0,20; b) el valor tal que la probabilitat d'obtenir un valor superior a aquest és igual a 0,30; c) la probabilitat d'obtenir un valor igual o inferior a 2,167; d) la probabilitat d'obtenir un valor superior a 16,622; e) la probabilitat d'obtenir un valor entre 4,255 i 9,803.

3.2. La distribució *t* de Student representa també una família de distribucions de probabilitat i, com en la distribució χ^2, cadascun dels membres d'aquesta família ve determinat per un paràmetre *k* que fa referència al nombre de graus de llibertat de la distribució. És habitual fer referència a qualsevol membre d'aquesta família de distribucions amb l'expressió t_k, on *k* expressa el nombre de graus de llibertat. Així, t_{10} representa la distribució *t* amb 10 graus de llibertat.

• En la figura inferior apareix representada la funció de densitat de probabilitat de les distribucions *t* amb 1, 2, 5, 10 i un valor molt gran de graus de llibertat. Com pot observar-se, la família de distribucions *t* és molt similar a la distribució normal, encara que més platicúrtica; no obstant això, a mesura que *k* és major, la distribució *t* tendeix a ser més mesocúrtica i per valors de *k* superiors a 30, les diferències amb la distribució normal són inapreciables.



• Podem obtenir fàcilment amb el programa MS Excel® els valors corresponents a la funció de densitat de probabilitat de qualsevol distribució t_k. Per a això, cal introduir en una casella qualsevol del full de càlcul la següent expressió amb els valors de *t* i *k* que ens interesse:

```
=1-DISTR.T(t;k;1)
```

Si això que es desitja és obtenir el valor de la distribució *t* que té una determinada probabilitat acumulada, la funció de MS Excel® a utilitzar és:

`=INV.T(probabilitat_acumulada;k)`

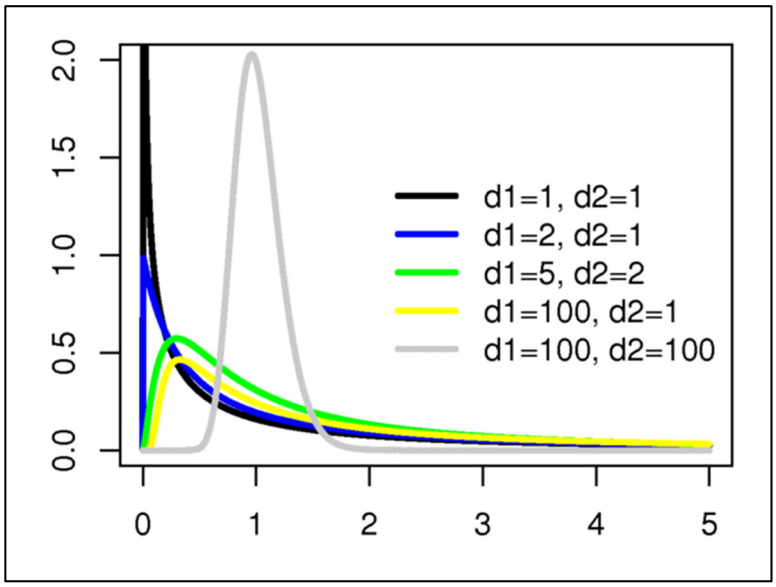
Per exemple, si escrivim en una casella `=1-DISTR.T(1,38;9;1)` i premem la tecla *Intro*, automàticament obtindrem el valor de probabilitat acumulada corresponent a una puntuació de 1,38 en la distribució *t* amb 9 graus de llibertat ($P_a\ (1,38)$ = 0,9).

Si el que es desitja és seguir el camí invers, això és, a partir d'un valor de probabilitat acumulada (per exemple, 0,9), obtenir la puntuació *X* a la qual correspon aqueixa probabilitat acumulada en una determinada distribució (per exemple, la distribució *t* amb 9 graus de llibertat), escriurem en una casella de MS Excel® la següent funció:

`=INV.T(0,9;9)`

Exercici 10: Siga *X* una variable la distribució de la qual és t_{18}. Obtingueu: a) el valor la probabilitat acumulada del qual és igual a 0,25; b) el valor tal que la probabilitat d'obtenir un valor superior a aquest és igual a 0,05; c) la probabilitat d'obtenir un valor inferior a -0,534; d) la probabilitat d'obtenir un valor superior a 1,33; e) la probabilitat d'obtenir un valor entre 0 i 0,688.

3.3. La distribució *F* de Snedecor representa una família de distribucions de probabilitat els membres de la qual venen caracteritzats per dos paràmetres, els coneguts com el nombre de graus de llibertat del numerador (*d1*) i el nombre de graus de llibertat del denominador (*d2*). És habitual fer referència a qualsevol membre d'aquesta família amb l'expressió $F_{d1,d2}$ (per exemple, $F_{5,10}$ representa la distribució *F* amb 5 i 10 graus de llibertat. En la figura inferior apareix representada la funció de densitat de probabilitat de les distribucions $F_{1,1}$, $F_{2,1}$, $F_{5,2}$, $F_{100,1}$ i $F_{100,100}$.



• Podem obtenir amb el programa MS Excel® els valors corresponents a la funció de densitat de probabilitat de qualsevol distribució $F_{d1,d2}$ mitjançant la següent expressió:

`=DISTR.F.N(X;d1;d2;1)`

Si això que es desitja és obtenir el valor de la distribució *F* que té una determinada probabilitat acumulada, la funció de MS Excel® a utilitzar és:

`=INV.F(probabilitat_acumulada;d1;d2)`

Per exemple , si escrivim en una casella `=DISTR.F.N(3,22;6;10;1)` i premem la tecla *Intro*, automàticament obtindrem el valor de probabilitat acumulada corresponent a una puntuació de 3,22 en la distribució *F* amb 6 i 10 graus de llibertat (P_a *(3,22)* = 0,95).

Si el que es desitja és seguir el camí invers, això és, a partir d'un valor de probabilitat acumulada (per exemple, 0,95), obtenir la puntuació *X* a la qual correspon aqueixa probabilitat acumulada en una determinada distribució (per exemple, la distribució *F* amb 6 i 10 graus de llibertat), escriurem la següent funció:

`=INV.F(0,95;6;10)`

Exercici 11: Siga *X* una variable la distribució de la qual és $F_{3,40}$. Obtingueu: a) el valor la probabilitat acumulada del qual és igual a 0,90; b) el valor tal que la probabilitat d'obtenir un valor superior a aquest és igual a 0,05; c) la probabilitat d'obtenir un valor inferior a 1; d) la probabilitat d'obtenir un valor superior a 2,7.

Exercici 12: Exercicis addicionals sobre les distribucions *khi-quadrat*, *t* i *F*:

12.1: Siga X una variable la distribució de la qual és χ^2 amb 7 graus de llibertat. Obtingueu: a) el valor la probabilitat acumulada del qual és igual a 0,10; b) el valor tal que la probabilitat d'obtenir un valor superior a aquest és igual a 0,05; c) la probabilitat d'obtenir un valor igual o inferior a 2,83; d) la probabilitat d'obtenir un valor superior a 18,48; e) la probabilitat d'obtenir un valor superior a 20,28.

12.2: Siga X una variable la distribució de la qual és t_{18}. Obtingueu: a) el valor la probabilitat acumulada del qual és igual a 0,99; b) el valor tal que la probabilitat d'obtenir un valor superior a aquest és igual a 0,05; c) la probabilitat d'obtenir un valor inferior a –1,734; d) la probabilitat d'obtenir un valor superior a 1,33; e) la probabilitat d'obtenir un valor entre 0 i 1,33.

12.3: Siga X una variable la distribució de la qual és $F_{3,40}$. Obtingueu: a) el valor la probabilitat acumulada del qual és igual a 0,99; b) el valor tal que la probabilitat d'obtenir un valor superior a aquest és igual a 0,05; c) la probabilitat d'obtenir un valor superior a 4,31.

Referències

Barón-López, J. (2005). *Bioestadística: métodos y aplicaciones*. Anotacions i material disponibles en http://www.bioestadistica.uma.es/baron/apuntes/

Botella, J., León, O. G., San Martín, R. i Barriopedro, M. I. (2001). *Análisis de datos en psicología I: teoría y ejercicios.* Madrid: Piràmide

Taules

Taula 1: Fragment de la taula de la distribució binomial

n	p r	.01	.05	.10	.15	.20	.25	.30	1/3	.35	.40	.45	.49	.50
2	0	.9801	.9025	.8100	.7225	.6400	.5625	.4900	.4444	.4225	.3600	.3025	.2601	.2500
	1	.0198	.0950	.1800	.2550	.3200	.3750	.4200	.4444	.4550	.4800	.4950	.4998	.5000
	2	.0001	.0025	.0100	.0225	.0400	.0625	.0900	.1111	.1225	.1600	.2025	.2401	.2500
3	0	.9703	.8574	.7290	.6141	.5120	.4219	.3430	.2963	.2746	.2160	.1664	.1327	.1250
	1	.0294	.1354	.2430	.3251	.3840	.4219	.4410	.4444	.4436	.4320	.4084	.3823	.3750
	2	.0003	.0071	.0270	.0574	.0960	.1406	.1890.	.2222	.2389	.2880	.3341	3674	.3750
	3	.0000	.0001	.0010	.0034	.0080	.0156	.0270	.0370	.0429	.0640	.0911	.1176	.1250
4	0	.9606	.8145	.6561	.5220	.4096	.3164	.2401	.1975	.1785	.1296	.0915	.0677	.0625
	1	.0388	.1715	.2916	.3685	.4096	.4219	.4116	.3951	.3845	.3456	.2995	.2600	.2500
	2	.0006	.0135	.0486	.0975	.1636	.2109	.2646	.2963	.3105	.3456	.3675	.3747	.3750
	3	.0000	.0005	.0036	.0115	.0256	.4609	.0756	.0988	.1115	.1536	.2005	.2400	.2500
	4	.0000	.0000	.0001	.0005	.0016	.0039	.0081	.0123	.0150	.0256	.0410	.0576	.0625
5	0	.9510	.7738	.5905	.4437	.3277	.2373	.1681	.1317	.1160	.0778	.0503	.0345	.0312
	1	.0480	.2036	.3280	.3915	.4096	.3855	.3602	.3292	.3124	.2592	.2059	.1657	.1562
	2	.0010	.0214	.0729	.1382	.2048	.2637	.3087	.3292	.3364	.3456	.3369	.3185	.3125
	3	.0000	.0011	.0081	.0244	.0512	.0879	.1323	.1646	.1811	.2304	.2757	.3060	.3125
	4	.0000	.0000	.0004	.0022	.0064	.0146	.0284	.0412	.0488	.0768	.1128	.1470	.1562
	5	.0000	.0000	.0000	.0001	.0003	.0010	.0024	.0041	.0053	.0102	.0185	.0283	.0312
6	0	.9415	.7351	.5314	.3771	.2621	.1780	.1176	.0878	.0754	.0467	.0277	.0176	.0156
	1	.0571	.2321	.3543	.3993	.3932	.3560	.3025	.2634	.2437	.1866	.1359	.1014	.0938
	2	.0014	.0305	.0984	.1762	.2458	.2966	.3241	.3292	.3280	.3110	.2780	.2437	.2344
	3	.0000	.0021	.0146	.0415	.0819	.1318	.1852	.2195	.2355	.2765	.3032	.3121	.3125
	4	.0000	.0001	.0012	.0055	.0154	.0330	.0595	.0823	.0951	.1382	.1861	.2249	.2344
	5	.0000	.0000	.0001	.0004	.0015	.0044	.0102	.0165	.0205	.0369	.0609	.0864	.0938
	6	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0014	.0018	.0041	.0083	.0139	.0156
7	0	.9321	.6983	.4783	.3206	.2097	.1335	.0824	.0585	.0490	.0280	.0152	.0090	.0078
	1	.0659	.2573	.3720	.3960	.3670	.3115	.2471	.2048	.1848	.1306	.0872	.0603	.0574
	2	.0020	.0406	.1240	.2097	.2753	.3115	.3177	.3073	.2985	.2613	.2140	.1740	.1641
	3	.0000	.0036	.0230	.0617	.1147	.1730	.2269	.2561	.2679	.2903	.2918	.2786	.2734
	4	.0000	.0002	.0026	.0109	.0287	.0577	.0972	.1280	.1442	.1935	.2388	.2676	.2734
	5	.0000	.0000	.0002	.0012	.0043	.0115	.0250	.0384	.0466	.0774	.1172	.1543	.1641
	6	.0000	.0000	.0000	.0001	.0004	.0013	.0036	.0064	.0084	.0172	.0320	.0494	.0547
	7	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0006	.0016	.0037	.0068	.0078
8	0	.9227	.6634	.4305	.2725	.1678	.1001	.0576	.0390	.0319	.0168	.0084	.0046	.0039
	1	.0746	.2793	.3826	.3847	.3355	.2670	.1977	.1561	.1373	.0896	.0548	.0352	.0312
	2	.0026	.0515	.1488	.2376	.2936	.3115	.2965	.2731	.2587	.2090	.1569	.1183	.1094
	3	.0001	.0054	.0331	.0839	.1468	.2076	.2541	.2731	.2786	.2787	.2568	.2273	.2188
	4	.0000	.0004	.0046	.0185	.0459	.0865	.1361	.1707	.1875	.2322	.2627	.2730	.2734
	5	.0000	.0000	.0004	.0026	.0092	.0231	.0467	.0683	.0808	.1239	.1719	.2098	.2188
	6	.0000	.0000	.0000	.0002	.0011	.0038	.0100	.0171	.0217	.0413	.0703	.1008	.1094
	7	.0000	.0000	.0000	.0000	.0001	.0004	.0012	.0024	.0033	.0079	.0164	.0277	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0002	.0007	.0017	.0033	.0039

Taula II: Probabilitats acumulades corresponents a la corba normal estàndard [P_a (X)].

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998