

ESTADÍSTICA APLICADA A L'EDUCACIÓ

Mercedes Torrado i Ruth Vilà (coords.)

Vanesa Berlanga
Jordi del Barrio
Francesc Martínez
M.^a Isabel Mateo
Javier Ormazábal
M.^a José Rubio
Antoni Ruiz
Robert G. Valls

Departament de Mètodes d'Investigació i Diagnòstic en l'Educació

ESTADÍSTICA APLICADA A L'EDUCACIÓ

Mercedes Torrado i Ruth Vilà (coords.)

Vanessa Berlanga
Jordi del Barrio
Francesc Martínez
M.^a Isabel Mateo
Javier Ormazábal
M.^a José Rubio
Antoni Ruiz
Robert G. Valls

Departament de Mètodes d'Investigació
i Diagnòstic en Educació

Universitat de Barcelona. Dades catalogràfiques

Estadística aplicada a l'educació. – (Textos docents ; 390)

ISBN 978-84-475-3817-1

I. Torrado Fonseca, Mercè II. Vilà Baños, Ruth III. Col·lecció:
Textos docents (Universitat de Barcelona ; 390)
1. Estadística educativa

© Publicacions i Edicions de la Universitat de Barcelona
Adolf Florensa, s/n
08028 Barcelona
Tel.: 934 035 430
Fax: 934 035 531
www.publicacions.ub.edu
comercial.edicions@ub.edu

ISBN	978-84-475-3817-1
DIPÒSIT LEGAL	B-16.433-2014
IMPRESSIÓ I ENQUADERNACIÓ	Gráficas Rey

És rigorosament prohibida la reproducció total o parcial d'aquesta obra. Cap part d'aquesta publicació, inclòs el disseny de la coberta, no pot ser reproduïda, emmagatzemada, transmesa o utilitzada per cap mitjà o sistema, sense l'autorització prèvia per escrit de l'editor.

Índex

Capítol 1

INTRODUCCIÓ

1.1. Conceptes bàsics d'estadística	9
1.1.1. L'estadística	9
1.1.2. Escales de mesura	9
1.1.3. Les variables	11
1.1.4. Individu, població i mostra	12
1.1.5. Paràmetre i estadístic	12
1.2. Funcionament bàsic de l'SPSS	13
1.2.1. Introducció a l'SPSS	13
1.2.2. Creació d'una matriu de dades	16
1.2.3. Funcions bàsiques	19
1.2.4. Opcions d'ajuda a l'SPSS	23
1.3. Exercicis resolts	25

Capítol 2

ESTADÍSTICA DESCRIPTIVA

2.1. Tabulació i representació gràfica	27
2.1.1. La tabulació	27
2.1.2. Representació gràfica	34
2.2. Indicadors estadístics	39
2.2.1. Indicadors de posició	40
2.2.2. Indicadors de tendència central	42
2.2.3. Indicadors de dispersió	45
2.2.4. Indicadors de forma	48
2.2.5. Els indicadors estadístics a l'SPSS	50
2.2.6. Com descriure una variable utilitzant els indicadors estadístics	52
2.3. Exercicis resolts	53

Capítol 3

PROBABILITAT I ESTADÍSTICA

3.1. Conceptes bàsics	55
3.1.1. Noció intuïtiva de probabilitat	55
3.1.2. Noció clàssica de probabilitat	55
3.1.3. Noció axiomàtica de probabilitat	56
3.1.4. Noció experimental de probabilitat	57
3.2. Llei de probabilitat de variables aleatòries discretes	58
3.3. Llei de probabilitat de variables aleatòries contínues	60
3.3.1. La Llei Normal	61
3.3.2. Proves de normalitat (Kolmogorov)	64
3.4. Taules	67
3.5. Exercicis resolts	70

Capítol 4

ESTADÍSTICA INFERENCIAL

4.1. Conceptes bàsics	73
4.2. Teoria de l'estimació estadística	76
4.2.1. Interval de probabilitat	78
4.2.2. Interval de confiança	79
4.2.3. Grandària de la mostra	82
4.3. Teoria de la decisió estadística	83
4.3.1. Errors i riscos	84
4.3.2. Potència d'un contrast	84
4.3.3. Grau de significació	84
4.3.4. Contrasts unilaterals i bilaterals	85
4.4. Exercicis resolts	86

Capítol 5

CORRELACIÓ I REGRESSIÓ

5.1. Conceptes bàsics	89
5.2. Correlació i regressió lineal simple	90
5.2.1. Correlació de Pearson (r_{xy})	92
5.2.2. Altres coeficients de correlació simple	97
5.3. Correlació parcial i regressió lineal múltiple	100
5.4. Exercicis resolts	106

Capítol 6

PROVES PARAMÈTRIQUES I NO PARAMÈTRIQUES

6.1. Condicions de parametricitat	111
6.2. Selecció de la mostra	112
6.3. Classificació de proves paramètriques i no paramètriques	113
6.3.1. Proves paramètriques	113
6.3.2. Proves no paramètriques	115
6.4. Exercicis resolts	119

REFERÈNCIES BIBLIOGRÀFIQUES	127
-----------------------------------	-----

Aquest és un material docent guia per iniciar-se en l'estadística aplicada en l'àmbit socioeducatiu. L'obra no pretén ser un manual, sinó, més aviat, una guia útil de suport en l'estudi i de resolució de dubtes més comuns i bàsics en investigadors novells. Per això, s'ha utilitzat un llenguatge directe i alhora rigorós que facilita la comprensió dels conceptes bàsics de l'anàlisi quantitativa de dades.

Al llarg dels sis capítols de l'obra es fa un recorregut que va des dels conceptes bàsics d'estadística i del programa SPSS, passant per l'estadística descriptiva, fins a la fonamentació de la inferència estadística, anàlisis correlacionals, i arribant fins i tot a proves de contrast de mitjanes i proporcions, paramètriques i no paramètriques. Es combina en tots ells l'explicació teòrica emmarcada amb exemples del context socioeducatiu i una aplicació directa amb l'SPSS fent èmfasi especial en la interpretació dels *outputs* i obviant els càlculs manuals.

Amb l'objectiu d'acompanyar el lector en el seu procés d'aprenentatge, al final de cada capítol hi ha tot un seguit d'exemples d'exercicis resolts a partir d'*outputs* de l'SPSS.

INTRODUCCIÓ

En aquest primer capítol es fa una aproximació a l'estadística definint els conceptes clau que la caracteritzen i que permetran abordar els propers capítols amb una perspectiva cap a la investigació quantitativa i l'anàlisi estadística de les dades més amplia; partint de l'estadística descriptiva per donar pas a les proves pròpies de l'estadística inferencial. Es veuran les diferents escales de mesura que permeten descriure les variables que s'utilitzen en les anàlisis de les investigacions.

En el segon apartat, **Funcionament bàsic de l'SPSS**, es troba una introducció al paquet estadístic SPSS (Statistical Package for the Social Sciences) per tal que es pugui iniciar un apropament a les funcions que aquesta eina d'anàlisi estadística de dades quantitatives ens ofereix. En posteriors capítols, es desenvoluparan més concretament els resultats que podem extreure d'aquest programa imprescindible per a les investigacions quantitatives en ciències socials.

Al final del capítol, es troben un seguit d'exercicis resolts que podeu utilitzar per acabar de revisar els coneixements adquirits en aquest primer bloc del temari.

1.1. Conceptes bàsics d'estadística

1.1.1. L'estadística

Definim l'estadística com la ciència que té per objecte l'estudi quantitatiu dels col·lectius. Dins l'àmbit educatiu, facilita el còmput i enumera els fets i els individus susceptibles d'enumerar-se o de mesurar-se, coordina i classifica les dades obtingudes amb l'objecte de determinar-ne les causes, les conseqüències i les tendències.

Quan parlem d'estadística, ho solem fer diferenciant dos grans blocs:

- **L'estadística descriptiva** és aquella que ens diu com és una mostra, mitjançant valors estadístics de les dades recollides, i obté informació sobre la freqüència, la posició i la variabilitat dels valors recollits.
- **L'estadística inferencial** permet fer inferències de les dades obtingudes a la mostra i generalitzar els resultats a la població origen o a poblacions similars. A partir de l'anàlisi de la mostra, es dedueixen propietats de la població.

1.1.2. Escales de mesura

La *teoria de mesura*, en general i de cada àmbit d'aplicació en particular, no és un problema que l'estadística ha de dilucidar ni resoldre; més aviat, és propi d'altres disciplines com la metrologia o la psicometria. No obstant, convé recordar alguns conceptes respecte a això.

Tot i que no és per unanimitat, es reconeix i s'accepta la classificació de Stevens sobre l'existència de quatre escales de mesura: **nominal**, **ordinal**, **d'interval** i **de raó o proporció**. Per a una millor comprensió, les podem referir o aplicar a un mateix fenomen que accepti

la mesura amb les quatre escales, com, per exemple, *la temperatura d'un conjunt de trenta elements*.

- **Escala nominal:** és aquella que utilitza categories de tipus qualitatiu. Alguns exemples de variables mesurades amb aquesta escala serien: religió, partit polític, sexe, color del cabell, etc. A cadascuna de les categories se li assignen atributs diferents (per exemple, nombres, noms, etc.) sense afirmar que un sigui superior a l'altre, només diferents. És a dir, que permet classificar els objectes o individus segons que siguin iguals o no respecte a una característica. És el nivell més simple i permet la classificació dels individus o objectes en categories merament descriptives, exhaustives i mútuament excloents.

Podem completar-la afegint una categoria d'*altres* o *varis*, sempre que no sigui possible establir a priori una classificació completa i compregui un petit percentatge.

D'altra banda, podem designar les categories amb números (1, 2, 3, etc.). En aquest cas, els números no indiquen valors quantitatius, sinó que són signes o codis d'identificació.

Amb el nostre **exemple:** podem mesurar la temperatura dels trenta elements, per exemple, en graus centígrads, Kelvin i Fahrenheit. Aquestes categories compleixen les condicions següents:

- Han d'estar clarament definides, sense equívocs ni interpretacions diferents.
 - Han d'ésser mútuament excloents, per evitar que algú pugui estar classificat en més d'una categoria.
 - La classificació ha d'ésser completa: tots els subjectes han de poder ser classificats en alguna categoria.
-
- **Escala ordinal:** és aquella que indica el major o menor grau quant a la classificació dels individus segons un criteri creixent o decreixent de la característica que es vol mesurar. D'altra banda, aquest tipus de mesura no ofereix cap informació sobre la magnitud de les diferències entre categories. No sabem la diferència numèrica entre les categories (alt/mitjà/baix); només sabem l'ordre que tenen (alt > mitjà > baix). Exemples de variables que poden mesurar-se amb aquesta escala: nivell econòmic, ordre d'arribada a l'escola, classe social, etc.
- Amb el nostre **exemple:** podem mesurar la temperatura de més a menys, o de menys a més. Proporciona més informació que l'escala nominal, ja que expressa la situació de cada element dins del conjunt; no obstant això, no implica que existeixi la mateixa diferència de temperatura entre dos elements consecutius, és a dir, entre el 1r i el 2n no cal que hi hagi la mateixa diferència que entre el 16è i el 17è.

- **Escala d'interval:** allò que caracteritza aquesta escala és l'existència d'una unitat de mesura constant que permet assignar un valor que possibilita interpretar la diferència entre dues mesures. Aquest nivell no només classifica i ordena, sinó que, a més, especifica les distàncies existents entre les diferents categories. En aquesta escala, *tant el zero com la unitat de mesura són arbitraris*, de manera que el valor zero no implica l'absència de la característica mesurada; la diferència entre els valors és constant, però no proporcional.

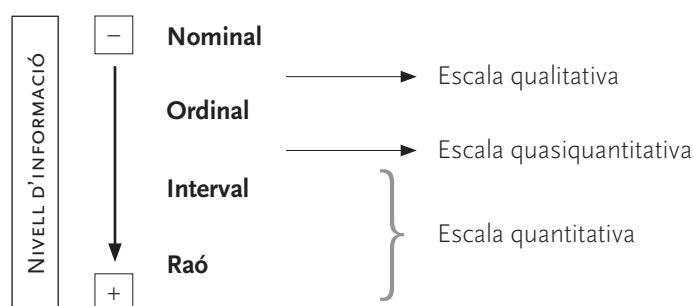
Exemples: la temperatura (0° no és absència de temperatura, ni 20° és el doble de calor que 10°), rendiment acadèmic (0 no és absència de coneixement, ni qui treu un 8 sap el doble de qui treu un 4), etc.

Amb el nostre **exemple:** l'escala Celsius constitueix una escala d'interval i indica la quantitat de graus centígrads (70, 67, 65, 64, ..., 35, ..., etc.) de cadascun dels elements. El zero no és absolut, sinó convencional (en el cas de l'escala Celsius, correspon a la

temperatura de fusió de l'aigua). Això vol dir que podem afirmar que entre 70° y 67° existeix la mateixa diferència de temperatura que entre 38° i 35°, però no que un element que té 70° estigui el doble de calent que el que en té 35°.

- **Escala de raó:** és similar a la d'interval, però a més se li pot assignar *el valor zero absolut* que representa l'absència total de la característica que es mesura i la proporcionalitat entre els valors. En ciències socials és molt difícil trobar casos en què es pugui aplicar. Exemples: pes (0 kg és absència de pes i 20 kg és el doble de pes que 10 kg), edat, longitud, etc.

El nostre **exemple:** En el cas de la temperatura, correspondria a l'escala de Kelvin (en què el «zero» és absolut i equival a -273 de l'escala centígrada).



Quan treballem amb el paquet estadístic SPSS, distingirem entre escales de mesura nominal, ordinal i d'escala, aquesta darrera fent referència tant a la d'interval com a la de raó.



1.1.3. Les variables

La *variable* és aquella propietat, fenomen o característica que varia i adopta diferents categories, ordres o valors numèrics. És qualsevol cosa que es pugui mesurar directament en un individu, com el pes, l'alçada, el lloc de naixement, etc.

Les variables se solen classificar segons diferents criteris. Des del punt de vista estadístic, podem distingir:

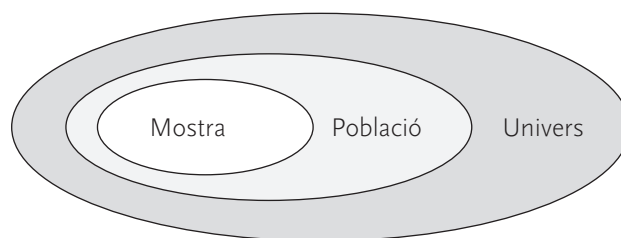
- **Variables qualitatives:** són les que estan mesurades amb categories de tipus qualitatiu (escala nominal). Són resultat de l'observació d'una qualitat i la seva naturalesa és no quantificable (sexe, estat civil, etc.). Les variacions de la variable reben el nom de *categories*. Segons el nombre de categories que presenta, es divideix en dicotòmica i politòmica.
 - **Variables qualitatives dicotòmiques:** les que tenen dues categories, ja siguin naturals (femení/masculí) o fruit d'una frontera arbitrària (nadiu/estranger).
 - **Variables qualitatives politòmiques:** quan prenen més de dues categories (bé/regular/malament). Per exemple: la professió, els estudis universitaris, etc.

- **Variables quantitatives:** són el resultat d'una mesura d'interval o de raó (estatura, edat, temps de resposta, etc.). Les diferents variacions de la variable s'anomenen *valors* de la variable.
 - **Variables quantitatives discontinües (discretes):** són variables que només poden prendre un nombre concret de valors entre els que constitueixen el seu abast. Per la seva naturalesa intrínseca només poden adoptar valors sencers. Per exemple: nombre de fills, llibres editats, subjectes a classe, etc.
 - **Variables quantitatives contínues:** les variables tenen una variabilitat que pot prendre qualsevol dels valors possibles entre els que constitueixen el seu abast. Parlem, per tant, d'un contínuum (valors sencers o fraccionats). Per exemple: pes, temps, edat, alçada, etc.

1.1.4. Individu, població i mostra

Delimitar els subjectes de la investigació ens permetrà concretar quins són els individus a partir dels quals es vol obtenir la informació. Aquí es presenta des del nivell superior, definint l'univers d'estudi, la població, la mostra i, finalment, el subjecte d'investigació o individu.

- **Univers:** conjunt d'individus que presenten la característica que ens interessa estudiar (exemple: l'alumnat de la secundària obligatòria).
- **Població:** dins d'aquest univers, el conjunt d'individus sobre els quals volem fer l'estudi (exemple: l'alumnat de l'ESO de Catalunya).
- **Mostra:** subconjunt d'individus de la població extret i seleccionat per algun mètode de mostreig. És el nivell en què realment fem l'estudi i obtenim les dades per a l'anàlisi. Treballem amb la mostra per tal de poder extrapolar els resultats a la població, sempre que la mostra sigui representativa (exemple: 1.000 alumnes de l'ESO dels diferents instituts catalans).
- **Individu:** element ben definit que porta associat un judici, un ordre o una quantitat (segons l'escala) que presenta la característica que volem estudiar. Poden ser persones, objectes o col·lectius (exemple: cada un dels alumnes de l'ESO).



1.1.5. Paràmetre i estadístic

Segons el tipus d'anàlisi que es vulgui realitzar, utilitzarem o bé estadístics, si els valors que obtenim fan referència a la mostra, o bé paràmetres, si ens refereixen valors de la població.

- **Paràmetres:** són els valors obtinguts mitjançant l'estadística que caracteritzen la població. Normalment, se simbolitzen mitjançant lletres gregues: μ , σ , etc.
- **Estadístics:** són els valors obtinguts mitjançant l'estadística que descriuen el fenomen en la mostra. Se solen simbolitzar mitjançant lletres llatines: $\bar{X}$, etc.

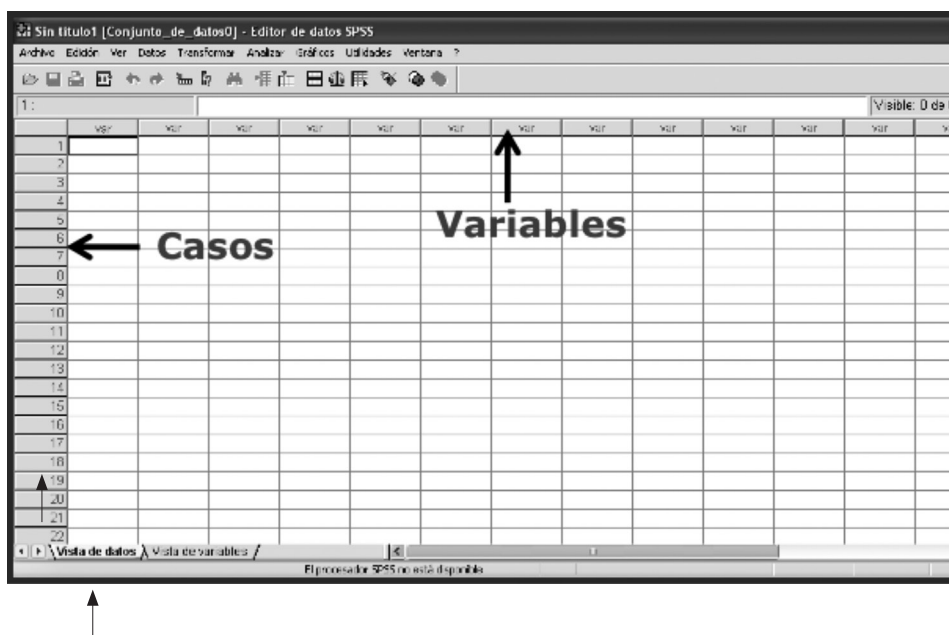
1.2. Funcionament bàsic de l'SPSS

1.2.1. Introducció a l'SPSS

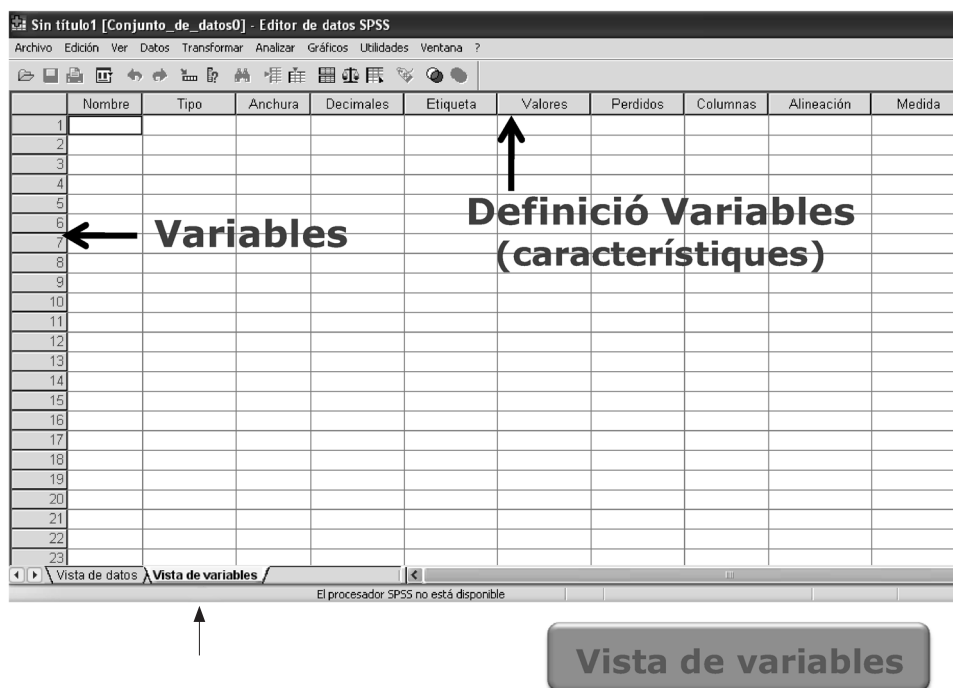
L'SPSS (Statistical Package for Social Sciences) és un paquet de programes per tal de dur a terme l'anàlisi estadística de les dades. Constitueix una aplicació estadística força potent, de la qual s'han anat desenvolupant diverses versions des del seu inici, als anys setanta. No obstant això, tot i que la forma ha anat variant, pel que fa al seu funcionament és molt similar entre les diferents versions.

Abans d'iniciar-nos en la utilització de les aplicacions de l'SPSS és important familiaritzar-nos amb algunes de les finestres que més farem servir. En entrar a l'SPSS, el primer que ens trobem és l'*editor de dades*, tal com es representa a la imatge posterior. Aquesta finestra visualitza, bàsicament, les dades que anirem introduint.

L'*editor de dades* inclou dues opcions: la *vista de les dades* i la *de les variables*. Aquestes opcions es poden seleccionar a partir de les dues pestanyes que es presenten a la part inferior. A la imatge posterior trobem activada la *vista de dades*, que conté el menú general i la matriu de dades. Aquesta matriu està estructurada ubicant els casos a les files i les variables a les columnes.



Per a definir les variables, s'utilitza la *vista de variables* tot clicant sobre la pestanya corresponent tal com presentem a la figura següent:



La vista de variables ens permet fer especificacions sobre les variables que conté la nostra matriu de dades. Com, per exemple, el nom de la variable, el nombre de decimals que té, l'escala de mesura, etc.

En aquesta finestra, les diferents variables es troben situades a les files, mentre que les característiques a definir es troben a les columnes.

Un altre tipus de finestra que s'utilitza amb freqüència és el *visor de resultats*. Aquest visor s'activa automàticament en demanar qualsevol tipus d'anàlisi (taules, gràfics, estadístics, etc.). Les operacions que es demanen es recullen de forma contínua en la finestra del visor de resultats, com podem veure en la figura.

Sexo

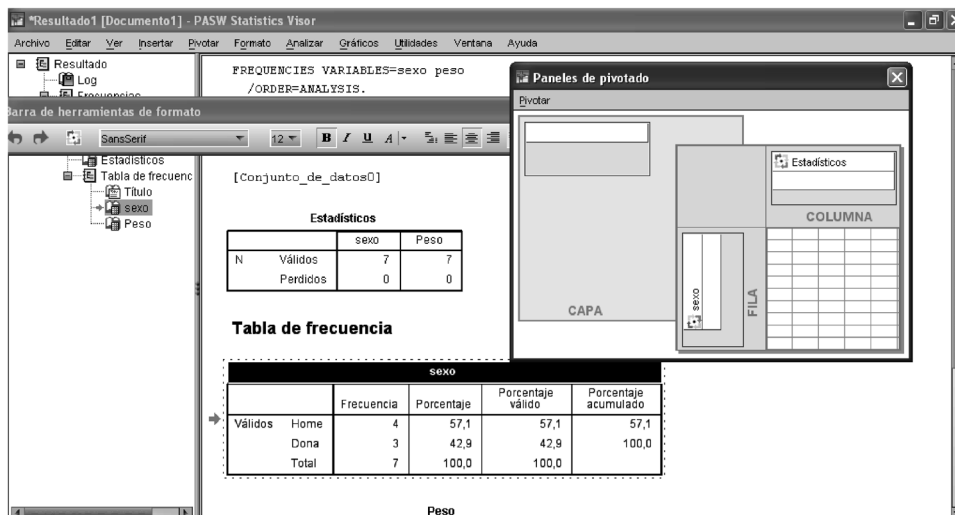
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Hombre	1	33,3	50,0	50,0
	Mujer	1	33,3	50,0	100,0
	Total	2	66,7	100,0	
Perdidos	Sistema	1	33,3		
Total		3	100,0		

Año en que terminó sus estudios

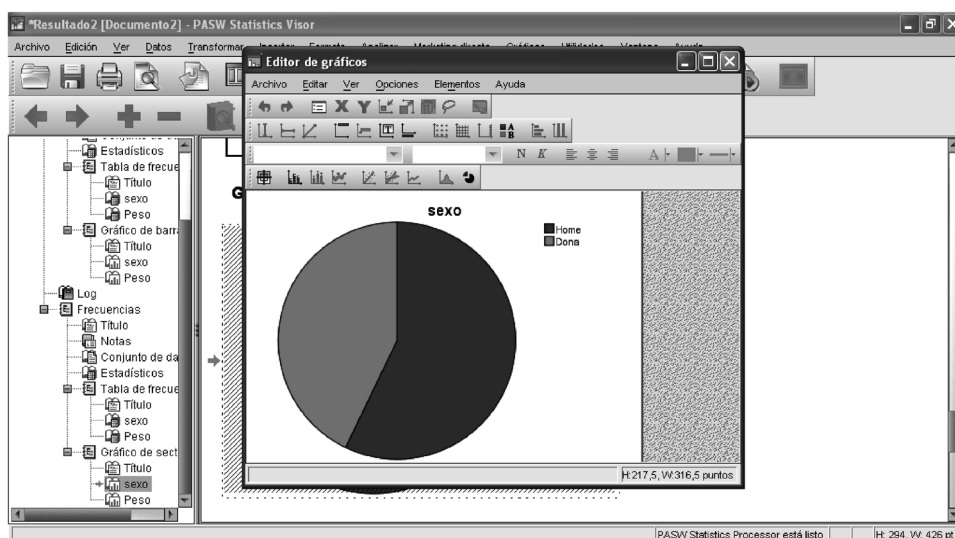
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1997	1	33,3	50,0	50,0
	1998	1	33,3	50,0	100,0
	Total	2	66,7	100,0	
Perdidos	Sistema	1	33,3		
Total		3	100,0		

La finestra de visor de resultats es troba dividida en dos compartiments: el navegador de resultats i els resultats o *outputs*. Aquest navegador ens permet desplaçar-nos d'una manera ràpida i eficaç per aquest recull de resultats continu. Des d'aquest compartiment, trobem un índex de tot el que s'ha anat adjuntant al visor, i se'ns permet, entre altres coses, esborrar, minimitzar o seleccionar elements concrets dels resultats.

D'altra banda, per editar els resultats obtinguts podem fer un doble clic en el resultat concret (taula o gràfic) i això obrirà automàticament un nou menú. Concretament, si desitem editar una taula, ens trobarem amb la **finestra de l'editor de taules**. Aquest editor obre un nou menú que conté opcions com editar texts, intercanviar files i columnes (pivotar) i diverses alternatives de format per a la taula i el text.



Finalment, si fem doble clic en un gràfic, entrarem en l'**editor de gràfics**, com podem observar en la figura següent. Aquest editor obre un nou menú que permet realitzar modificacions en els gràfics, com per exemple en els colors, les fonts, els eixos, les etiquetes i algunes opcions alternatives.



Aquestes diferents finestres corresponen també a diversos tipus d'arxius. Concretament, des de l'editor de dades, obtenim *arxius de dades* que contenen tant les dades introduïdes

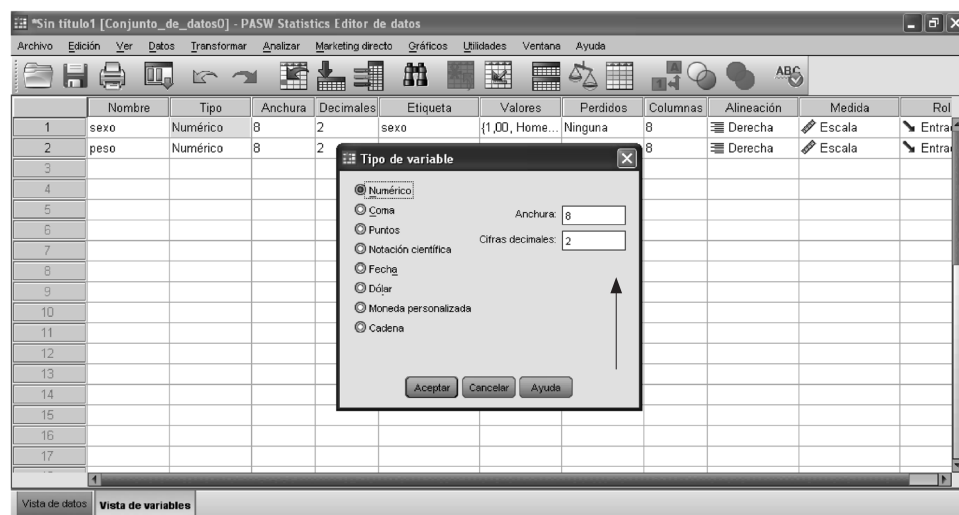
com tota la informació sobre les variables. L'SPSS identifica aquests fitxers amb l'**extensió SAV** (nom del fitxer .SAV). En canvi, des de l'editor de resultats, obtenim *arxius de resultats* que tenen com a **extensió SPV** (nom del fitxer .SPV).

1.2.2. Creació d'una matriu de dades

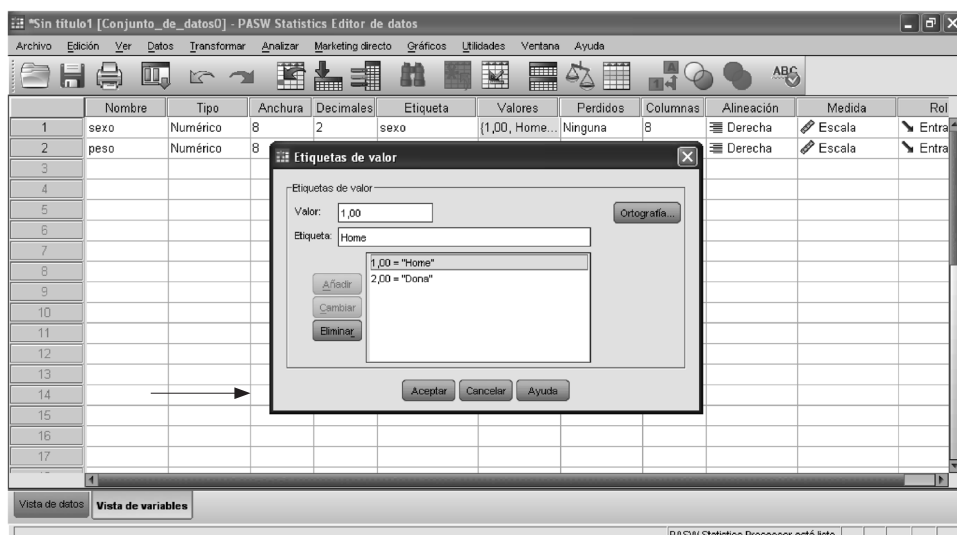
Crear una matriu de dades suposa la definició de les variables de l'estudi, així com la introducció de les dades obtingudes. Per definir les variables, hem d'accedir al visor de variables dins de la finestra de l'editor de dades.

Concretament, podem definir els aspectes següents de cada una de les variables de l'estudi:

- El **nom de la variable**, l'utilitzarem per visualitzar-lo en la matriu i, per tant, ens interessa un nom curt que ens identifiqi cada una de les variables.
- El **tipus de variable** ens permet distingir si la variable és numèrica o alfanumèrica.

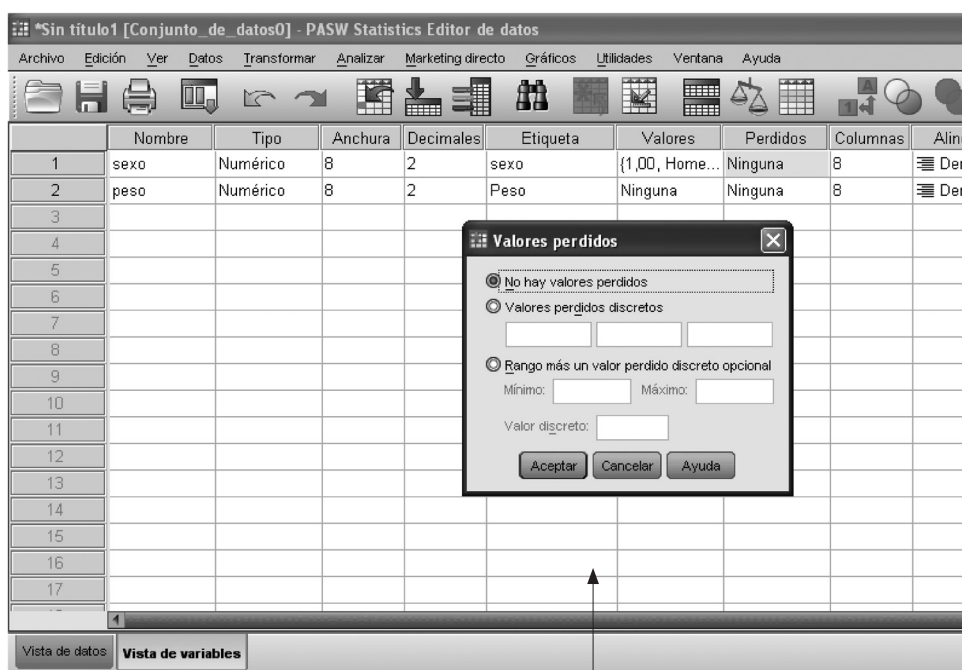


- A l'**amplada** de la variable podem definir el nombre de dígitos que inclouen aquesta variable en la matriu de dades.
- Els **decimals** suposen definir el nombre de decimals que conté la variable.
- A l'**etiqueta de la variable** podem proporcionar un nom més llarg i específic que identifiqi el contingut de la variable que generem. Aquesta etiqueta és la que apareixerà en les sortides de resultats. En cas de no definir-la, s'utilitza per defecte el *nom de la variable*.
- Els **valors** de la variable s'utilitzen per a aquelles variables qualitatives que disposen de categories. Per exemple, el sexe té les categories «home» i «dona». En aquest cas, cliquem l'opció de valors, tal com s'indica a la figura de la pàgina següent, i en el quadre de diàleg que sorgeix podem especificar els *valors* numèrics i la seva corresponent *etiqueta de valor*. Després de cada una de les especificacions és necessari prémer *afegir*.



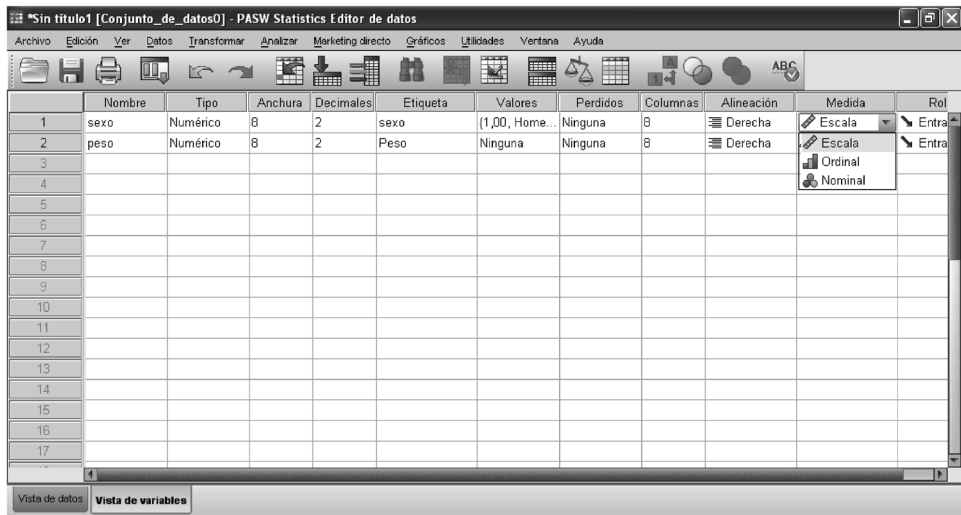
- Els valors *perduts* són aquells que no es tenen en compte per a l'anàlisi dels resultats. L'SPSS n'identifica dos tipus: els perduts pel sistema (qualsevol casella en blanc dins de la matriu de dades) i els perduts per l'usuari. En aquest darrer cas, és l'usuari qui defineix quins són aquests valors; ho fa clicant a l'opció de valors perduts tal com es mostra a la imatge posterior.

En el quadre de diàleg que apareix, podem escollir entre diferents opcions: definir valors perduts que no siguin consecutius (valors perduts discrets), o bé concretar un rang de valors perduts més un valor discret opcional.



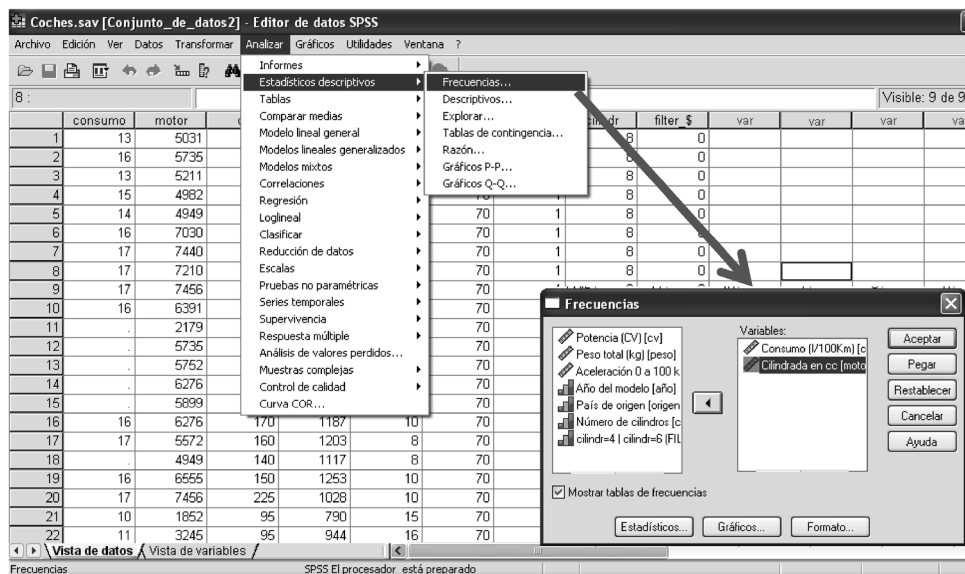
- L'opció **columna** ens permet definir o modificar l'amplada que ocupa la columna per a cada variable dins la vista de dades i no afecta en cap cas l'amplada definida prèviament.
- L'**alineació** fa referència a com se situen les dades a la columna corresponent dins la vista de dades (a la dreta, a l'esquerra o centrades).

- Finalment, hem d'especificar el tipus *d'escala de mesura* que farem servir per a cada variable. Si marquem sobre l'opció *mesura*, apareix un menú en el qual podem escollir l'escala corresponent: escala, ordinal o nominal.



Un cop hem definit totes les característiques de les variables de l'estudi podem començar a introduir les dades dins l'editor de dades. Abans que comencem a analitzar les dades i a extreure resultats, és important que procedim a la *depuració de la matriu*, ja que aquest procediment ens permetrà detectar si hem comès alguna errada en introduir les dades i poder arreglar-la. Un exemple d'això seria el fet que trobéssim un 3 dins les respostes de la variable grup, essent les possibles: 1= matí i 2=tarda.

Per tal de dur a terme aquesta depuració, el que farem serà demanar a l'SPSS que ens generi les *taules de freqüències* de cada una de les variables per a poder comprovar que no més inclouen els codis pertinents. Ho farem seguint les passes marcades a la figura de sota, clicant l'opció *freqüències* dins les diferents possibilitats que ofereix l'anàlisi dels *estadístics descriptius*. Un cop clicada aquesta opció, ens apareix un quadre de diàleg en què hem d'especificar les variables de les quals demanem la taula de freqüències, havent de demanar-la de totes per tal de dur a terme l'esmentada depuració.

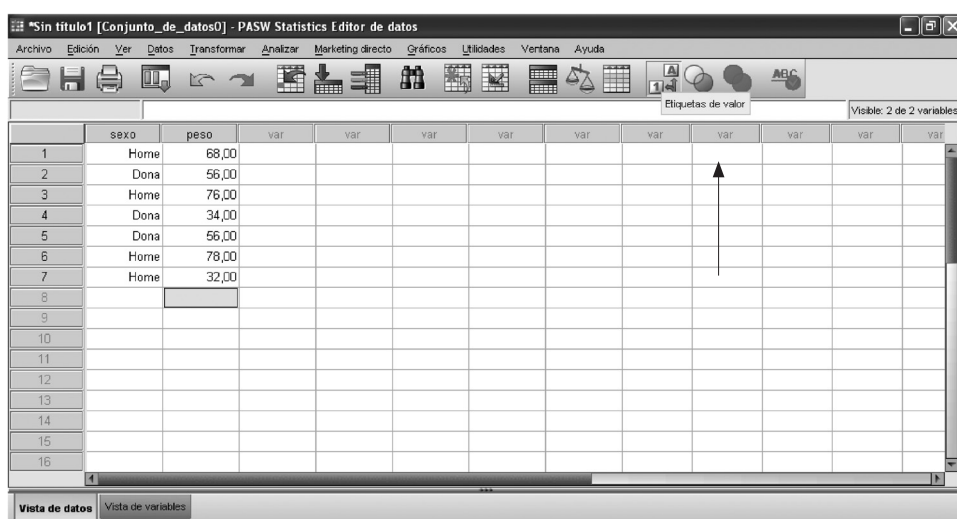


Després de demanar les taules de freqüències, ens aniran apareixent al visor de resultats i haurem de revisar que no hi hagi cap codi incorrecte, localitzant, sempre que n'hi hagi, en quin cas es troba l'errada per així poder modificar la resposta errònia per la correcta.

1.2.3. Funcions bàsiques

Dins les funcions bàsiques que treballarem en aquest apartat, hem de remarcar la que ens permet visualitzar, dins la vista de dades de l'editor de dades, les *etiquetes dels valors de les variables*; és a dir, que en comptes d'aparèixer els codis definits per a les diferents categories de les variables el que ens apareixerà quan cliquem la icona assenyalada, seran les etiquetes corresponents.

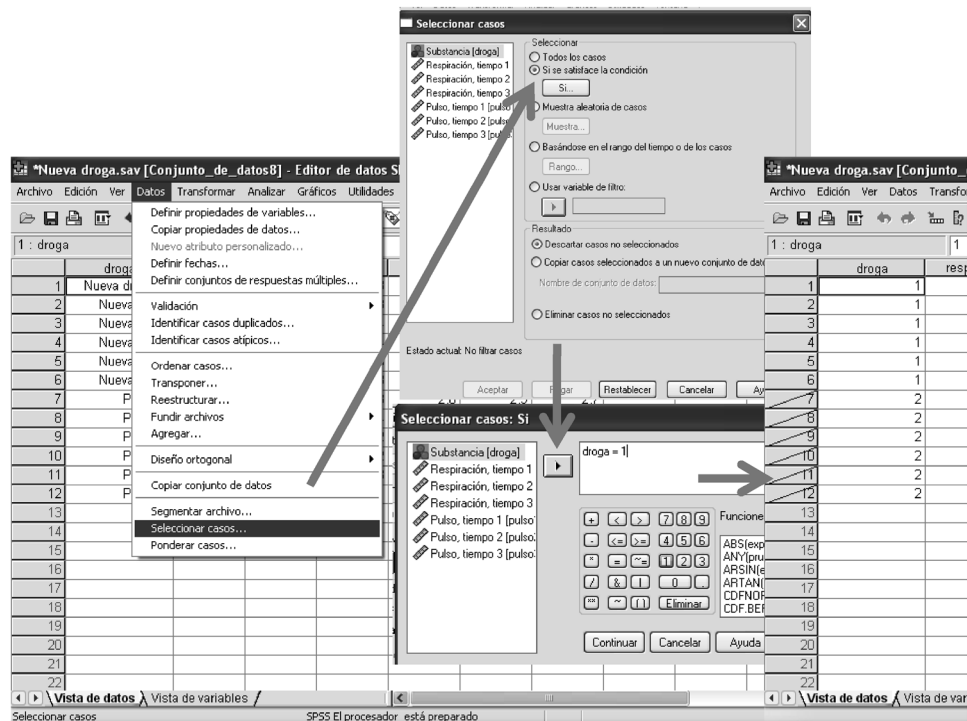
Per exemple, pel que fa a la variable sexe, si n'hem definit els valors com a: home=1 i dona=2, a la vista de dades ens apareixeran els codis 1 i 2, però si cliquem la icona assenyalada, en comptes dels codis apareixeran les etiquetes Home i Dona. Això ens permetrà conèixer amb més facilitat el contingut de la matriu.



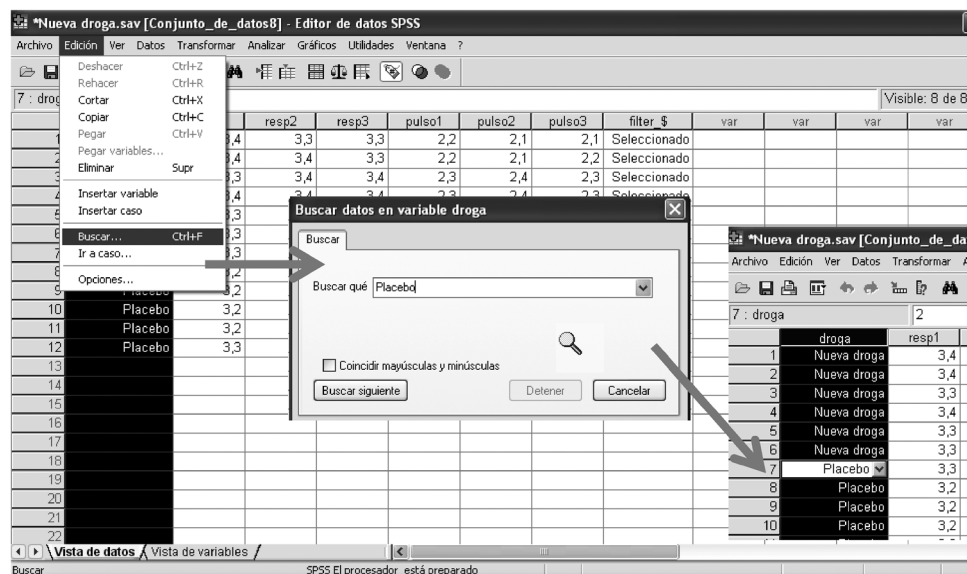
EDICIÓ DE DADES

Quan parlem d'edició de dades ens referim a diferents opcions: *selecció, canvi, tallar, copiar i enganxar dades, inserir o esborrar variables o casos, cercar casos o valors de la variable, etc.*

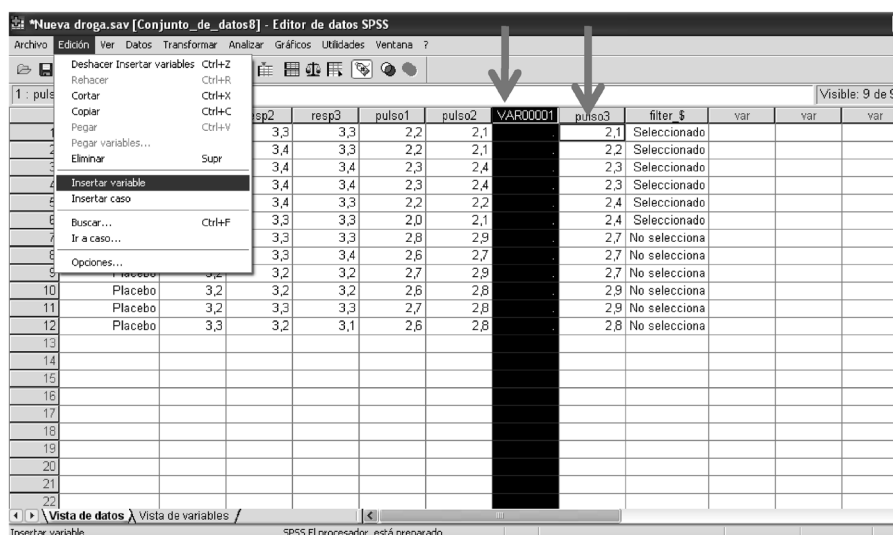
D'entre les opcions abans esmentades, en aquest apartat explicarem les que considerem bàsiques per al treball amb l'SPSS. Començarem per la *selecció de casos*. Tal com s'indica a la figura de sota, el que fem primer és escollir l'opció «seleccionar casos» del menú, que obre un nou quadre de diàleg en el qual clicant l'opció «Si se satisface la condición», obrirà un nou quadre que ens permetrà especificar quins són els casos que volem seleccionar.



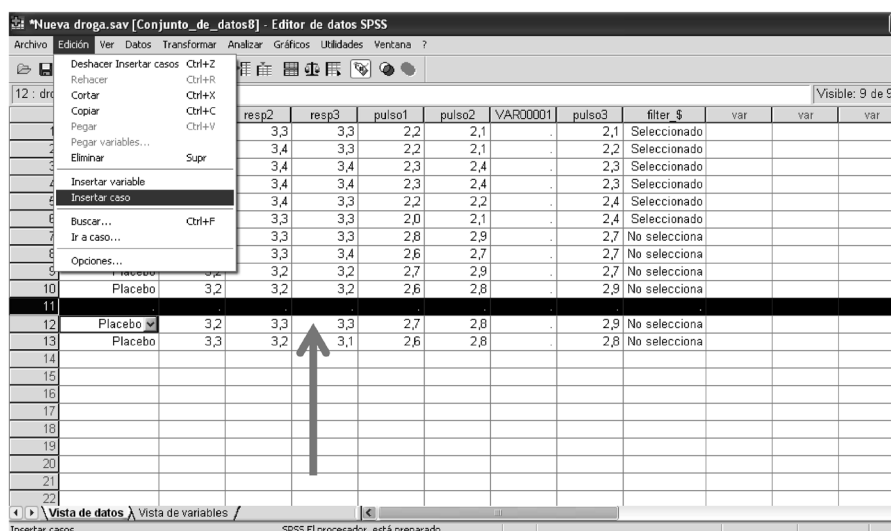
Una altra opció similar a l'anterior serà la **cerca de casos**, fent-la servir quan els casos que volem cercar siguin pocs. El que farem serà escollir aquesta funció dins el menú Edició; un cop remarcada la variable de la qual volem fer la cerca, havent-la seleccionat clicant sobre el seu nom dins la vista de dades, llavors s'obrirà un quadre de diàleg en què especificarem el cas que volem cercar.



Finalment, farem esment a com inserir variables i casos dins una matriu ja elaborada. La figura següent ens indica com **inserir una nova variable**, tenint en compte que es generarà a l'esquerra d'on estava situat el cursor en el moment de demanar la funció. Un cop inserida, s'haurà de definir tal com hem explicat a l'apartat 1.2 i posteriorment introduir els valors.



Pel que fa a *inserir nous casos*, la figura de sota ens indica com fer-ho, tenint en compte que s'insertaran a la mateixa fila on estava situat el cursor en el moment de demanar la funció.



CREACIÓ DE VARIABLES NOVES

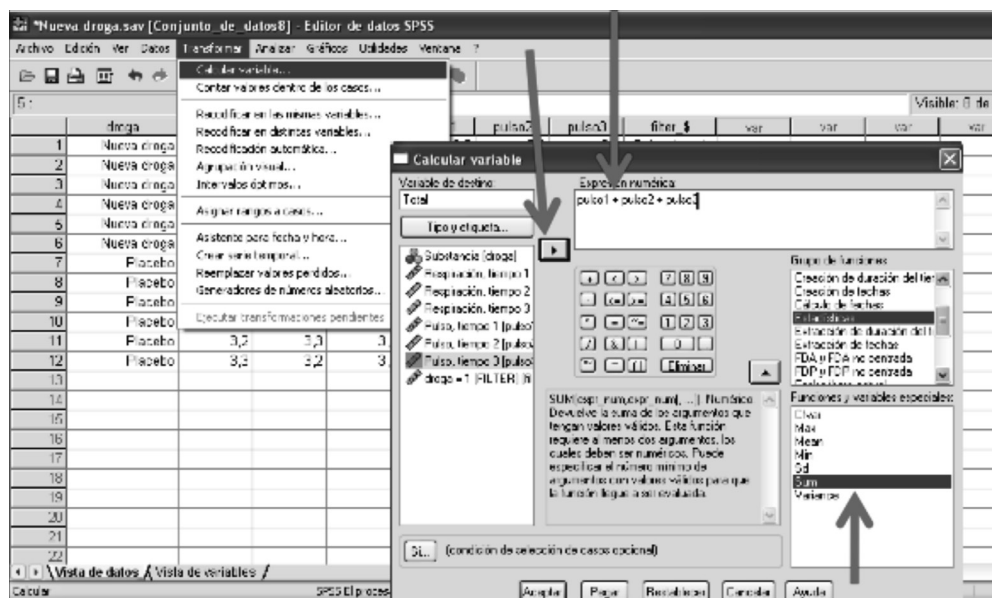
Un cop creada la matriu, tenim dues opcions d'una gran utilitat per modificar les variables existents i aconseguir que siguin adequades per als nostres propòsits: **calcular noves variables o bé recodificar-les**.

Efectivament, és molt possible que necessitem crear una nova variable, fruit del càlcul d'alguna o algunes de les variables que ja tenim creades. Per exemple, imaginem que la nostra matriu consta de tres variables que són les puntuacions de tres preguntes o ítems d'un instrument de mesura. És molt probable que necessitem el total que ha obtingut cada subjecte en aquest instrument; per tant, l'operació que volem fer és *calcular una nova variable* que serà el total de puntuació de cada subjecte en l'instrument de mesura.

Com ho fem? En l'opció *transformar* del menú, hi trobem la funció *calcular*; un cop la cliquem, s'obre el desplegable que veiem a la figura posterior, on observem que tenim dues opcions:

- Utilitzar el símbols de la suma, la resta, la multiplicació, etc. que tenim al teclat i que també apareixen en el desplegable.
- Utilitzar les funcions que el menú desplegable ens ofereix, situades en el requadre de la part dreta.

En tots dos casos seleccionarem les variables entre les que apareixen al requadre de l'esquerra, que són totes les que s'han definit prèviament. En el nostre exemple farem servir les tres variables definides i les anirem sumant entre elles per aconseguir el total de l'instrument. No obstant això, també podríem haver utilitzat la funció *sumatori* entre les funcions establertes.



D'altra banda, és molt probable que també necessitem, en altres ocasions, modificar variables canviant els valors o codis, fent el que anomenem una **recodificació**. Imaginem que en una recerca hem recollit dades sobre el rendiment acadèmic de l'alumnat, essent aquest el resultat obtingut en un examen en puntuacions de 0-10, però per al nostre estudi necessitem aquestes puntuacions en les categories següents: suspès, aprovat, notable i excel·lent. Per a transformar els valors del rendiment acadèmic haurem de fer la recodificació de la variable.

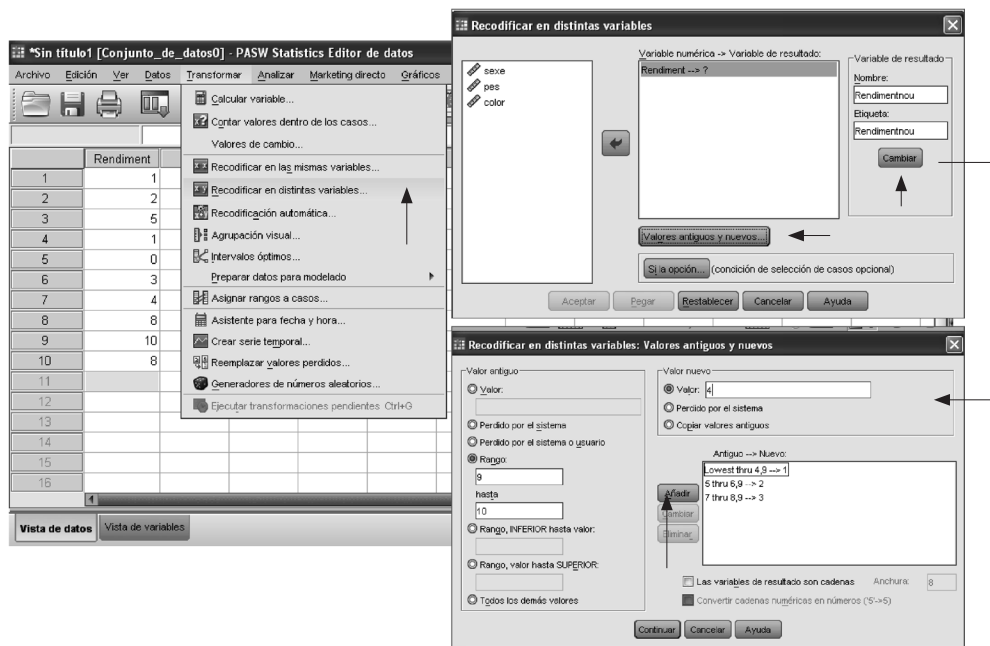
Com ho fem? Per recodificar les variables haurem d'accedir a l'opció transformar del menú, essent aquest primer pas el mateix que hem seguit per a calcular noves variables. Després de clicar la funció recodificar ens apareixeran dues opcions:

- Crear la variable sobre la qual volem recodificar, esborrant el seu contingut anterior (*en las mismas variables*).
- Crear la variable en una nova columna (*en distintas variables*). Si decidim crear una nova variable a part, li haurem de donar un nom. Per a aquest fi és important que activem l'opció *cambiar*.

Seguidament, haurem d'indicar quins són els valors de cada codi que volem recodificar i en quins valors es transformaran. És a dir, haurem de dir que un suspès és aquella persona que ha obtingut una puntuació que oscil·la entre 0 i 4,9, etc. Per fer-ho, hem d'accedir a l'opció de «*valores antiguos y nuevos*». La finestra que s'obre està dividida en dues parts verticals.

La part esquerra que defineix els valors antics, és a dir, els intervals de les puntuacions en el nostre exemple (0-4,9, 5-6,9, 7-8,9, 9-10).

La part dreta que defineix els nous valors (suspès, aprovat, notable, excel·lent). No obstant això, l'SPSS només ens permet introduir els valors numèrics, per exemple: 1, 2, 3 i 4; i és per això que, en aquest cas, en un segon moment haurem de definir la variable i posar les etiquetes corresponents (1 serà suspès; 2, aprovat, etc.).

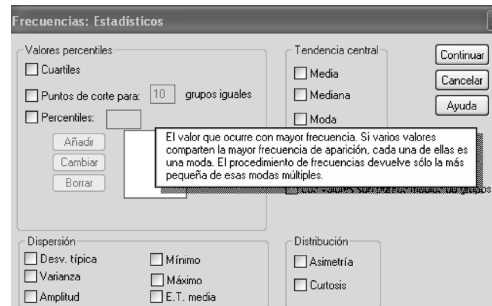


Per definir aquests intervals de puntuacions podem utilitzar l'opció de rang de valors, o bé altres alternatives: per exemple, el suspès pot ser definit a partir de rang del menor fins a ..., posant únicament el 4,9 com a puntuació màxima; o bé l'excel·lent pot ser identificat a través de rang de ... fins al major, posant en aquest cas el 9 com a puntuació mínima. En tot cas, cada vegada que posem els valors antics i nous d'una categoria hem de clicar a afegir perquè es mantinguin els valors. Un cop s'hagin afegit totes les categories a la petita pantalla que es troba a la part dreta de la finestra podrem continuar l'operació i veurem com es crea una nova variable a la darrera columna de la matriu.

1.2.4. Opcions d'ajuda a l'SPSS

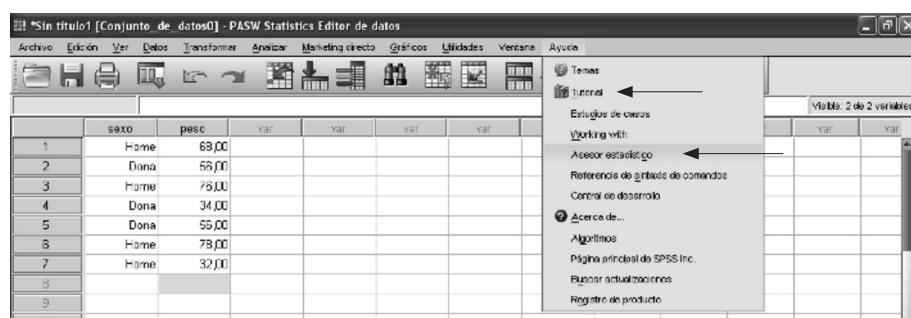
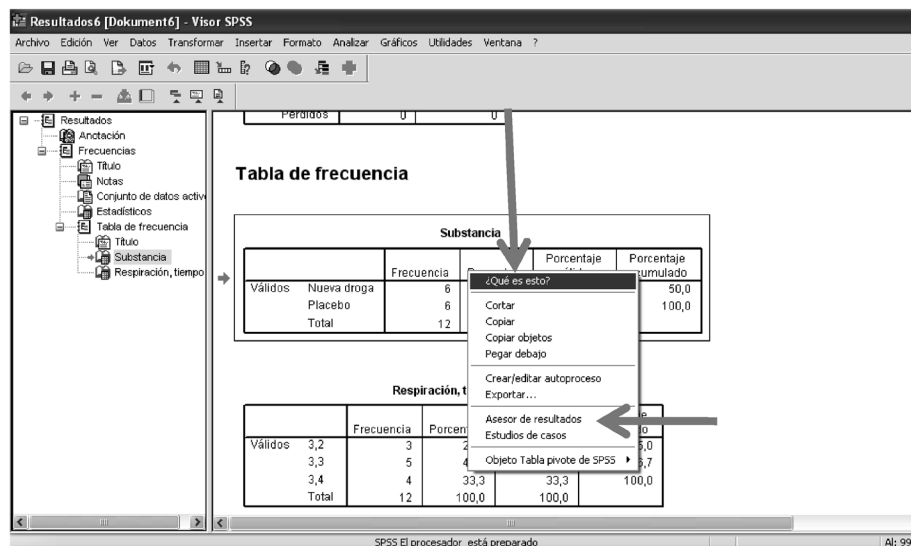
L'SPSS disposa de diverses opcions d'ajuda que poden ser molt útils en el seu aprenentatge i ús. Per començar, hem de parlar del **botó d'ajuda** que apareix a cada finestra i que permet l'accés als diferents continguts, índexs i opcions de recerca de qualsevol terme específic.

D'altra banda, també hem de parlar de la possibilitat d'obtenir ajuda en els diferents **quadres de diàleg**. Per exemple, si no recordem el concepte de la moda, només hem de fer un petit clic amb el botó dret del ratolí sobre el seu enunciat i ens sorgeix la seva definició.



Seguint el mateix procediment, podrem obtenir informació d'ajuda sobre els termes d'una taula o un gràfic; el que haurem de fer serà activar-la clicant-hi a sobre i, després de prémer el botó dret del ratolí, seleccionar l'opció **¿Qué es esto?** del menú que es desplega per tal que apareguin les definicions dels termes. També hem de fer esment a l'**assessor de resultats**, que apareixerà prement el botó dret del ratolí sobre una taula i llavors seleccionant l'opció assessor de resultats dins el menú que es desplega; d'aquesta manera, obtindrem informació sobre la interpretació dels resultats.

Una altra opció d'ajuda força important és l'**assessor estadístic**, que apareix al menú d'ajuda general i proporciona un mètode d'assistència per a trobar el procediment de SPSS més adequat per al tipus de resultat que volem obtenir. Aniran apareixent una sèrie de preguntes senzilles referides a les dades de què disposem i a allò que volem aconseguir i finalment ens proposarà el procediment més adequat a seguir. Finalment, parlarem del **Tutorial**, que és una opció que apareix al menú general d'ajuda i que ens ofereix una explicació pas a pas de diferents temes de funcionament de l'SPSS que ens permetrà escollir aquests temes dins d'un índex.



1.3. Exercicis resolts

1. Quan la finalitat de l'estadística consisteix a obtenir una sèrie de conclusions sobre algun aspecte rellevant de la població, a partir d'observacions en mostres, davant de quin tipus d'estadística ens trobem?

SOLUCIÓ:

Si els resultats obtinguts fan referència a característiques de la població d'estudi, voldrà dir que estem realitzant anàlisis pròpies de l'estadística inferencial; els resultats obtinguts, els anomenarem paràmetres, en tant que es diferencien dels estadístics descriptius, resultats referents a l'estadística descriptiva.

2. Indica quines variables són qualitatives i quines quantitatives:

- a) Professió que t'agradaria exercir.
- b) Nombre de gols marcats pel teu equip en la darrera temporada.
- c) Nombre d'alumnes de la teva facultat.
- d) El color dels ulls dels teus companys de classe.
- e) Coeficient intel·lectual dels membres de la teva família.

SOLUCIONS:

- a) Variable qualitativa
- b) Variable quantitativa
- c) Variable quantitativa
- d) Variable qualitativa
- e) Variable quantitativa

3. De les variables següents, indica quines són discretes i quines contínues:

- a) Nombre d'accions venudes cada dia a la Borsa.
- b) Temperatures registrades cada hora en un observatori.
- c) Nombre de fills de cinquanta famílies.
- d) Cens anual dels espanyols.
- e) Nota dels aprovats en una assignatura.

SOLUCIONS:

- a) Variable discreta
- b) Variable contínua
- c) Variable discreta
- d) Variable discreta
- e) Variable contínua

4. A continuació, es presenta una petita part d'un conjunt de dades sobre els estudiants d'últim curs d'Educació Primària; identifica el tipus de variable i el nivell de mesura de cadascuna de les variables:

Codi	Edat	Pes (kg)	Sexe	Especialitat docent	Nota mitjana de l'expedient
001	22	55	D	Musical	7,6
002	25	87	H	Llengua anglesa	9,2
003	28	73	D	Magisteri general	6,5
004	22	68	D	Magisteri general	8,9
005	24	79	H	Educació física	7,2
006	26	60	D	Llengua anglesa	6,7
007	25	59	D	Magisteri general	7,4

SOLUCIONS:

Codi: quantitativa discreta ordinal

Edat: quantitativa continua de raó

Pes: quantitativa continua de raó

Sexe: qualitativa dicotòmica nominal

Especialitat docent: qualitativa politòmica nominal

Nota mitjana de l'expedient: quantitativa contínua d'interval

L'estadística descriptiva és una aproximació objectiva al coneixement d'una mostra d'individus en un moment determinat, que estudia els valors de les dades recollides i els organitza mitjançant taules, representacions gràfiques i indicadors numèrics:

- Les taules permeten comparar numèricament els valors de les variables (la freqüència dels diversos valors, les relacions entre els valors de les diferents variables, etc.).
- Les representacions gràfiques ofereixen aquestes comparacions d'una manera sintètica i visual.
- Finalment, els indicadors són valors numèrics de referència que permeten comparar múltiples aspectes de diverses mostres (els seus valors centrals, la distribució dels diferents valors, la seva homogeneïtat, la forma de la distribució, etc.).

2.1. Tabulació i representació gràfica

2.1.1. La tabulació

La tabulació és el procediment d'organitzar les dades en una taula. Forma part d'una tècnica d'ordenació de dades. Concretament, les taules de freqüències consten dels aspectes següents, que poden variar en funció del tipus de variable:

- **Freqüència absoluta:**
 - **Individual (fi):** és el nombre de vegades que apareix una determinada dada.
 - **Acumulada (fa):** és la suma de les freqüències absolutes individuals de les dades anteriors més l'actual.

$$fa_j = \sum_{k=1}^j fi_k$$

Per exemple: $fa_4 = fi_1 + fi_2 + fi_3 + fi_4$

L'última freqüència absoluta acumulada coincideix amb el total de la mostra (n).

- **Freqüència relativa:**
En proporció:
 - **Individual (pi):** és el nombre de vegades que apareix una dada (fi) en relació amb el total de la mostra (n). És el tant per u:

$$pi = \frac{fi}{n}$$

- **Acumulada (pa):** és la freqüència absoluta acumulada (fa) dividida per la grandària de la mostra (n):

$$pa = \frac{fa}{n}$$

O bé podem dir que és la suma de les freqüències relatives individuals de les dades anteriors més l'actual.

$$pa_j = \sum_{k=1}^j pi_k$$

Per exemple: $pa_4 = pi_1 + pi_2 + pi_3 + pi_4$

L'última freqüència relativa en proporció acumulada val sempre 1.

En percentatge:

- **Individual (Pi):** és la freqüència relativa (pi) en base a 100 (percentatge).

$$Pi = \frac{fi}{n} \cdot 100 = pi \cdot 100$$

- **Acumulada (Pa):** és la suma de les freqüències relatives individuals de les dades anteriors més l'actual.

$$Pa_j = \sum_{k=1}^j Pi_k$$

Per exemple: $Pa_4 = Pi_1 + Pi_2 + Pi_3 + Pi_4$

L'última freqüència relativa en percentatge acumulada val sempre 100.

EXEMPLE:

En un institut de secundària cal fer previsions, a mitjan curs, sobre el nombre de grups que hi haurà a cada nivell el curs següent, la qual cosa depèn del nombre de matèries que suspengui cada alumne/a. Per fer-les, es compta amb dades sobre els resultats de l'alumnat després de la primera avaluació:

Amb els setanta-sis alumnes de primer d'ESO s'obtenen les dades següents:

Nombre de matèries suspeses	Quantitat d'alumnes
Cap	24
Una	9
Dues	2
Tres	7
Quatre	12
Cinc	9
Sis	5
Set	2
Vuit	2
Nou	4

Tabulació d'aquestes dades:

X	fi	fa	pi	pa	Pi	Pa
0	24	24	0,316	0,316	31,6	31,6
1	9	33	0,118	0,434	11,8	43,4
2	2	35	0,029	0,461	2,9	46,1
3	7	42	0,092	0,553	9,2	55,3
4	12	54	0,158	0,711	15,8	71,1
5	9	63	0,118	0,829	11,8	82,9
6	5	68	0,066	0,895	6,6	89,5
7	2	70	0,026	0,921	2,6	92,1
8	2	72	0,026	0,947	2,6	94,7
9	4	76	0,053	1,000	5,3	100,0

Aquesta taula permet fer diverses previsions:

- Els alumnes que en aquests moments passarien de curs automàticament, ja que tenen com a molt dues matèries suspeses, que serien uns trenta-cinc alumnes (només el 46,1% del total).
- El total d'alumnes que estarien en condicions de passar de curs al juny, afegint als anteriors els que tenen fins a quatre matèries suspeses (ja que poden millorar lleugerament els seus resultats durant els trimestres següents), que serien uns cinquanta-quatre alumnes (el 71,1% del total).
- El total d'alumnes que podrien acabar passant de curs gràcies a les recuperacions de setembre, que podrien ser els que tenen fins a sis matèries suspeses al primer trimestre, és a dir, uns seixanta-vuit alumnes (pràcticament el 90% del total).

Veiem, doncs, que les freqüències acumulades permeten agafar diferents ventalls de casos segons el criteri que es vulgui aplicar.

TAULES DE FREQÜÈNCIA AMB SPSS

L'SPSS proporciona taules de freqüències de les variables que consten dels aspectes següents:

- Els valors vàlids o les categories.
- La seva freqüència absoluta individual (*frecuencia*).
- La freqüència relativa individual en percentatge (*porcentaje*).
- La freqüència relativa individual en percentatge sense comptabilitzar els valors perduts (*porcentaje válido*).
- La freqüència relativa acumulada en percentatge (*porcentaje acumulado*) basada en el percentatge vàlid.

Presentem la taula obtinguda amb SPSS sobre la variable *Nombre de matèries suspeses*. Per demanar aquesta taula des de l'SPSS es fa mitjançant l'opció *Analizar / Estadísticos descriptivos / Frecuencias*.

NOMBRE DE MATÈRIES SUSPESES

		Freqüència	Percentatge	Percentatge vàlid	Percentatge acumulat
Vàlids	0	24	31,6	31,6	31,6
	1	9	11,8	11,8	43,4
	2	2	2,6	2,6	46,1
	3	7	9,2	9,2	55,3
	4	12	15,8	15,8	71,1
	5	9	11,8	11,8	82,9
	6	5	6,6	6,6	89,5
	7	2	2,6	2,6	92,1
	8	2	2,6	2,6	94,7
	9	4	5,3	5,3	100,0
Total		76	100,0	100,0	

A la taula es presenta com queden les freqüències de percentatge vàlid i percentatge acumulat en el cas d'una variable (nombre de fills) quan hi ha valors perduts.

NOMBRE DE FILLS

		Freqüència	Percentatge	Percentatge vàlid	Percentatge acumulat
Vàlids	0	419	27,6	27,8	27,8
	1	255	16,8	16,9	44,7
	2	375	24,7	24,9	69,5
	3	215	14,2	14,2	83,8
	4	127	8,4	8,4	92,2
	5	54	3,6	3,6	95,8
	6	24	1,6	1,6	97,3
	7	23	1,5	1,5	98,9
	8 o més	17	1,1	1,1	100,0
	Total	1.509	99,5	100,0	
Perduts	No contesta	8	0,5		
Total		1.517	100,0		

ALTRES TAULES

A més a més de les taules de freqüències també podem fer ús de les taules de contingència i d'altres en les quals creuar més d'una variable.

Una **taula de contingència** és una taula en la qual situar les freqüències a l'interior, en funció de dues variables amb dues o més categories cada una.

EXEMPLE:

En un centre educatiu s'ha fet una enquesta a l'alumnat per tal de conèixer com valora el seu rendiment i quines considera que en són les causes. Les dades obtingudes a partir de les respostes que han facilitat els 218 alumnes participants en l'estudi es poden organitzar en una taula de contingència que relacioni aquestes dues variables:

TAULA DE CONTINGÈNCIA RESULTATS ESPERATS I PER QUIN MOTIU

Recompte

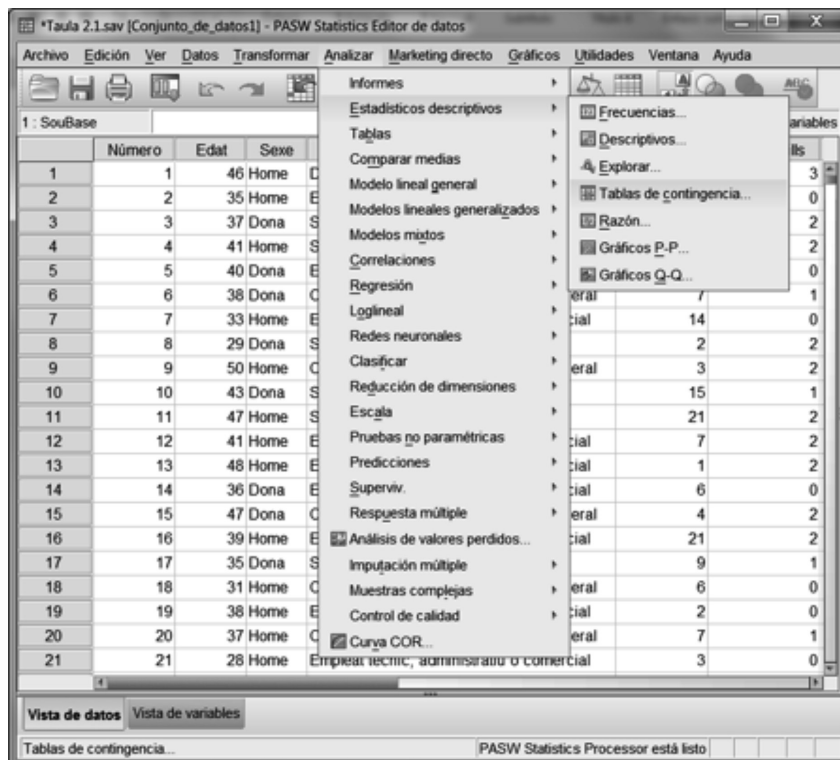
		Per quin motiu			
		Per la meua actitud	Ho trobo difícil	Perquè m'esforço	M'agrada o m'interessa
Resultats esperats	Suspendré bastant	27	4	2	0
	Suspendré poc	23	10	21	1
	Aprovaré amb notes justes	20	5	46	16
	Aprovaré amb bones notes	3	0	18	21
Total		73	19	87	38
					217

D'aquesta taula es poden extreure algunes conclusions:

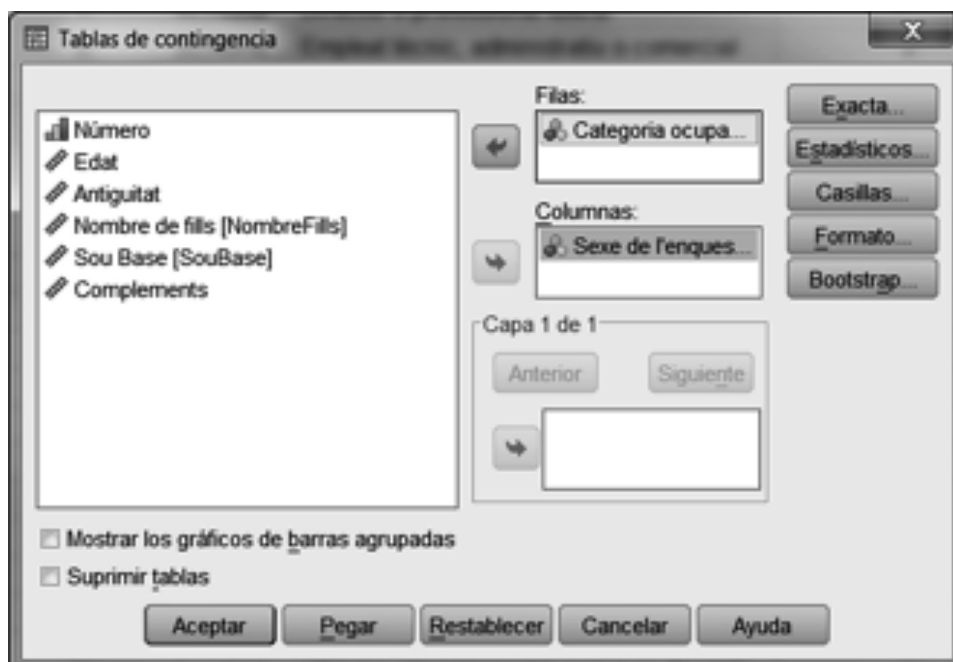
- La majoria d'alumnes que esperen treure molt mals resultats ho atribueixen fonamentalment a la seva actitud.
- Entre els alumnes que preveuen suspendre poc hi ha una major diversitat d'opinions sobre les causes, ja que alguns ho atribueixen a la seva actitud, d'altres al seu esforç (probablement en el sentit que sense esforç suspendrien encara més) i uns altres a la dificultat de les matèries.
- D'altra banda, l'alumnat que pensa aprovar amb notes justes ho atribueix sobretot al seu esforç.
- Finalment, l'alumnat que confia a aprovar amb bones notes ho relaciona amb el fet que li agrada o l'interessa o bé amb el seu esforç.

En resum, les taules de contingència ens evidencien les relacions entre els diferents valors de dues variables.

L'SPSS ens ofereix crear taules de contingència. A la figura, podeu veure el procés pel qual es demana una taula de contingència a l'SPSS.



I llavors:



També podem representar estadístics a l'interior d'una taula, en funció de diferents variables categòriques.

EXEMPLE:

Volem saber el valor mitjà (anomenat mitjana aritmètica)¹ de la nota global que han obtingut els alumnes d'un institut en funció de l'etapa educativa en què estan (ESO, cicles formatius de grau mitjà, batxillerat o cicles formatius de grau superior). Aquest tipus de taula és el que presentem tot seguit:

		Nota global
		Mitjana
Etapa educativa	Educació secundària obligatòria	5,15
	Cicle formatiu de grau mitjà	5,68
	Batxillerat	5,13
	Cicle formatiu de grau superior	6,22

També podem demanar més variables per fer subgrups, com per exemple el *sex*. I així obtindrem la mitjana de la nota global, segons l'etapa educativa, d'homes i dones, com es pot veure a la taula següent:

¹ La mitjana aritmètica és el resultat de sumar tots els valors i dividir-lo entre el nombre de valors; és un valor mitjà. Aquest indicador estadístic i d'altres es tractaran en l'apartat següent en aquest mateix tema.

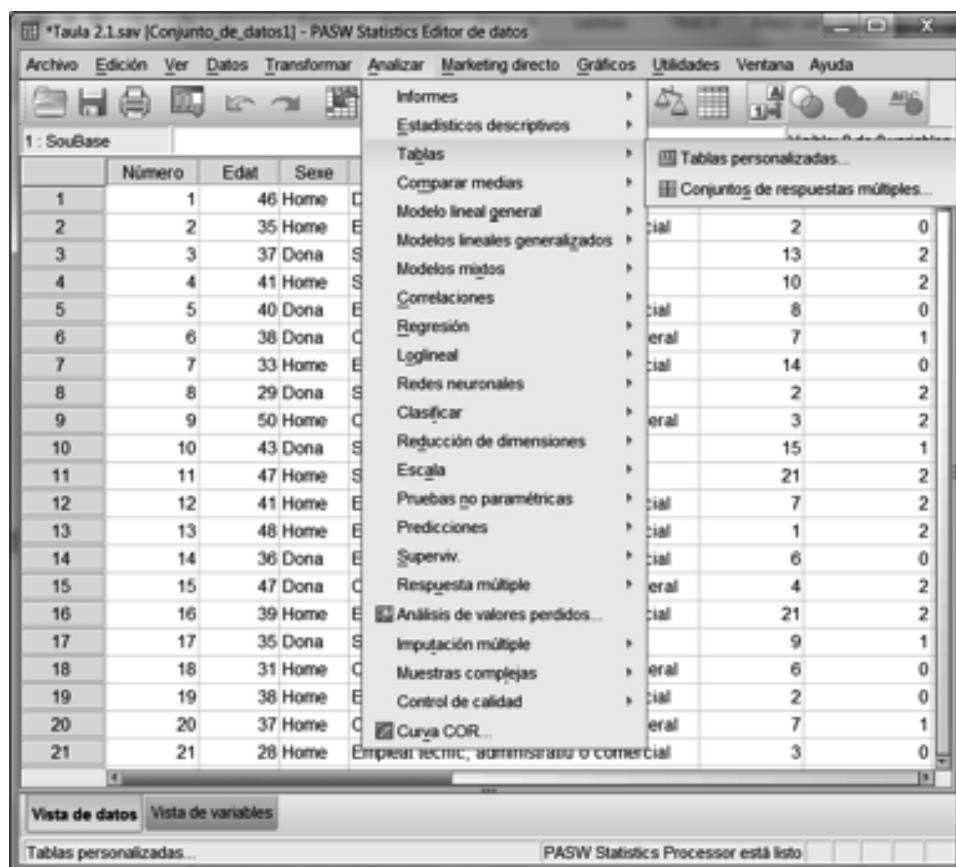
				Nota global
				Mitjana
Etapla educativa	Educació secundària obligatòria	Sexe	Home	4,76
			Dona	5,63
	Cicle formatiu de grau mitjà	Sexe	Home	5,71
			Dona	5,61
	Batxillerat	Sexe	Home	4,80
			Dona	5,64
	Cicle formatiu de grau superior	Sexe	Home	6,30
			Dona	6,16

Podem arribar a demanar taules tan complexes com ens interressi.

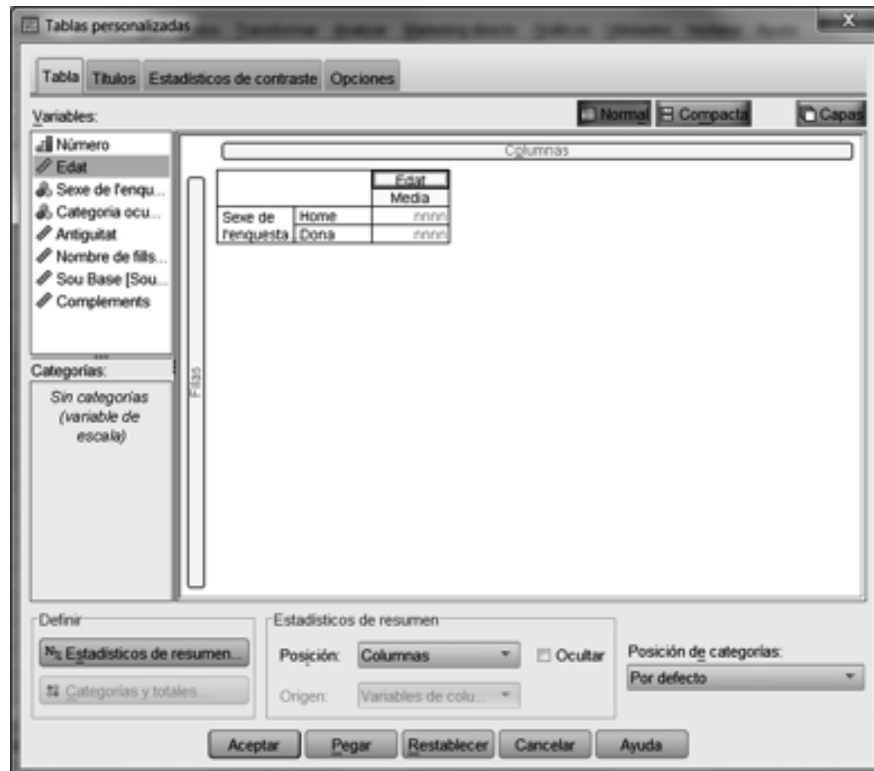
L'SPSS ens ofereix crear aquesta mena de taules. En la figura podeu veure un dels possibles processos per demanar una taula a través de les *taules personalitzades*.

Els subgrups, els fariem arrossegant al rectangle *Filas* alguna variable qualitativa (l'etapa educativa en l'exemple) i al rectangle *Columnas* la variable quantitativa a resumir (la nota global en l'exemple). En l'exemple que es veu a la figura es demana la mitjana aritmètica de la variable *edat* en funció del sexe; obtindrem així una taula que ens indiqui les mitjanes d'edat d'homes i dones.

Des d'aquesta opció, podem fer taules amb els estadístics que volem.



I llavors:



2.1.2. Representació gràfica

Les tècniques de **representació gràfica** són un altre instrument per a presentar la informació original. La pèrdua d'informació és major, però presenta l'avantatge d'establir diferències entre dues distribucions a simple vista.

Representació gràfica:

Ordenades:
freqüències
individuals o
acumulades

Abscisses: categories o
valors de la variable

Per a la representació gràfica és necessari diferenciar si la variable és qualitativa, quantitativa contínua o discontinua. A més, la correcta realització d'un gràfic necessita que s'hi inclogui la font d'origen de les dades i el títol que indiqui la variable a la qual correspon la informació.

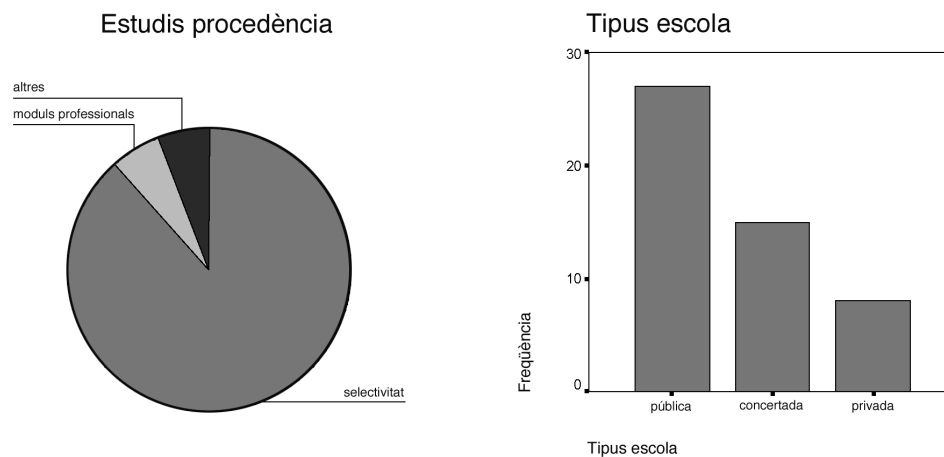
a) Variables qualitatives:

Per a les variables qualitatives, la representació permet els gràfics següents:

- **Diagrama de barres:** gràfic de barres separades entre si, amb la mateixa distància i amplitud de base, amb longituds proporcionals a les freqüències.

- **Ciclograma o diagrama de sectors:** cercle que es divideix en sectors circulars proporcionals a la freqüència de cada categoria.
- **Diagrama de rectangles:** la superfície d'un rectangle es reparteix proporcionalment a les freqüències.
- **Pictograma:** dibuixos amb la mida proporcional a les freqüències.

Alguns dels gràfics qualitatius que ofereix l'SPSS són el de barres i el de sectors, tal com podem veure al gràfic. Aquests es poden aconseguir des de l'opció del menú *Gràfics*. A la figura, podem consultar altres vies per demanar aquests gràfics.



b) Variables quantitatives discretes:

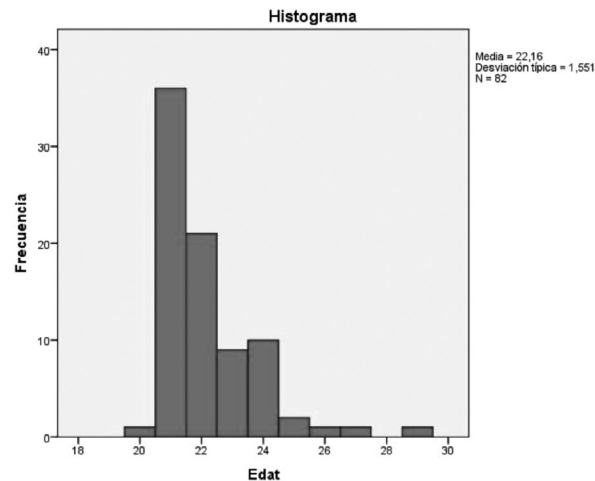
Les variables quantitatives discretes o discontinues es representen gràficament, entre d'altres, amb el *diagrama lineal*, format per blocs separats sobre cadascun dels valors de la variable que s'eleva fins a la freqüència o el nombre de vegades que es repeteix. Són molt semblants al diagrama de barres, però, en lloc de rectangles simètrics, s'utilitzen línies verticals.

c) Variables quantitatives contínues:

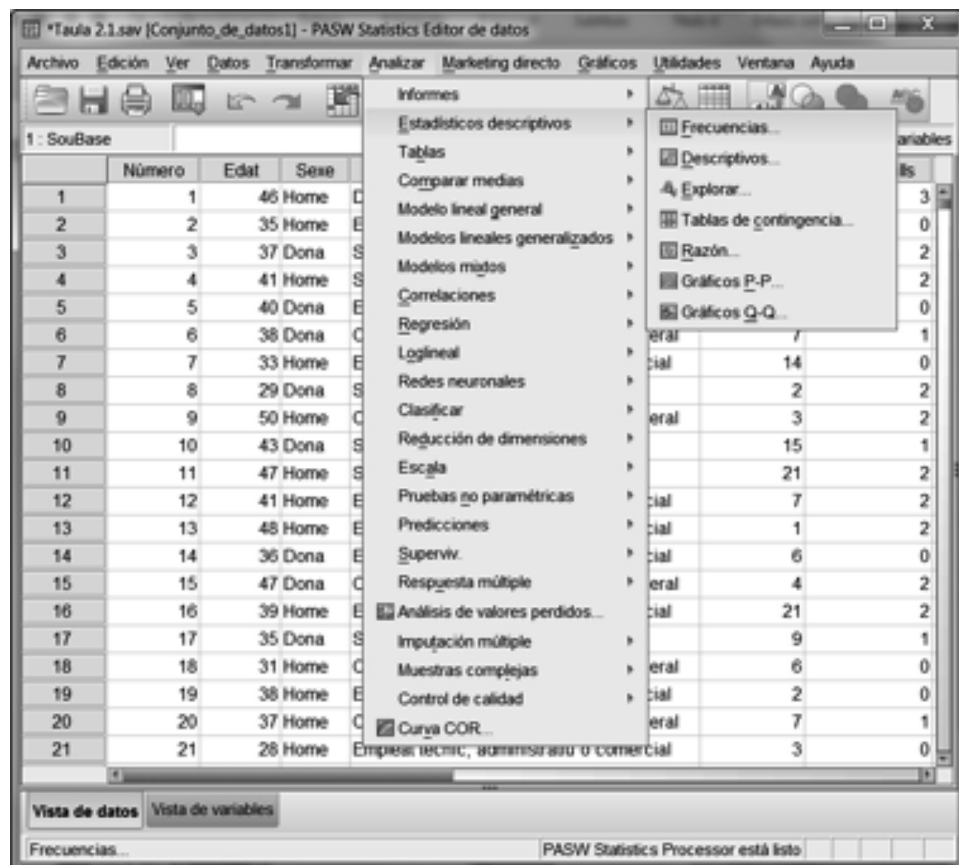
Les variables quantitatives contínues es poden representar gràficament amb els gràfics següents:

- **Polígon de freqüències:** és un gràfic de les freqüències de classe en relació amb la marca de classe (punt mig d'un interval simbolitzat per x'), en què els punts són units en una línia. A l'eix de les ordenades es poden representar les freqüències absolutes, les relatives o les percentuals, i a l'eix de les abscisses, les marques de classe.
- **Histograma:** consisteix en un conjunt de rectangles units, amb un eix d'ordenades format per les freqüències absolutes, relatives o percentuals, i a l'eix de les abscisses es representen el límit inferior del primer interval i els límits superiors de la resta dels intervals.
- **Polígon de freqüències acumulades (Ogiva):** és un gràfic que representa les freqüències acumulades. Es representa igual que el polígon, però en l'eix de les ordenades hi ha les freqüències absolutes acumulades o les freqüències percentuals acumulades. A les abscisses, el límit inferior (li) del primer interval i els superiors (Li) de la resta d'intervals. És important per representar gràficament el nombre de casos que es troben per sobre o per sota d'un determinat valor. Exemple: conèixer el percentatge d'aprovat o dels qui hagin tret menys d'un notable.

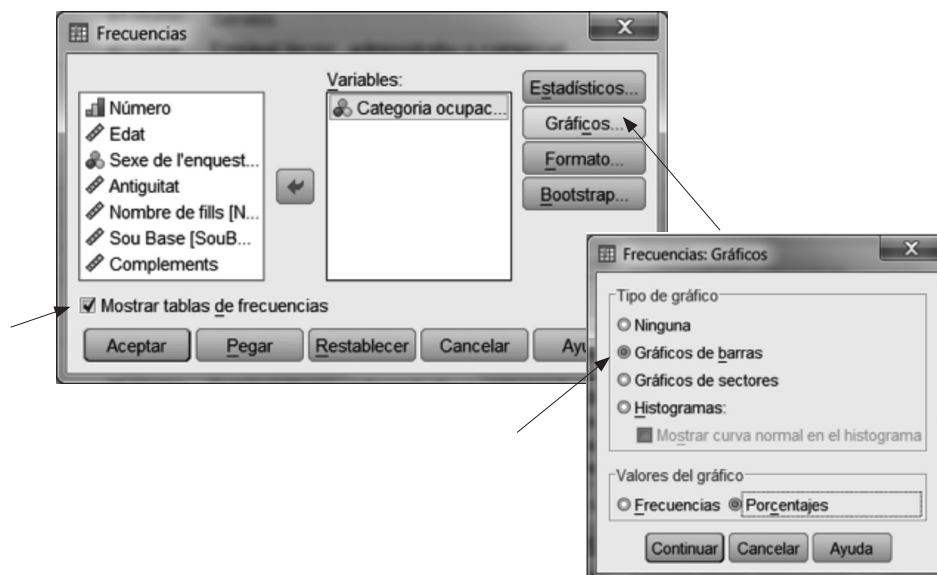
Un dels gràfics quantitativs més habituals que ofereix l'SPSS és l'histograma que es pot aconseguir des de l'opció del menú *Gràfics*. A la figura següent podeu consultar altres vies per demanar aquest gràfic.



Finalment, a la figura següent queda reflectit un altre procediment a seguir per tal de sol·licitar els gràfics i les taules exposats en aquest tema. Com podem observar, hem d'accedir a l'opció *Estadísticos descriptivos* del menú i, dins d'aquesta, seleccionar *Frecuencias*, per tal de demanar tant la taula com els gràfics corresponents a les variables que ens interessin.



I llavors:



ALTRES GRÀFICS

Als anys vuitanta, amb la proposta de noves tècniques d'estadística descriptiva assistida per ordinador, es va començar a parlar d'*Anàlisi exploratori de dades* (EDA).

Aquest és un conjunt de tècniques innovadores, tant a nivell de gràfics com d'anàlisi, que té com a objectiu aconseguir un coneixement previ de les dades a partir d'un examen visual. Per tant, entre d'altres aspectes, recull la utilitat i la necessitat de plantejar la construcció de diferents tipus de gràfics que siguin bons sintetitzadors de la informació proporcionada per les dades.

Des d'aquesta perspectiva, es considera el diagrama de caixa com un gràfic que permet obtenir una impressió global dels aspectes més rellevants de la distribució d'una variable.

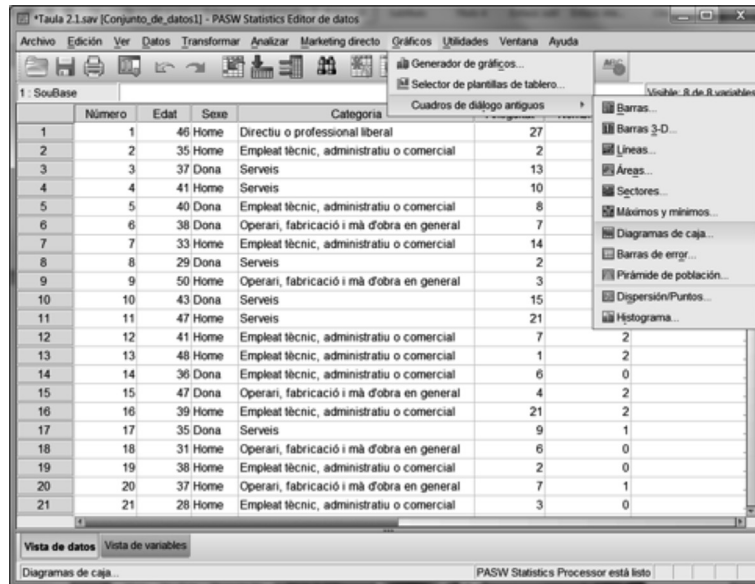
En aquest gràfic es representen estadístics bàsics de la distribució: mediana, quartils i valors extrems. És a dir, que amb una simple observació del gràfic podem obtenir una sèrie d'indicadors estadístics:² amplitud, rang interquartílic, mediana, simetria (posició que ocupa la mediana respecte als quartils).

Per tal d'obtenir aquest gràfic seguirem les passes assenyalades a la figura. En primer lloc, hem d'accedir a l'opció *Gràfics* del menú i seleccionar-hi el diagrama desitjat (*diagrama de caixes*).

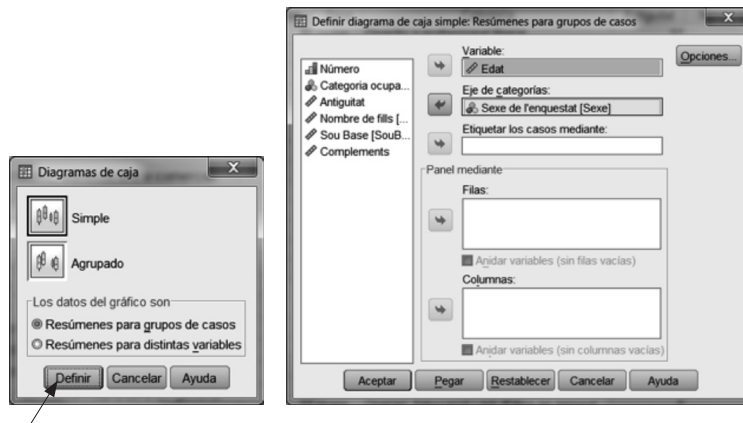
Un cop hàgim seleccionat el *Diagrama de cajas* del menú, se'ns obrirà un nou quadre de diàleg en el qual haurem d'especificar si farem un gràfic simple o agrupat, essent la diferència entre ambdós que amb l'últim les dades es poden agrupar en funció dels valors d'una tercera variable.

Escollirem l'opció *Simple* i tot seguit haurem de decidir-nos entre *Resúmenes para grupos de casos* o bé *Resúmenes para distintas variables*. La primera opció ens permet definir les variables amb les quals pretenem elaborar el gràfic, essent necessari que una sigui quantitativa (*Variable*) i l'altra qualitativa (*Eje de categorías*). Amb la segona opció, podrem definir una o més variables sense necessitat de representar-les en funció d'una altra, sinó a elles mateixes.

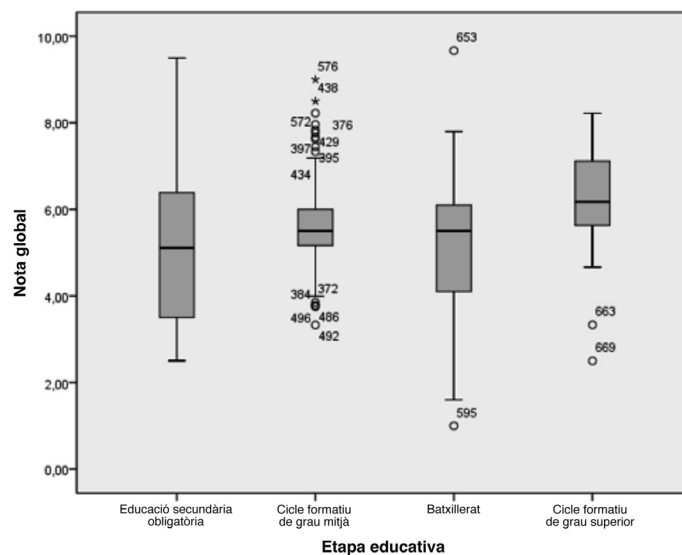
² Aprofundirem sobre els indicadors estadístics en el punt 2.2 d'aquest mateix tema.



I llavors:



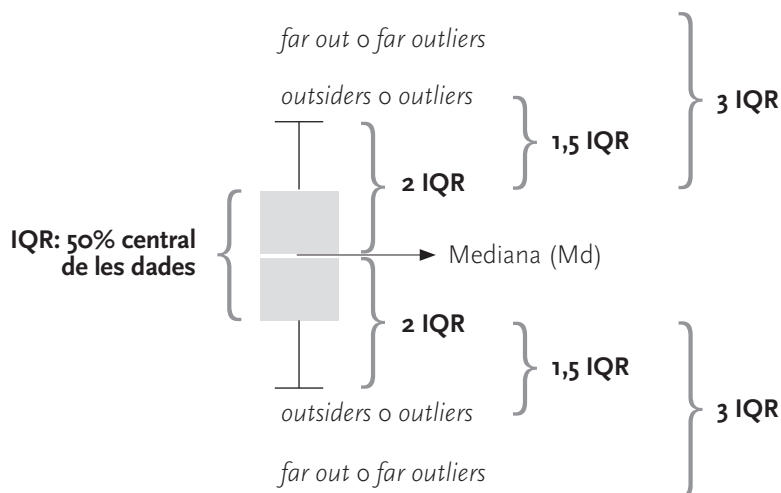
Al gràfic podem observar un diagrama de caixa elaborat amb les dues variables Nota global i Etapa educativa.



Per tal de poder interpretar un diagrama de caixa haurem de tenir coneixement dels indicadors estadístics que veurem en el proper apartat. La informació, però, que podem obtenir d'aquest tipus de gràfic és:

- D'una banda, la *posició de la mediana*, tendència central del grup. És marcada per la línia que es troba dins la caixa, havent-nos de fixar en si es troba al centre (distribució simètrica) o bé si està pròxima al límit inferior (asimetria positiva) o superior (asimetria negativa) de la caixa.
- D'altra banda, ens trobem que els límits d'aquesta caixa estan definits pels *quartils 1 i 3* (centils 25 i 75), obtenint també d'aquestes dades el *rang interquartílic*.
- També hem de tenir en compte que les línies que surten de la caixa central fan referència als *valors màxim i mínim* de la distribució.
- Pel que fa als *casos extrems*, en funció de la distància que prenguin respecte del cinquanta per cent central de la distribució, distingirem entre dos tipus: els remots i els extrems. Aquells casos que estan més allunyats s'anomenen remots (*far outliers*) i estan representats per un asterisc (*); en canvi, aquells que estan més propers, tot i trobar-se lluny de la resta del grup, s'anomenen extrems (*outliers*) i es representen per un cercle (O). Sempre, acompanyant aquests símbols, hi anirà el número d'identificació dels subjectes concrets.

Tots aquests elements es troben resumits a la figura següent, amb l'objectiu de facilitar-ne la comprensió i l'aplicació pràctica.



Finalment, hem de dir que aquests gràfics són molt útils per a comparar la distribució dels valors entre diferents grups; per exemple, en el gràfic podem comparar fàcilment les distribucions de l'ESO, els Cicles formatius de grau mitjà, el Batxillerat i els Cicles formatius de grau superior.

2.2. Indicadors estadístics

Una vegada que sabem com es recullen, organitzen i presenten les dades, ens queda estudiar com es resumeixen. Les representacions gràfiques ofereixen a l'investigador/a un mètode elemental per a l'anàlisi de les propietats de l'objecte d'estudi; es necessiten altres mètodes més precisos per a l'estudi de les característiques més importants de les distribucions de freqüències.

Bàsicament, amb un sol indicador no podem arribar a descriure a la perfecció tota una sèrie numèrica, sinó que com a mínim, a l'hora de sintetitzar una distribució, haurem de donar resposta a les qüestions següents:

- Com es troben els valors ordenats? (indicadors de posició)
- Com es concentren els valors? (indicadors de tendència central)
- Com es dispersen o quina variabilitat hi ha? (indicadors de dispersió)
- Quina forma adopta la distribució? (indicadors de forma)

A continuació presentem cada un d'aquests elements.

2.2.1. Indicadors de posició

Els *indicadors de posició* ens mostren la posició que ocupa un valor de la variable en la distribució. Divideixen la distribució en diverses parts iguals (per exemple, en cent parts) i indiquen el lloc que correspon a un valor concret en relació amb el grup. Per exemple, ens ajuden a respondre qüestions com les següents:

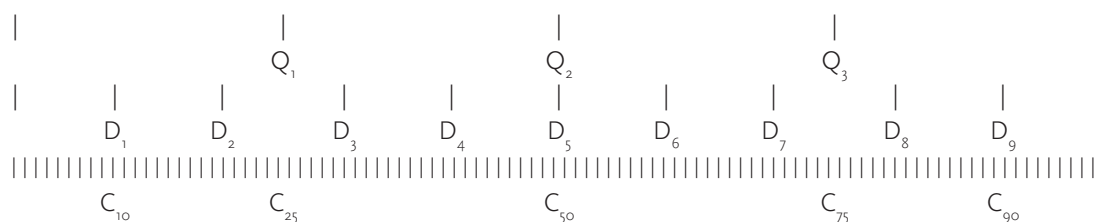
- Quin percentatge d'alumnes no passa de l'aprobat (és a dir, d'una puntuació de 5)?
- Quin percentatge està per sobre del notable?
- I entre el notable i l'aprobat?
- Etc.

Segons el nombre de parts en què dividim una distribució ordenada, podem trobar-nos amb els indicadors següents:

- **Percentils o centils (C):** es divideixen en cent parts iguals i donen la situació de l'individu dins del grup en tant per cent. Indiquen el percentatge de la distribució que queda per sota d'un determinat valor.
- **Decils (D):** el mateix que els centils, però dividint la distribució en deu parts iguals.
- **Quartils (Q):** es divideix la distribució en quatre parts iguals i cada part és equivalent al 25% de la distribució.

Un exemple: si cerquem el percentil 48 del rendiment acadèmic d'un grup d'estadística i el resultat que obtenim és 6,35; això vol dir que el 48% de l'alumnat d'aquest grup ha obtingut una nota que és igual o inferior a 6,35.

En la figura següent podem veure les equivalències entre els diferents indicadors de posició, abans esmentats.



Com es calculen?

1) Localitzar l'interval en el qual es troba el centil buscat (Pa). Essent aquest el primer en què el percentatge acumulat sigui igual o superior al centil que busquem.

2) Aplicar una fórmula:

$$C_{?} = li + \frac{\frac{? \cdot n}{100} - fa_{i-1}}{fi} \cdot i \quad D_{?} = li + \frac{\frac{? \cdot n}{10} - fa_{i-1}}{fi} \cdot i \quad Q_{?} = li + \frac{\frac{? \cdot n}{4} - fa_{i-1}}{fi} \cdot i$$

En què:

? = Percentil, decil o quartil que cerquem

li = Límit inferior de l'interval

fa_{i-1} = Freqüència absoluta acumulada de l'interval anterior al cercat

fi = Freqüència absoluta individual de l'interval

i = Amplitud de l'interval

EXEMPLE:

Els resultats en un prova, dins del procés d'avaluació d'un grup format per vint alumnes, són els següents:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
6	8,75	5	8,5	3	4,25	4,25	9,5	4,75	3,25	8	6,5	5,25	5,75	4,25	8,5	4,25	6	6	8,25

Si els ordenem de gran a petit, queden de la manera següent:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3,25	4,25	4,25	4,25	4,25	4,75	5	5,25	5,75	6	6	6	6,5	8	8,25	8,5	8,5	8,75	9,5

Ara podem localitzar els centils que ens interessin; per exemple, el 10 i el 90:

$$C_{10} = \frac{10 \cdot 20}{100} = 2 \quad C_{90} = \frac{90 \cdot 20}{100} = 18$$

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3,25	4,25	4,25	4,25	4,25	4,75	5	5,25	5,75	6	6	6	6,5	8	8,25	8,5	8,5	8,75	9,5

Aquests centils són, per tant, 3,25 i 8,5 en aquesta distribució.

També podem determinar els decils, com per exemple el decil 2 o el 8:

$$D_2 = \frac{2 \cdot 20}{10} = 4 \quad D_8 = \frac{8 \cdot 20}{10} = 16$$

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3,25	4,25	4,25	4,25	4,25	4,75	5	5,25	5,75	6	6	6	6,5	8	8,25	8,5	8,5	8,75	9,5

Aquests decils són, doncs, 4,25 i 8,25 en aquesta distribució.

Finalment, podem determinar tots els quartils:

$$Q_1 = \frac{1 \cdot 20}{4} = 5$$

$$Q_2 = \frac{2 \cdot 20}{4} = 10$$

$$Q_3 = \frac{3 \cdot 20}{4} = 15$$

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3,25	4,25	4,25	4,25	4,25	4,75	5	5,25	5,75	6	6	6	6,5	8	8,25	8,5	8,5	8,75	9,5

Els tres quartils són, en aquesta distribució, 4,25, 5,75 i 8, respectivament.

Amb l'SPSS podem demanar tots els quartils, tots els decils (*puntos de corte para 10 grupos iguales*), els percentils que desitgem i altres indicadors (*puntos de corte para ? grupos iguales*). Més endavant, a la figura, podreu veure el procés pel qual es demanen tots els indicadors d'una distribució amb aquest paquet estadístic.

No obstant això, els *outputs* sempre ens ofereixen percentils, ja que l'equivalència entre els indicadors és força senzilla.

ESTADÍSTICS

Nota global		
N	Vàlids	706
	Perduts	0
Percentils	10	3,167
	20	3,944
	25	4,589
	30	4,944
	40	5,167
	50	5,474
	60	5,736
	70	6,089
	75	6,314
	80	6,667
	90	7,350

2.2.2. Indicadors de tendència central

S'anomenen *indicadors de tendència central* d'una distribució perquè fan referència al lloc on es concentren les dades (normalment, en la part central de la distribució: valors mitjans, valors centrals...).

Hi ha diverses tipologies, les més comunes de les quals són: la mitjana aritmètica, la mediana i la moda. Cadascuna té avantatges i desavantatges segons les dades i l'objectiu que es persegueix.

- La **mitjana aritmètica** ($\bar{X}$): es correspon amb la idea usual que tenim del valor mitjà i equival a la suma de totes les dades d'una distribució dividida pel nombre de casos. Es representa pel símbol $\bar{X}$ per a la mostra i μ per a la població.

Com es calcula?

Amb la calculadora científica: tecla $\bar{X}$

Es calcula segons es troben les dades:

- 1) Dades sense agrupar: se sumen tots els valors de les dades i es divideixen entre el nombre de dades.

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

- 2) Dades agrupades: cada dada es multiplica per la seva freqüència absoluta i la suma es divideix pel nombre total de casos.

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

- 3) Dades agrupades en intervals: cada marca de classe es multiplica per la seva freqüència absoluta i la suma es divideix pel nombre total de casos.

$$\bar{X} = \frac{\sum X_i' \cdot f_i}{n}$$

- La **mediana (Md)**: és el valor que ocupa la posició central d'una sèrie estadística ordenada. En altres paraules, és el valor que deixa igual nombre de valors per sobre que per sota, deixant a cada extrem el 50% de la distribució.

Com es calcula?

Atès que aquest indicador equival al centil 50, la forma de calcular la mediana seria fer servir la fórmula assenyalada per als centils i substituir ? (centil, decil o quartil que cerquem) per 50.

- La **moda (Mo)**: és la dada que té la freqüència més alta, o bé que es repeteix un major nombre de vegades. Segons el nombre de dades que tenen la freqüència més alta, podem parlar de distribucions:
 - *Unimodals*. Una única dada amb la freqüència més alta.
 - *Bimodals*. Dues dades no consecutives amb la freqüència més alta.
 - *Multimodals*. Més de dues dades no consecutives amb la freqüència més alta.
 - *Amodals*. No apareix cap moda clara.

Si hi ha dues dades que corresponen a la freqüència més alta, però són consecutives, es fa la mitjana per a calcular la moda. La distribució és, en aquest cas, unimodal.

EXEMPLE:

Farem servir els resultats de la prova que hem utilitzat anteriorment, ja ordenats:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3,25	4,25	4,25	4,25	4,25	4,75	5	5,25	5,75	6	6	6	6,5	8	8,25	8,5	8,5	8,75	9,5

Podem calcular la mitjana aritmètica:

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

$$\overline{X} = \frac{3 + 3,25 + 4 \times 4,25 + 4,75 + 5 + 5,25 + 5,75 + 3 \times 6 + 6,5 + 8 + 8,25 + 2 \times 8,5 + 8,75 + 9}{20}$$

$$\overline{X} = \frac{120}{20} = 6$$

Per tant, la mitjana aritmètica d'aquesta distribució és 6.

Per determinar la mediana, hem de trobar el segon quartil (que també és el cinquè decil i el cinquantè centil):

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3,25	4,25	4,25	4,25	4,25	4,75	5	5,25	5,75	6	6	6	6,5	8	8,25	8,5	8,5	8,75	9,5

Així, doncs, la mediana d'aquesta distribució és 5,75.

Per acabar, trobarem la moda buscant el valor més repetit:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3,25	4,25	4,25	4,25	4,25	4,75	5	5,25	5,75	6	6	6	6,5	8	8,25	8,5	8,5	8,75	9,5

És a dir, que tenim una distribució unimodal amb una moda igual a 4,25.

Amb l'SPSS podem demanar aquests tres indicadors de tendència central. Més endavant, en la figura corresponent podreu veure el procés pel qual es demanen tots els indicadors d'una distribució amb aquest paquet estadístic.

No obstant això, en el cas de la moda ens dona un únic valor, tant si es tracta de distribucions unimodals, bimodals, com multimodals. Concretament, quan estem davant de distribucions amb més d'una moda, en l'*output* se'ns indica quan es tracta de distribucions multimodals i se'ns ofereix el valor modal més petit.

ESTADÍSTICS

Nota global avaluació final

N	Vàlids	706
	Perduts	0
Mitjana		5,391
Mediana		5,474
Moda		2,5

2.2.3. Indicadors de dispersió

Per a descriure una distribució necessitem saber alguna cosa més que la tendència d'agrupació de les dades o la posició que adopten. Les mesures de tendència central no ens donen informació sobre els valors més elevats o els més baixos, sinó només sobre on tendeixen a agrupar-se les dades, però no com es dispersen.

En aquest sentit, les mesures de dispersió ens informen de com es distribueixen els valors al voltant dels centrals (si en són lluny o propers). És a dir, ens indiquen la homogeneïtat o heterogeneïtat de les dades. Per tant, amb aquests valors podem arribar a descriure com és la variabilitat o dispersió d'un grup.

Estudiarem els indicadors següents:

- **Rang, recorregut o amplitud (R):** és la diferència entre els valors màxim i mínim de la distribució.

Com es calcula?

Utilitzem les fórmules següents:

$$R = V_{\text{màx}} - V_{\text{mín}}$$

$$\text{Intervals: } R = V_{\text{màx}} - V_{\text{mín}} + 1$$

Aquesta mesura és senzilla de calcular i ens dóna una visió intuïtiva de la dispersió de la distribució. No obstant això, no té en consideració els valors intermedis, únicament es basa en els extrems. Per aquest motiu, el rang normalment s'utilitza en combinació amb altres indicadors que tinguin en compte un major volum d'informació. Per exemple, amb el rang interquartílic (RQ, RI) que veurem més endavant.

- **Variància (S^2):** la importància d'aquest indicador és fonamental per a l'estadística inferencial. Per ara, podem afirmar que, a major variància, major dispersió. La variància és la mitjana de les diferències al quadrat de cada valor respecte a la mitjana de la distribució.

Com es calcula?

Amb la calculadora: tecla σ^2 (S és l'estadístic i σ , el paràmetre).

Càlcul similar a la desviació mitjana, però amb els quadrats de les diferències:

$$S^2 = \frac{\sum (X - \bar{X})^2}{n}$$

Dades sense agrupar

$$S^2 = \frac{\sum (X - \bar{X})^2 \cdot f_i}{n}$$

Dades agrupades

$$S^2 = \frac{\sum (X' - \bar{X})^2 \cdot f_i}{n}$$

Dades agrupades en intervals

Com podem veure, les unitats de la variància no són les mateixes que les de la mostra, ja que estem elevat les diferències al quadrat. Això implica certes dificultats en interpretar la magnitud de la dispersió del grup. Per superar aquesta dificultat, disposem de la *desviació típica*, que veurem a continuació.

- La **desviació estàndard, típica o tipus (S , S_x)**: és la mesura de dispersió més utilitzada en l'estadística descriptiva. En expressió matemàtica, és l'arrel quadrada de la variància.

En una corba normal,³ la desviació típica equival a la distància compresa entre la mitjana i el punt d'inflexió (que modifica la corba de còncava a convexa).

Com es calcula?

Amb la calculadora científica: tecla σ .

O bé calculant la variància i la seva arrel quadrada:

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

- **Rang interquartílic (RQ, RI)**: és la diferència entre els quartils tercer i primer. És un índex de dispersió que ens pot servir per evitar l'existència de buits d'informació. El rang interquartílic ens permet obtenir informació quan els valors extrems no són gaire significatius. Ens dóna, bàsicament, informació sobre el 50% central de la distribució.

Com es calcula?

Utilitzem la fórmula següent:

$$RI = Q_3 - Q_1$$

- **Coefficient de variació (CV) de Pearson**: expressa la relació que hi ha entre la desviació típica d'una mostra i la seva mitjana aritmètica.

És una unitat de mesura relativa, no depèn de les unitats de mesura de la variable i, per tant, permet la comparació entre diferents mostres. Es tracta de saber quina mitjana de dues mostres és més representativa, en relació amb les seves desviacions típiques.

Un CV gran ens indica que la mostra està molt desviada; per tant, la mitjana no és gaire representativa. Si és petita, indica que la mitjana és un bon indicador i que la mostra es troba força concentrada al voltant d'aquesta mesura.

Com es calcula?

Utilitzem la fórmula següent:

$$CV = \frac{S_x}{\bar{X}} \cdot 100$$

³ Una corba normal correspon a la representació gràfica d'una variable que segueix la Llei Normal. En el tema següent aprofundirem en la llei normal.

EXEMPLE:

Seguirem amb la mateixa distribució dels exemples anteriors:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3,25	4,25	4,25	4,25	4,25	4,75	5	5,25	5,75	6	6	6	6,5	8	8,25	8,5	8,5	8,75	9,5

Podem obtenir el rang, recorregut o amplitud amb l'expressió següent:

$$R = V_{\max} - V_{\min} = 9,5 - 3 = 6,5$$

El rang d'aquesta distribució és 6,5.

Per calcular la variància, aplicarem la fórmula següent:

$$s^2 = \frac{\sum (X - \bar{X})^2 \cdot f_i}{n}$$

La variància d'aquesta distribució és 3,68.

Per calcular la desviació estàndard, típica o tipus només cal fer l'arrel quadrada de la variància:

$$S = \sqrt{s^2} = \sqrt{3,68} = 1,92$$

La desviació típica d'aquesta distribució és 1,92.

Per tal de determinar el rang interquartílic, haurem de recuperar el valor del primer i tercer quartils, obtinguts anteriorment:

$$RI = Q_3 - Q_1 = 8 - 4,25 = 3,75$$

El rang interquartílic és 3,75.

Per acabar, calcularem el coeficient de variació de Pearson utilitzant la fórmula:

$$CV = \frac{S_x}{\bar{X}} \cdot 100 = \frac{1,92}{6} \cdot 100 = 32$$

El coeficient de variació de Pearson val 32 en aquesta distribució.

Amb l'SPSS podem demanar la majoria dels indicadors de dispersió i els obtindrem en forma de taula. Més endavant, en la figura corresponent, podreu veure el procés pel qual es demanen tots els indicadors d'una distribució amb aquest paquet estadístic.

No obstant això, en el cas de la moda ens dona un únic valor, tant si es tracta de distribucions unimodals, bimodals o multimodals. Concretament, quan estem davant de distribucions amb més d'una moda, en l'*output* se'ns indica quan es tracta de distribucions multimodals i se'ns ofereix el valor modal més petit.

Nota global avaluació final

N	Vàlids	706
	Perduts	0
Desviació típica		1,5502
Variància		2,403
Rang		8,7
Mínim		1,0
Màxim		9,7

2.2.4. Indicadors de forma

Els indicadors de forma són aquells que ens donen informació sobre la forma de la distribució. Ens indiquen si els valors es distribueixen uniformement a dreta i esquerra de la mitjana aritmètica (simetria) i l'alçada de la corba que representa aquesta distribució (curtosi). A continuació, detallem un xic més aquests indicadors:

- **Asimetria:** indica la distribució dels valors al voltant de la mitjana aritmètica (a la seva dreta i esquerra).

Com es calcula?

Es pot valorar a nivell gràfic a partir de la representació dels valors, o bé de forma més exacta, tot i ser una fórmula aproximada, a partir dels coeficients d'asimetria de Pearson (As):

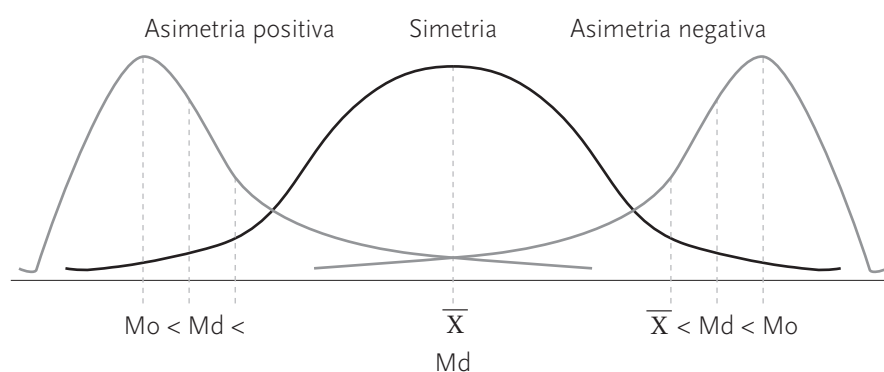
$$As_1 = \frac{\bar{X} - Mo}{S} \quad As_2 = \frac{3 \cdot (\bar{X} - Md)}{S}$$

La interpretació del resultat és la següent:

- $As < -0,5 \rightarrow$ la distribució és asimètrica a la dreta (asimetria negativa).
- $-0,5 < As < 0,5 \rightarrow$ es considera simètrica, perquè molt poques vegades el coeficient serà igual a zero.
- $As > 0,5 \rightarrow$ la distribució és asimètrica a l'esquerra (asimetria positiva).

L'SPSS utilitza una fórmula exacta (coeficient d'asimetria de Pearson) mantenint, però, la mateixa interpretació dels resultats.

Gràficament, ho podem representar com es veu a la figura:



- **Curtosi:** és l'estudi de l'apuntament de la corba en la part central de la distribució, en relació amb la corba normal.

Com es calcula?

Podem mesurar-la analíticament a través del coeficient següent de curtosi percentílic:

$$Cu = \frac{Q_3 - Q_1}{2 \cdot (C_{90} - C_{10})}$$

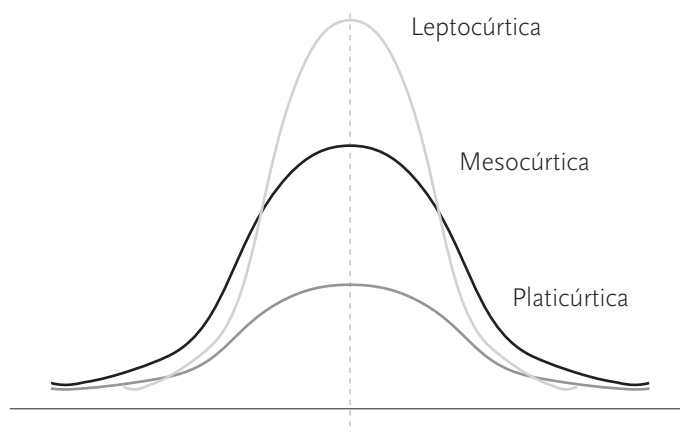
La interpretació del resultat és la següent:

- $Cu < 0,263 \rightarrow$ Platicúrtica (plana)
- $0,263 < Cu < 0,363 \rightarrow$ Mesocúrtica (normal)
- $Cu > 0,363 \rightarrow$ Leptocúrtica (aixecada)

Amb l'SPSS s'utilitza una altra fórmula i s'interpreta de la següent manera: és mesocúrtica quan el coeficient és 0; leptocúrtica quan és més gran de 0 i platicúrtica si és menor de 0. Però pot acceptar-se amb els marges de flexibilitat següents:

- $Cu < -0,5 \rightarrow$ Platicúrtica
- $-0,5 < Cu < 0,5 \rightarrow$ Mesocúrtica
- $Cu > 0,5 \rightarrow$ Leptocúrtica

Gràficament, ho podem representar com es veu a la figura:



EXEMPLE:

Continuarem amb la distribució dels exemples anteriors.

Podem calcular l'asimetria i la curtosi:

$$As_2 = \frac{3 \cdot (\bar{X} - Md)}{S} = \frac{3 \cdot (6 - 5,75)}{1,92} = 0,39$$

$$Cu = \frac{Q_3 - Q_1}{2 \cdot (C_{90} - C_{10})} = \frac{8 - 4,25}{2 \cdot (8,5 - 3,25)} = 0,39$$

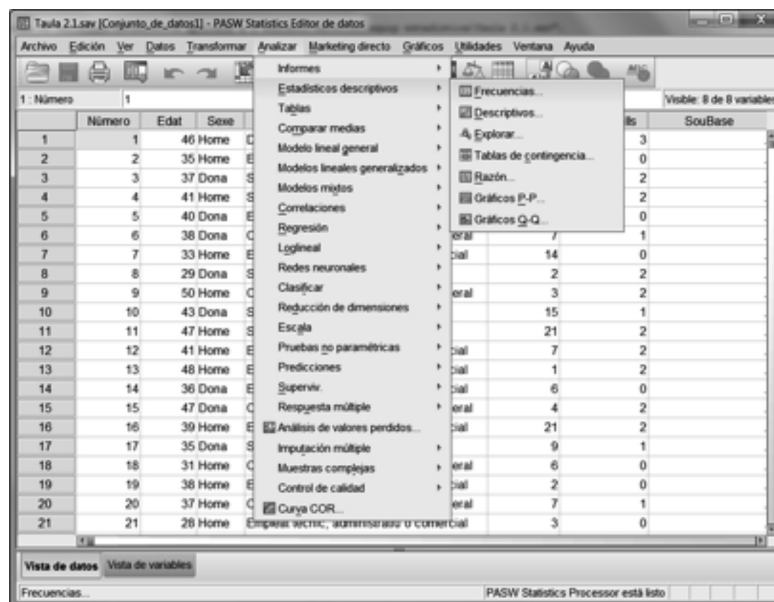
Ens indicaria una distribució simètrica i mesocúrtica.

Amb l'SPSS també podem demanar els indicadors de forma esmentats, de manera semblant als altres indicadors comentats. Més endavant, en la figura corresponent, podreu veure el procés pel qual es demanen tots els indicadors d'una distribució amb aquest paquet estadístic.

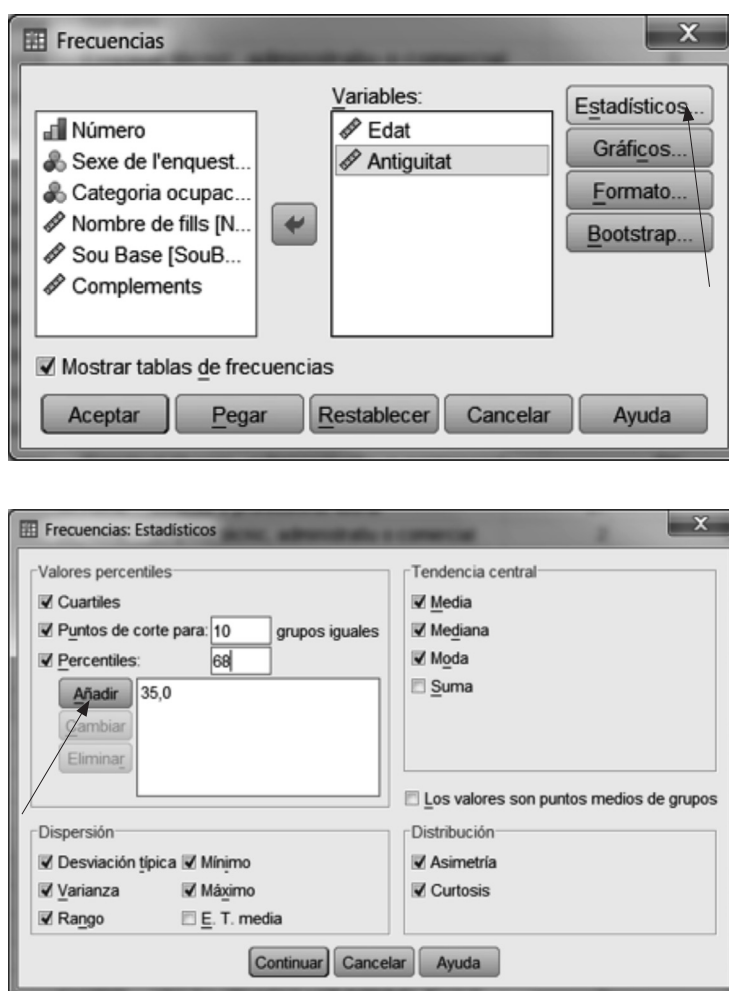
ESTADÍSTICS		
		Nota global
N	Valids	706
	Perduts	0
Asimetria		-,060
Error típic d'asimetria		,092
Curtosi		-,206
Error típic de curtosi		,184

2.2.5. Els indicadors estadístics a l'SPSS

Bona part dels indicadors que hem anat explicant al llarg d'aquest tema es calculen des del paquet estadístic SPSS. Hi ha diverses maneres d'obtenir aquests càlculs des de l'SPSS: una de les més senzilles és des de l'opció *frecuencias*, seguint les passes que es detallen a la figura següent:



I llavors:



Proposem una atenció especial en l'organització del quadre de diàleg *Frecuencias: Estadísticos*. Aquest quadre es divideix en quatre parts, distingint els quatre tipus d'indicadors que hem treballat:

- A la part superior esquerra s'identifiquen alguns *indicadors de posició*: quartils, decils⁴ i percentils.⁵
- A la part superior dreta s'identifiquen alguns *indicadors de tendència central*: mitjana aritmètica, mediana i moda.
- A la part inferior esquerra s'identifiquen alguns *indicadors de dispersió*: la desviació típica, la variància, el rang, així com els valors mínim i màxim de la distribució.
- A la part inferior dreta s'identifiquen alguns *indicadors de forma*: l'asimetria i la curtosi.

Una vegada sabem com demanar i calcular aquests indicadors estadístics, cal aprofundir en com interpretar els resultats obtinguts, aspecte que proposem en l'apartat següent.

⁴ Recordem que els decils poden obtenir-se indicant «10» en els: *puntos de corte para* ___ *grupos iguales*.

⁵ Podem demanar qualsevol percentil concret, tenint en compte d'*afegir* cada vegada que en demanem un de nou.

2.2.6. Com descriure una variable utilitzant els indicadors estadístics

En aquest apartat ens proposem posar a la pràctica tots els conceptes que s'han anat treballant al llarg d'aquest tema (els indicadors estadístics). D'aquesta manera, utilitzarem un exemple per poder aplicar els indicadors de tendència central, de posició, de dispersió i de forma. Pretenem extreure una descripció coherent i integrada de les dades, facilitant la interpretació dels resultats dels diferents càlculs estadístics aplicats.

De les diferents variables de la matriu, hem seleccionat la variable edat. Com que es tracta d'una variable quantitativa contínua, podem utilitzar els diferents indicadors estadístics treballats per fer la descripció de l'edat del grup. Concretament, fent ús de l'SPSS demanem els indicadors següents:

- Tendència central: mitjana, mediana i moda
- Posició: decils i quartils
- Dispersió: desviació típica, variància, rang o amplitud, valor mínim i valor màxim.
- Forma: asimetria i curtosi

A la taula següent trobem els resultats que s'obtenen de l'SPSS sobre els indicadors abans esmentats.

Edat		
N	Vàlids	1.188
	Perduts	42
Mitjana		17,9545
Mediana		18,0000
Moda		18,00
Desviació típica		,90656
Variància		,82186
Asimetria		1,673
Error típic d'asimetria		,071
Curtosi		5,354
Error típic de curtosi		,142
Rang		9,00
Mínim		15,00
Màxim		24,00
Percentils	10	17,0000
	20	17,0000
	25	17,0000
	30	18,0000
	40	18,0000
	50	18,0000
	60	18,0000
	70	18,0000
	75	18,0000
	80	18,0000
	90	19,0000

D'aquestes dades, en podem realitzar la proposta de descripció següent. En primer lloc, s'ha de destacar que disposem d'informació sobre l'edat de 1.188 dels 1.230 subjectes; tenim, doncs, quaranta-dos valors perduts.⁶ La mitjana del grup és de 17,96 anys, i els 18 anys és l'edat més freqüent. De fet, el valor que deixa a dreta i esquerra el 50% de la distribució també és el dels 18 anys. Aquesta proximitat entre els tres indicadors de tendència central podria fer-nos pensar que es tracta d'una distribució simètrica. No obstant això, el coeficient d'asimetria ens indica que la distribució és asimètrica positiva. Aquest fet, doncs, pot ser causat perquè es tracta d'un grup molt poc dispers, fet que podem verificar calculant el coeficient de variació de Pearson (5%), tot i que també podem observar el resultat del coeficient de curtosi que ens indica que la distribució és leptocúrtica, és a dir, que la majoria dels individus es troben al voltant de la mitjana. És atesa aquesta poca dispersió del grup que tot i que l'edat oscil·la entre els 15 i els 24 anys, el 80% té entre 17 i 18 anys. Concretament, la mitjana de les diferències entre les edats de les persones del grup no arriba a un any.

2.3. Exercicis resolts

1. Motivació i actitud dels estudiants: la prova SSHA (Survey of Study Habits and Attitudes) és una prova psicològica que valora la motivació i l'actitud dels estudiants vers la carrera. Una universitat privada sotmet els seus estudiants a la prova SSHA una mostra de cinquanta alumnes de primer curs. Els resultats són:

154	109	137	115	140	140	154	178	101	103
103	126	137	165	165	129	178	148	126	115
109	129	178	137	115	101	148	137	154	140
178	126	165	129	101	154	126	154	137	140
129	103	101	165	178	148	137	154	140	103

- Fes la tabulació de les dades següents: la taula ha de contenir les freqüències absolutes individual (**fi**) i acumulada (**fa**); les freqüències relatives en proporció individual (**pi**) i acumulada (**pa**), i en percentatge individual (**Pi**) i acumulat (**Pa**).
- Digues quina és la mitjana d'aquestes dades mitjançant la fórmula del càlcul de mitjanes.
- Calcula de nou la mitjana, ometent el valor atípic 178 en el càlcul. Quin resultat obtenim?
- Calcula la mediana i la moda de les dades. Quin és el valor central de la distribució? La mediana és major o menor que la mitjana?
- Calcula la variància i la desviació típica de la distribució anterior.
- Calcula els indicadors de forma de la distribució anterior, comenta si la distribució és simètrica o no.

⁶ Recordem que hi ha valors perduts pel sistema (caselles buides) i valors perduts per l'usuari (definit per l'usuari com a no vàlids).

SOLUCIONS:

a) Tabulació

X	fi	fa	pi	pa	Pi	Pa
101	4	4	0,08	0,08	8,0	8,0
103	4	8	0,08	0,16	8,0	16,0
109	2	10	0,04	0,20	4,0	20,0
115	3	13	0,06	0,26	6,0	26,0
126	4	17	0,08	0,34	8,0	34,0
129	4	21	0,08	0,42	8,0	42,0
137	6	27	0,12	0,54	12,0	54,0
140	5	32	0,10	0,64	10,0	64,0
148	3	35	0,06	0,70	6,0	70,0
154	6	41	0,12	0,82	12,0	82,0
165	4	45	0,08	0,90	8,0	90,0
178	5	50	0,10	1,00	10,0	100,0

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

$$\bar{X} = \frac{(101 \cdot 4) + (103 \cdot 4) + (109 \cdot 2) + (115 \cdot 3) + (126 \cdot 4) + (129 \cdot 4) + (137 \cdot 6) + (140 \cdot 5) + (148 \cdot 3) + (154 \cdot 6) + (165 \cdot 4) + (178 \cdot 5)}{50} = 134,72$$

$$c) \bar{X} = \frac{404 + 309 + 218 + 345 + 504 + 516 + 822 + 700 + 444 + 924 + 660}{45} = 129,91$$

$$d) \text{Me} = 137 + 137/2 = 137$$

$$\text{Mo} = 137 \text{ i } 154 \text{ (bimodal)}$$

$$e) S^2 = \frac{\sum (X - \bar{X})^2}{n} = 135,48/50 = 2,71$$

$$S = \sqrt{S^2} = \sqrt{2,71} = 1,65$$

$$f) \text{As} = 3 \cdot \left(\frac{X - \text{Md}}{S} \right) = 3 \cdot (134,72 - 137)/1,65 = -4,14$$

És < -0,5 i, per tant, parlarem d'una distribució asimètrica a la dreta o negativa en la qual trobem un major nombre de dades agrupades en els valors més elevats de la distribució.

En aquest capítol es presenten alguns conceptes bàsics sobre la noció de la probabilitat i s'aprofundeix en el càlcul de probabilitats amb variables quantitatives contínues. Finalitza el capítol amb la llei normal i la prova de normalitat Kolmogorov-Smirnov.

3.1. Conceptes bàsics

L'estadística moderna, és a dir, la que no es limita a la simple descripció d'unes dades, està basada en la teoria de la probabilitat. El seu estudi s'inicià al Renaixement i es va anar desenvolupant de manera que al segle XVIII ja es tenia elaborat un corpus propi i suficient. Sense entrar en un coneixement profund d'aquesta teoria, sí que és convenient exposar alguns conceptes que ajuden a comprendre millor l'estadística inferencial.

3.1.1. Noció intuïtiva de probabilitat

En la nostra vida quotidiana, tots utilitzem en major o menor grau el concepte de probabilitat. El metge recepta una medicina al malalt i li diu que al cap d'un parell de dies es trobarà millor; quasi tothom sap que, com més columnes s'omplen d'una travessa, existeixen més possibilitats d'encertar-ne una de catorze, o, almenys, una de dotze. També se sap que, com més números es juguen a la loteria, és més probable que toqui algun premi, que es tenen més probabilitats d'aprovar un examen si s'ha preparat el 80% dels temes que si tan sols se n'ha estudiat la meitat, etc.

Però poques vegades ens parem a pensar com podem formular matemàticament aquestes probabilitats i calcular-ne el valor. Així, el metge ens acostuma a dir: « Si es pren aquesta medecina, té un 90% de probabilitats de curar-se»; també és veritat que pocs jugadors de traveses ni tan sols saben que, combinant els resultats d' 1×2 , es poden obtenir 3^{14} columnes diferents, etc.

En qualsevol cas, podem afirmar que la majoria de la gent té una noció intuïtiva de la probabilitat, noció que utilitza normalment en la seva vida quotidiana, però de fet quasi ningú arriba a formular-la matemàticament.

3.1.2. Noció clàssica de probabilitat

Sense ànim d'entrar en la discussió que pot originar la teoria clàssica, en un principi podem afirmar que, per definició:

La probabilitat, a priori, que esdevingui un succés X és igual al quocient entre el nombre de casos favorables i el nombre de casos possibles, malgrat ser tots ells igualment possibles.

$$pr_{(x)} = \frac{h}{n}$$

On la **h** és el nombre de casos favorables i la **n** el nombre de casos possibles.

EXEMPLE:

Si tenim una urna amb deu boles exactament iguals en tot menys en el color, cinc de les quals són blanques, tres són negres i dues són vermelles i traiem una bola a l'atzar, quina és:

a) La probabilitat d'extreure una bola blanca (l'anomenarem B)?

Plantejament: casos possibles: $n = 10$; casos favorables: $h_{(B)} = 5$

Aplicació fórmula: $pr_{(B)} = h_{(B)}/n = 5/10 = 0,5$

Conclusió: la probabilitat d'extreure una bola blanca és 0,5. Empíricament podem comprovar que, si fem un gran nombre d'extraccions, tornant a col·locar a l'urna cada vegada la bola extreta (**mostreig no exhaustiu**), aproximadament el 50% de la vegades haurem extret una bola blanca (com més extraccions realitzem, més s'acostarà aquesta proporció al 50%).

b) La probabilitat d'extreure una bola negra (l'anomenarem N)?

Plantejament: casos possibles: $n = 10$; casos favorables: $h_{(N)} = 3$

Aplicació fórmula: $pr_{(N)} = h_{(N)}/n = 3/10 = 0,3$

Conclusió: la probabilitat d'extreure una bola negra és, doncs, 0,3.

3.1.3. Noció axiomàtica de probabilitat

Podem definir d'alguna manera la probabilitat d'un succés en funció d'uns axiomes o principis que es compleixen en tots els casos.

Axiomes:

1) La probabilitat d'un succés és un nombre real comprès entre zero i un.

$$0 \leq pr_{(x)} \leq 1$$

2) La probabilitat d'un succés segur (Ω , és a dir, el que conté tots els casos possibles) és igual a un i la probabilitat d'un succés impossible ($\emptyset$, és a dir, el que no conté cap cas possible) és igual a zero.

$$\begin{aligned} pr_{(\Omega)} &= 1 \\ pr_{(\emptyset)} &= 0 \end{aligned}$$

3) La probabilitat d'un succés qualsevol (x) és igual a la suma de les probabilitats dels successos elementals que el componen (els successos elementals són sempre incompatibles entre si).

$$\text{Si } x = \{x_1, x_2\} \text{ } pr_{(x)} = pr_{(x_1)} + pr_{(x_2)}$$

Consegüentment, per arribar a definir la probabilitat d'un succés qualsevol és necessari definir la probabilitat de cada un dels successos elementals del conjunt Ω , complint els postulats o axiomes anteriors.

Per això, hem de repartir el valor 1 entre tots els successos elementals possibles, d'acord amb certes hipòtesis que no són sempre les mateixes, sinó que depenen de l'experiència aleatòria o del fenomen estudiat.

La definició axiomàtica de probabilitat només dóna els mateixos resultats que la definició clàssica quan tots els successos elementals de l'experiència tenen una probabilitat idèntica (dau equilibrat, moneda perfecta, urna amb boles que només es diferencien pel color, etc.).

3.1.4. Noció experimental de probabilitat

Independentment de la teoria de probabilitat, necessitem recolzar-nos en certes hipòtesis per poder atribuir probabilitats a cada un dels esdeveniments elementals associats a una experiència aleatòria. En aquest sentit, ens podem trobar amb tres situacions:

- Unes vegades disposem d'aquestes hipòtesis.
- Altres vegades aquestes hipòtesis són més o menys arbitràries.
- D'altres, no les tenim.

En aquests dos últims casos, podem estimar la probabilitat d'un succés X a partir d'un estudi estadístic d'observació sobre la freqüència relativa del succés, sempre que l'experiència aleatòria es repeteixi diverses vegades en les mateixes condicions.

Si fem l'experiència n vegades i h_x vegades observem el succés X , la freqüència relativa (o proporció) de X val:

$$fr_{(x)} = \frac{h_x}{n}$$

Aquesta freqüència relativa de X es pot considerar com una estimació ($\wedge$) de la probabilitat de X . Es demostra experimentalment que, com més vegades es repeteix l'experiència (major és n), més tendeix la freqüència relativa del succés X cap a la probabilitat real de X . Així, doncs, la noció experimental de probabilitat d'un succés recolza, d'una banda, en la noció estadística experimental de freqüència relativa (proporció) d'aquest succés en una sèrie n de proves i, d'altra banda, en el fenomen experimental de regularitat estadística o llei dels grans nombres.

a) Freqüència relativa d'un succés

La freqüència observada $h_{(x)}$ d'un succés aleatori (X), al llarg d'un nombre n de proves, és el nombre de vegades que aquest succés s'ha presentat en aquestes n proves.

b) Regularitat estadística o llei dels grans nombres

Les freqüències relatives de qualsevol succés no són sempre les mateixes i, quan el nombre de proves (n) és petit, poden presentar grans oscil·lacions. Però, si augmentem el nombre de proves i segons la mesura en què ho fem, les freqüències relatives tendeixen a estabilitzar-se al voltant d'un valor fix que és la probabilitat de tal succés.

Quan manquem d'hipòtesis que ens permetin saber amb exactitud la probabilitat de cada succés elemental o quan les hipòtesis no són segures, és més correcte utilitzar la noció experimental de probabilitat. Això es pot il·lustrar amb l'exemple següent:

EXEMPLE

Si ens preguntem per la probabilitat de néixer home a Espanya en un determinat any, tindrem la temptació de basar-nos en la hipòtesi d'equiprobabilitat dels dos sexes, segons la qual:

$$\text{pr}(V) = \frac{1}{2} = 0,5$$

Però aquesta hipòtesi no és exacta, perquè realment aquell any van néixer més homes que dones; almenys això és el que ens diuen les xifres de l'Anuari Estadístic d'Espanya, segons el qual hi va haver 73.064 naixements, 37.289 dels quals van ser homes. A partir d'aquestes dades, podem estimar la probabilitat (experimental) de néixer home, que serà:

$$\hat{\text{pr}}(V) = \frac{37.289}{73.064} = 0,51$$

Aquesta estimació s'acosta més a la probabilitat real que la que deduïm de la hipòtesi primera d'equiprobabilitat d'ambdós sexes.

3.2. Llei de probabilitat de variables aleatòries discretes

Variable aleatòria: és aquella que es defineix en associar a cada succés elemental d'una experiència aleatòria un nombre real segons una determinada llei.

Variable aleatòria discreta: si, en associar a cada esdeveniment elemental un nombre real, la variable aleatòria només pot prendre determinats valors dins de certs límits.

EXEMPLE:

Sigui una població composta per 200 boles, que només es diferencien en el nombre, que es distribueixen així:

30 boles amb el número 1
50 boles amb el número 2
20 boles amb el número 3
60 boles amb el número 4
40 boles amb el número 5

Com és obvi, es tracta d'una variable aleatòria discreta, ja que només pot prendre els valors 1, 2, 3, 4, 5. A cada extracció a l'atzar, la variable prendrà un d'aquests cinc valors.

Es defineix una llei de probabilitat d'una variable aleatòria quan associem a cada valor la seva probabilitat corresponent. En aquest cas, aplicant la fórmula general de probabilitat:

$$\text{fr}_{(x)} = \frac{h_x}{n}$$

Seguint amb l'exemple:

- La probabilitat d'obtenir el valor 1 és: $pr_{(x=1)} = 30/200 = 0,15$
- La probabilitat d'obtenir el valor 2 és: $pr_{(x=2)} = 50/200 = 0,25$
- La probabilitat d'obtenir el valor 3 és: $pr_{(x=3)} = 20/200 = 0,10$
- La probabilitat d'obtenir el valor 4 és: $pr_{(x=4)} = 60/200 = 0,30$
- La probabilitat d'obtenir el valor 5 és: $pr_{(x=5)} = 40/200 = 0,20$

Amb aquestes probabilitats hem definit la llei de probabilitat de la variable aleatòria discreta X. La suma de les probabilitats associades a tots els possibles valors de la variable X val sempre 1.

El conjunt de tots els possibles valors d'una variable aleatòria X (discreta, en aquest cas) i de les probabilitats associades a cada un dels possibles valors rep el nom de *Llei de probabilitat de la variable X*.

LA FUNCIÓ DENSITAT I LA FUNCIÓ DISTRIBUCIÓ

a) Funció densitat o densitat de probabilitat ($pr_{(x_i)}$).

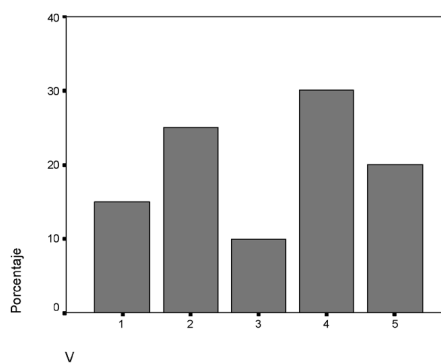
És la funció que indica la probabilitat que la variable aleatòria X prengui un valor igual a x_i . És a dir,

$$pr_{(x_i)} = pr_{(X = x_i)}$$

S'expressa generalment en forma de taula. En l'exemple que hem proposat seria:

X	$Pr_{(x)}$
1	0,15
2	0,25
3	0,10
4	0,30
5	0,20
	1,00

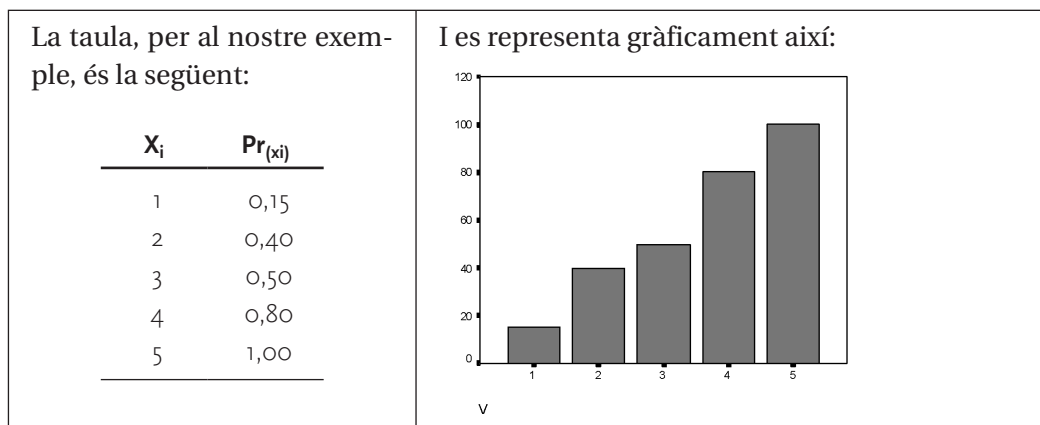
I es representa gràficament així:



b) Funció distribució ($Pr_{(x_i)}$).

És la funció que indica la probabilitat que la variable aleatòria X prengui un valor igual o menor que x_i . És a dir,

$$Pr_{(x_i)} = pr_{(X \leq x_i)}$$



També podem expressar les dues funcions en una mateixa taula:

x_i	$pr_{(x_i)}$	$Pr_{(x_i)}$
1	0,15	0,15
2	0,25	0,40
3	0,10	0,50
4	0,30	0,80
5	0,20	1,00

MODELS DE PROBABILITAT DE VARIABLES ALEATÒRIES DISCRETES

Hi ha diversos models de probabilitat de variables aleatòries discretes. En recerca educativa —en general en ciències humanes o socials—, les més freqüents són la *Llei binomial o de Bernouilli* i la de *Poisson* (és un cas particular de la binomial, aplicable a successos «rars», que tenen una probabilitat d'ocurrència molt baixa). La *Llei binomial* convergeix molt ràpidament en la *Llei Normal*, tant més ràpidament com més simètrica.

3.3. Llei de probabilitat de variables aleatòries contínues

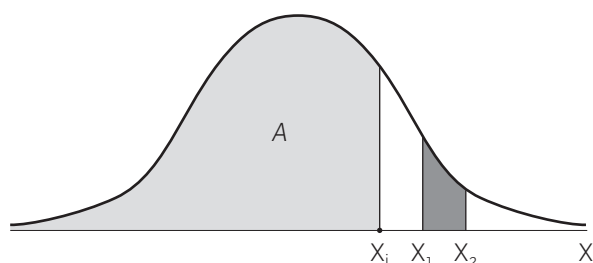
Variable aleatòria contínua: una variable és contínua si, en associar a cada esdeveniment elemental un nombre real, la variable aleatòria pot prendre qualsevol valor dins de certs límits.

Per exemple, la variable talla, fins i tot dins d'un interval petit com és un centímetre, pot prendre qualsevol valor. Entre 173 cm i 174 cm, la variable talla pot tenir infinits valors.

Això implica que hem de matisar el concepte de *funció densitat o densitat de probabilitat*: no és la probabilitat que la variable adopti un determinat valor ($pr_{(x = x_i)}$), perquè el quocient entre casos favorables i possibles —com que hi ha infinits valors possibles— ens donaria sempre zero.

Per això la funció densitat es defineix com la probabilitat que la variable adopti un valor compès dins d'un interval ($pr_{(x_1 \leq x \leq x_2)}$).

El concepte de funció distribució és el mateix: la probabilitat que la variable tingui un determinat valor o inferior ($Pr_{(x_i)} = pr_{(x \leq x_i)}$).

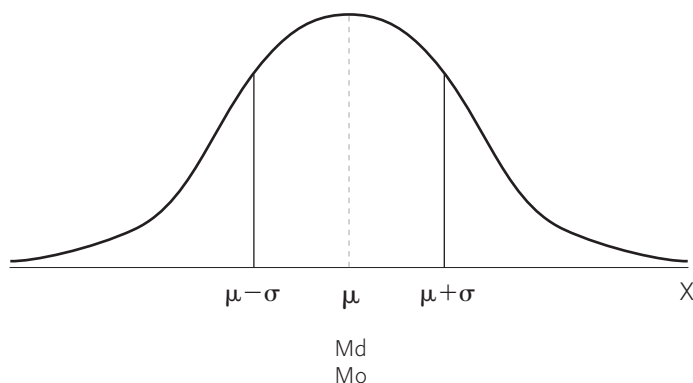


MODELS DE PROBABILITAT DE VARIABLES ALEATÒRIES CONTÍNUES

Existeixen multitud de models de probabilitat de variables aleatòries contínues. Les més freqüentment utilitzades són la Llei Normal, la Llei X², la Llei de Student-Fisher i la Llei F de Snedecor. La més coneguda i més àmpliament utilitzada és la Llei Normal.

3.3.1. La Llei Normal

Concepte: Es diu així perquè es va observar, ja fa molt temps, que la representació gràfica de les variables contínues, quan la n era prou gran, normalment tendia a adoptar una determinada figura en forma de campana (per això es diu també campana de Gauss), tal com la figura següent:



Van ser els matemàtics Laplace i Gauss qui, ja al segle XIX, van formular, independentment, la funció matemàtica d'aquesta gràfica.

$$f_{(x)} = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma_x^2}}$$

Característiques:

- És una corba simètrica respecte a un eix en el qual hi ha la mitjana, la mediana i la moda.
- Té dos punts d'inflexió (on hi ha un canvi de curvatura):

$$1) x = \mu - \sigma_x, \quad 2) x = \mu + \sigma_x.$$

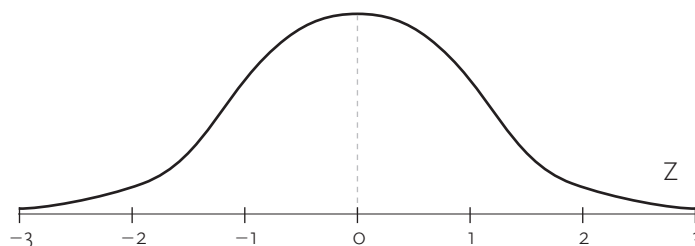
- És asimptòtica respecte a l'eix d'abscisses (va fins a l'infinit, gairebé tangent, però sense tallar-lo).

- Queda definida per dos paràmetres: mitjana (μ) i variància (σ^2).
- En termes de probabilitat, l'àrea total sota la corba és igual a 1.
- Es pot resoldre, a partir de la fórmula, qualsevol qüestió de probabilitat que es plantegi, resolent en cada cas la integral o confeccionant la taula corresponent. Aquesta taula només serviria per a una mitjana i variància determinades, perquè hi ha tantes lleis normals com possibles valors de la mitjana i la variància.

Però si fem servir les puntuacions estàndard o típiques, totes les variables que segueixen la Llei Normal, sigui quina sigui la seva mitjana i la seva variància, es converteixen en una única variable, que sempre té per mitjana 0 i per variància 1, anomenada Llei Normal estàndard.

LLEI NORMAL ESTÀNDAR («TÍPICA», «REDUÏDA», «CENTRADA I REDUÏDA»)

La Llei Normal estàndard sempre té una mitjana igual a 0 i una variància (també la desviació típica) igual a 1. Té les mateixes característiques que qualsevol Llei Normal, llevat que oscil·la, al voltant de 0, i va des de $-\infty$ fins a $+\infty$. Gràficament; seria la següent:

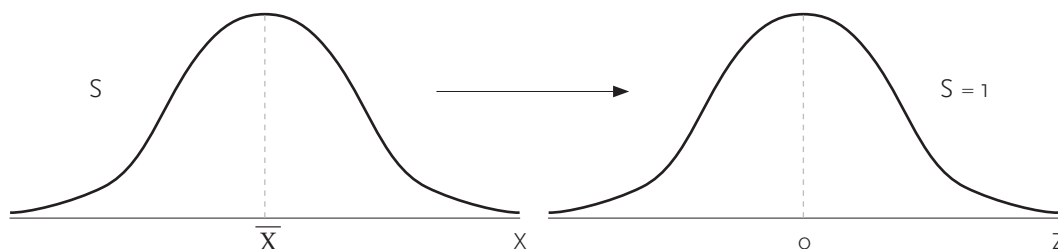


Per a aquesta distribució, sí que hi ha una taula que ens permet solucionar problemes de probabilitat sense haver de resoldre integrals. Hi ha diverses modalitats de presentació. La que utilitzarem ens dona la funció distribució, és a dir, l'àrea que queda per sota d'una determinada puntuació Z.

APLICACIONS A DISTRIBUCIONS QUE SEGUEIXEN LA LLEI NORMAL

També es pot utilitzar la taula de la *Llei Normal estàndard* per resoldre problemes de probabilitat de qualsevol variable X que segueixi la Llei Normal, sigui quina sigui la seva mitjana i la seva variància, sense necessitat de calcular la integral corresponent.

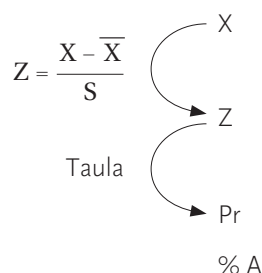
Només cal convertir la puntuació directa (X_i) en puntuació estàndard (Z_i):



Mitjançant la fórmula:

$$Z = \frac{X - \bar{X}}{S}$$

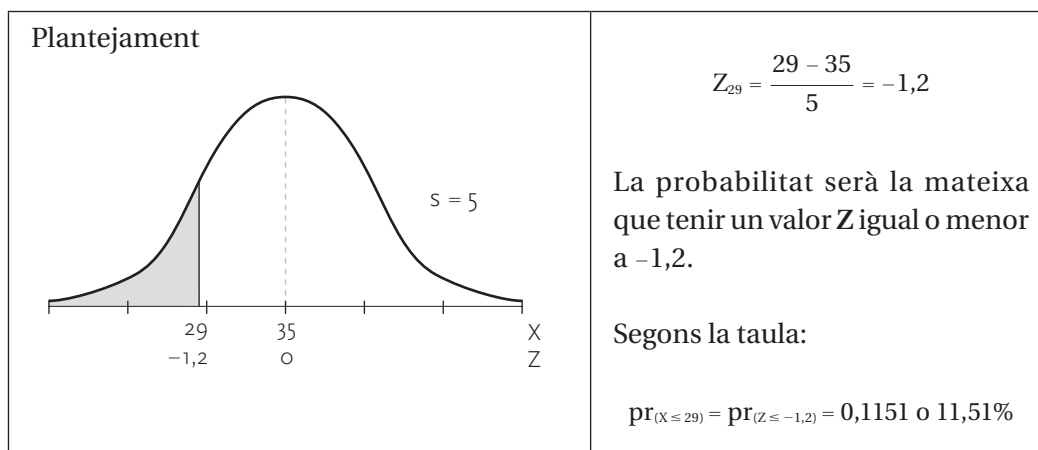
I seguir l'algorisme següent:



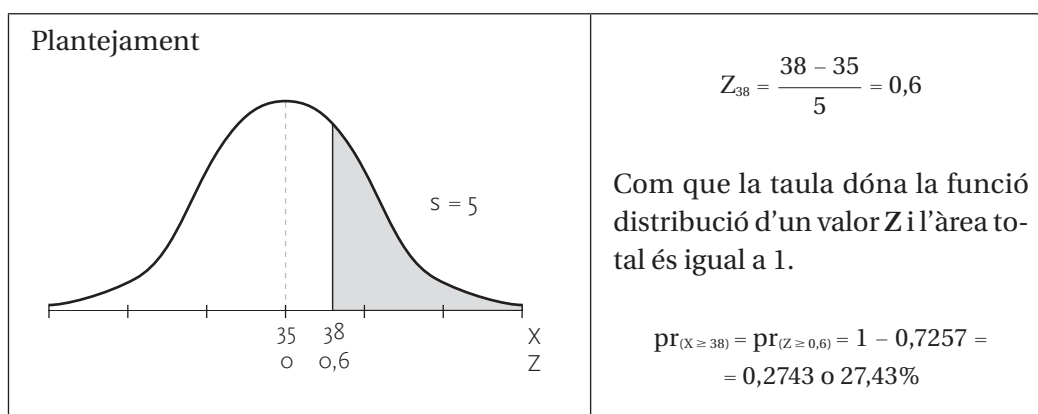
Sigui, **per exemple**, la variable rendiment acadèmic de ciències, essent aquesta una variable que segueix la llei normal, amb una mitjana de 35 i una variància de 25 (desviació típica = 5).

Ens plantegem les preguntes següents:

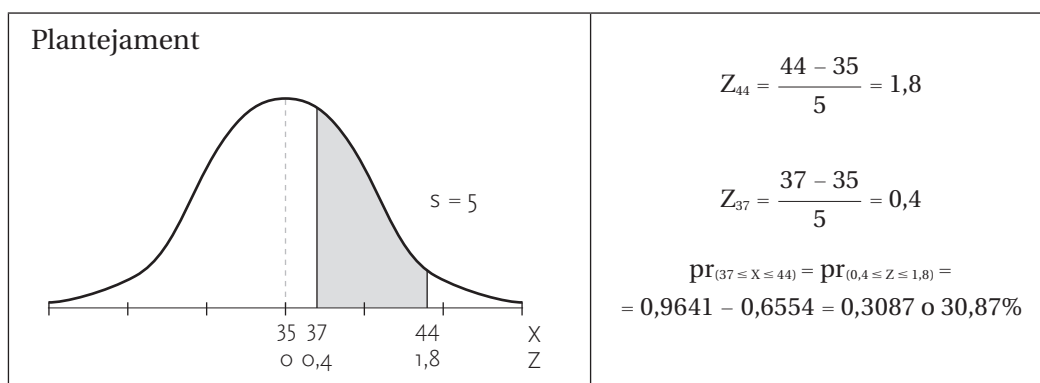
- a) Quina és la probabilitat que un subjecte escollit a l'atzar tingui un valor igual o menor a 29?



- b) Quina és la probabilitat que un subjecte escollit a l'atzar tingui un valor igual o major a 38?



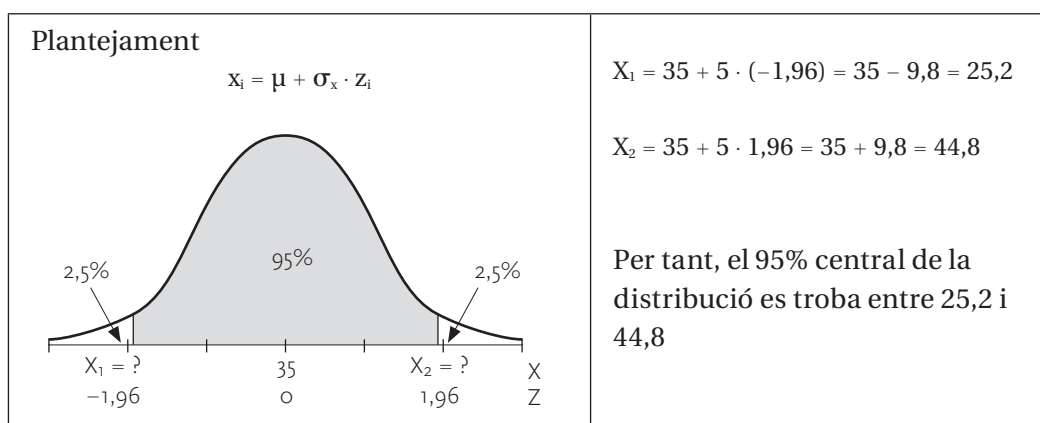
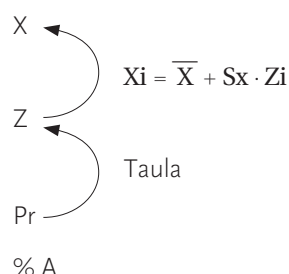
- c) Quina és la probabilitat que un subjecte escollit a l'atzar tingui un valor comprès entre 37 i 44?



- d) Entre quins valors estarà comprès el 95% (central) de la distribució?

El problema és l'invers dels tres anteriors. Es tracta de calcular l'interval (límits inferior i superior, X_1 i X_2) que conté un determinat percentatge de la distribució. Per això, primer hem de buscar els valors z (Z_1 i Z_2) que contenen aquest 95% central ($-1,96$ i $1,96$) i després calcular la puntuacions X corresponents a aquests valors Z .

En aquest cas, seguirem l'algorisme següent:



3.3.2. Proves de normalitat (Kolmogorov)

Per poder realitzar les aplicacions anteriors (i per aplicar diverses proves estadístiques), la condició és que **la distribució segueixi la Llei Normal** (en realitat, que provingui d'una població en la qual la variable segueix la Llei Normal). Per això, hem de realitzar una *prova de normalitat*.

Hi ha diverses proves i la d'ús més freqüent és *la prova de Kolmogorov*. Estudia, d'una banda, la distribució empírica o observada, i d'altra banda, la distribució teòrica normal. Poden ser més o menys semblants, però difícilment una distribució en una mostra serà igual que la distribució teòrica normal. Com menors siguin les diferències, més s'assemblarà la distribució empírica a la teòrica.

El que fa la prova de Kolmogorov és calcular i analitzar les diferències entre la *distribució empírica o observada* i la *distribució teòrica (normal)*; per a no estudiar totes les diferències, es calculen les àrees acumulades, tant de la distribució empírica (p_a) com de la teòrica (s_a) i s'estudia només la diferència màxima $|d_i \text{ màx.}|$. Aquesta diferència màxima es pren en valor absolut (perquè el que importa no és si la diferència és en més o en menys, sinó la magnitud d'aquesta diferència) i es compara amb el valor de la taula corresponent (Taula de la prova de *Kolmogorov*). La taula dona els valors màxims admissibles en funció del valor de «n» i del nivell de significació α (se sol utilitzar el 0,05). Poden ocórrer dos casos:

- Si $|d_i \text{ màx.}| \leq$ que el valor de la taula, direm que es pot acceptar com a normal.
- Si $|d_i \text{ màx.}| >$ que el valor de la taula, direm que no es pot acceptar com a normal.

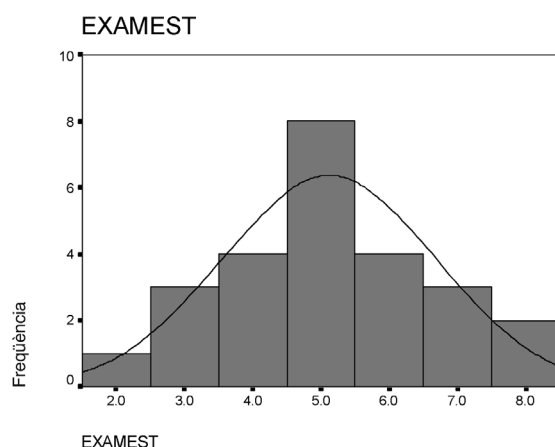
Vegem-ho amb un exemple: disposem de les puntuacions en un examen d'Estadística d'un grup de 25 alumnes i volem verificar (amb un nivell de confiança o un valor $\alpha = 0,05$) si aquesta distribució prové d'una població que segueix la llei normal en aquesta variable.

Les puntuacions són:

5, 4, 7, 5, 4, 6, 5, 6, 3, 5, 5, 2, 5, 4, 6, 8, 5, 4, 7, 3, 5, 7, 3, 6, 8

x_i	f_i	f_a	p_i	p_a	$z_i = (L_i - \bar{X})/s_x$	s_a	$d_i = p_a - s_a $
2	1	1	0,04	0,04	-1,71	0,0436	0,0036
3	3	4	0,12	0,16	-1,06	0,1446	0,0154
4	4	8	0,16	0,32	-0,40	0,3446	0,0246
5	8	16	0,32	0,64	0,25	0,5987	0,0413
6	4	20	0,16	0,80	0,90	0,8159	0,0159
7	3	23	0,12	0,92	1,55	0,9394	0,0194
8	2	25	0,08	1	2,21	0,9864	0,0136
25			1				

(En negreta, la distribució empírica o observada, i en cursiva, la distribució teòrica normal. També en negreta la $|d_i \text{ màx.}|$. O diferència màxima en valor absolut.)

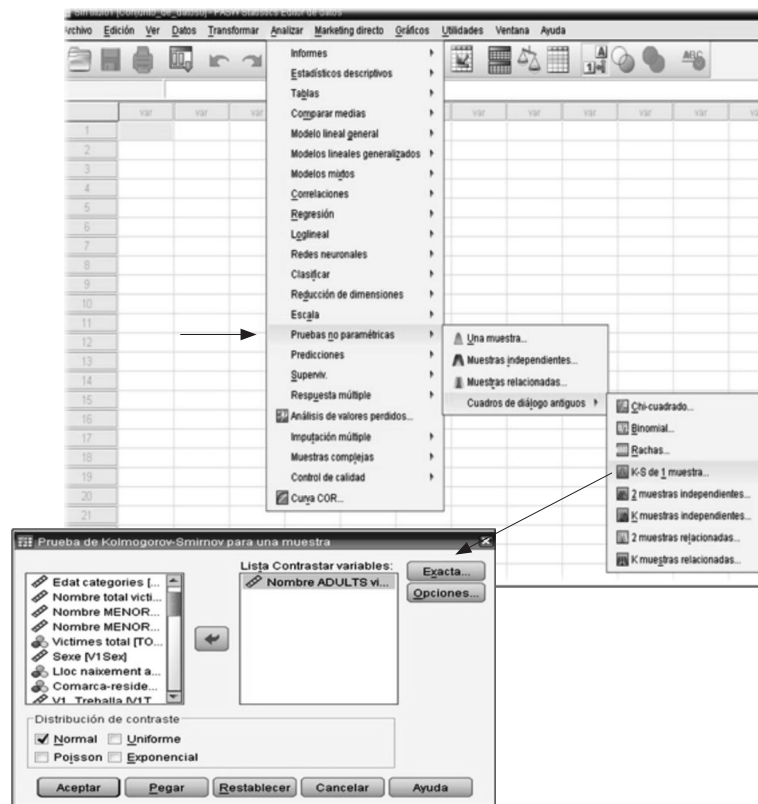


Si busquem a la taula «Prova de Kolmogorov» (columna Lilliefors) el valor corresponent per a una $n = 25$ i $\alpha = 0,05$, trobem 0,180 (diferència màxima admissible per acceptar que en la població origen la variable segueix una llei normal).

Com que la nostra diferència màxima: 0,0413 és menor que 0,180, podem acceptar que segueix la Llei Normal.

LA PROVA DE NORMALITAT AMB L'SPSS

Amb l'SPSS també podem verificar la normalitat d'una determinada distribució, seguint, per a això, els passos marcats en la figura següent:



Com podem observar, quan accedim al quadre de diàleg de la prova de Kolmogorov-Smirnov, hem de seleccionar aquelles variables a les quals volem sotmetre a aquesta prova, així com el tipus de distribució amb què volem contrastar les variables, i en el cas que ens ocupa la distribució de contrast és la normal, ja que el nostre objectiu, tal com hem comentat anteriorment, és verificar la normalitat de la distribució.

Com interpretem l'output?

Per interpretar el resultat obtingut d'aplicar la prova de normalitat, ens hem de centrar en la diferència absoluta, ja que fa referència a la diferència màxima en valor absolut. Podem observar que tant comptem en l'*output* amb la diferència màxima (o extrema) positiva com amb la negativa.

Des del moment que obtenim la diferència màxima, el procés a seguir serà el mateix que hem comentat anteriorment, és a dir, la comparem amb el valor de la taula de la prova de Kolmogorov i es pren la decisió en funció de:

- Si $|di\ màx.| \leq$ que el valor de la taula, direm que es pot acceptar com a normal.
- Si $|di\ màx.| >$ que el valor de la taula, direm que no es pot acceptar com a normal.

PROVA DE KOLMOGOROV-SMIRNOV PER A UNA MOSTRA

		Puntuació
N		48
Paràmetres normals ^{a, b}	Mitjana	10,00
	Desviació típica	5,174
Diferències més extremes	Absoluta	,106
	Positiva	,088
	Negativa	-,106
Z de Kolmogorov-Smirnov		,735
Sig. asintòt. (bilateral)		,652

a. La distribució de contrast és la normal.

b. S'han calculat a partir de les dades.

En l'*output* mostrat, observem que la diferència màxima absoluta és 0,106; tot seguit busquem a la taula «Prova de Kolmogorov» (columna Lilliefors) el valor corresponent per a una $n = 48$ i $\alpha = 0,05$ i trobem que quan la mida de la mostra és superior a 35 hem d'aplicar

$$\frac{0,886}{\sqrt{n}},$$

de manera que la diferència màxima admissible per acceptar que en la població origen la variable segueix una llei normal és 0,128.

Com que la nostra diferència màxima: 0,106 és menor que 0,128, podem acceptar que segueix la Llei Normal.

Finalment, hem de fer referència al fet que no hi ha una única interpretació de la normalitat mitjançant l'*output* obtingut amb el SPSS, ja que també es pot arribar a la mateixa conclusió interpretant el valor de significació (*sig. asintòt. (bilateral)*) que apareix en el resultat. Aquest concepte, l'introduïrem en el tema d'Estadística Inferencial.

3.4. Taules

- Taula de la Llei Normal.
- Taula de Kolmogorov-Smirnov.

TAULA DE LA LLEI NORMAL

	0	1	2	3	4	5	6	7	8	9
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002	.0002
-3.3	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0005	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0008	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0014	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0238	.0233
-1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0300	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0570	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4620	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Font: DOMÈNECH, J. M. (1988). *Tablas de estadística*. Barcelona: Herder.

	0	1	2	3	4	5	6	7	8	9
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9430	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9648	.9656	.9664	.9671	.9678	.9686	.9693	.9700	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9762	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9874	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9986	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9996
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998

TAULA DE KOLMOGOROV-SMIRNOV

Mida de la mostra (n)	Massey		Lilliefors	
	$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$
1	0,975	0,995	---	---
2	0,812	0,929	---	---
3	0,708	0,828	---	---
4	0,624	0,733	0,381	0,417
5	0,565	0,669	0,337	0,405
6	0,521	0,618	0,319	0,364
7	0,486	0,577	0,300	0,348
8	0,457	0,543	0,285	0,331
9	0,432	0,514	0,271	0,311
10	0,410	0,490	0,258	0,294
11	0,391	0,468	0,249	0,284
12	0,375	0,450	0,242	0,275
13	0,361	0,433	0,234	0,268
14	0,349	0,418	0,227	0,261
15	0,338	0,404	0,220	0,257
16	0,328	0,392	0,213	0,250
17	0,318	0,381	0,206	0,245
18	0,309	0,371	0,200	0,239
19	0,301	0,363	0,195	0,235
20	0,294	0,356	0,190	0,231
25	0,270	0,320	0,180	0,203
30	0,240	0,290	0,161	0,187
sup. a 35	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$	$\frac{0,886}{\sqrt{n}}$	$\frac{1,031}{\sqrt{n}}$

Massey s'aplica quan els paràmetres són coneguts.

Lilliefors s'aplica quan es desconeixen els paràmetres.

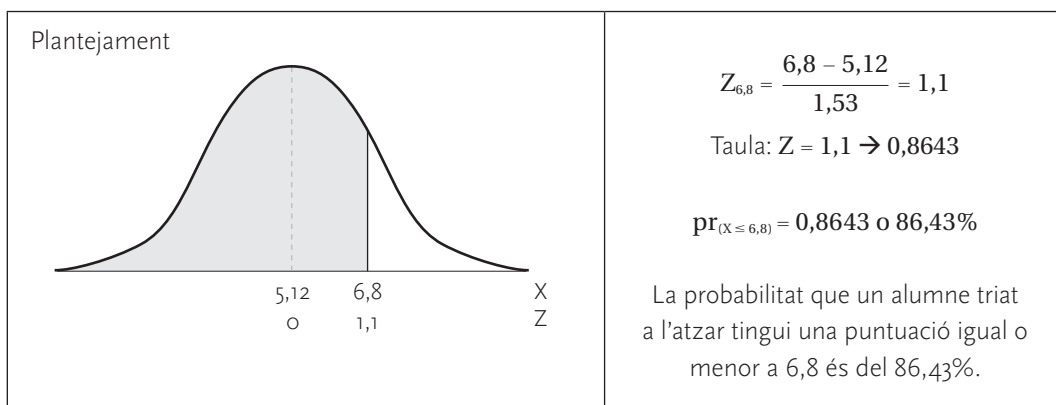
3.5. Exercicis resolts

1. Un grup de matemàtiques de primer de batxillerat segueix la Llei Normal i presenta una mitjana de 5,12 i una desviació típica de 1,5315. Resol les qüestions següents:

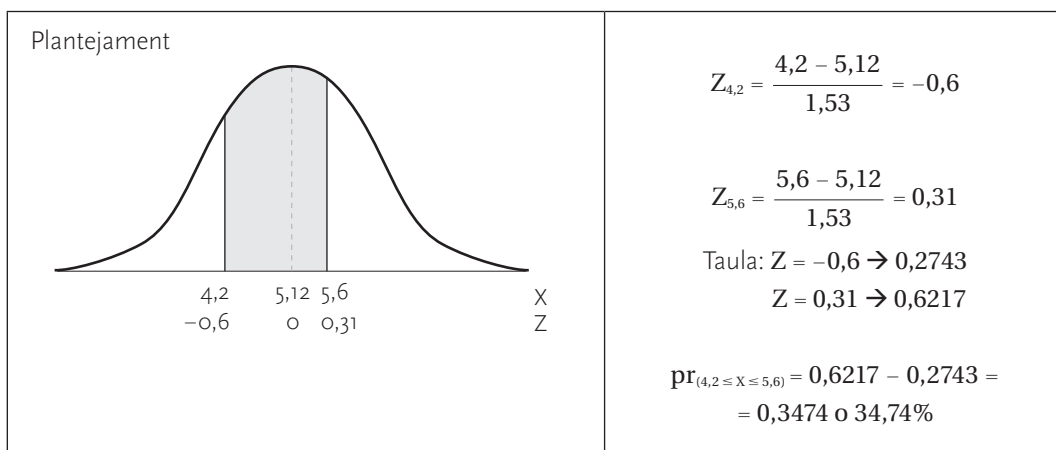
- Probabilitat que si escollim un alumne a l'atzar obtingui una puntuació igual o inferior a 6,8.
- Cerca novament la probabilitat, però per obtenir una puntuació entre 4,2 i 5,6.
- Entre quins valors està comprès el 95% de la distribució?

SOLUCIONS:

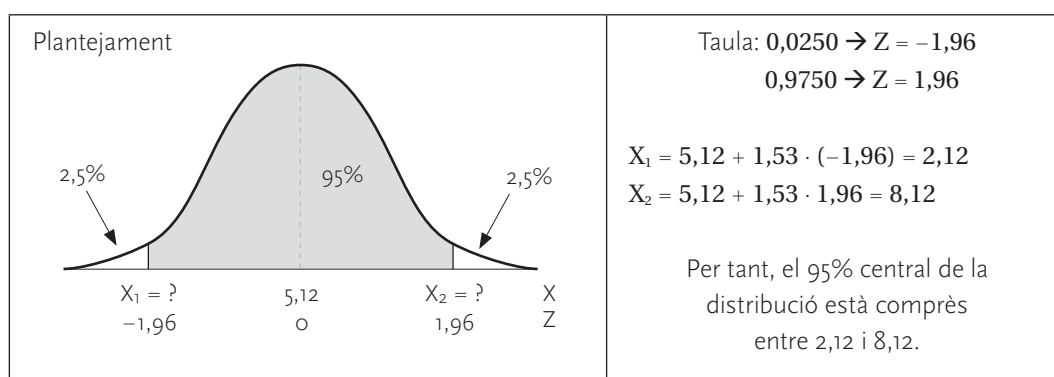
- Probabilitat que si escollim un alumne a l'atzar obtingui una puntuació igual o inferior a 6,8.



- Cerca novament la probabilitat, però per obtenir una puntuació entre 4,2 i 5,6.



c) Entre quins valors està comprès el 95% de la distribució?



2. En realitzar la prova de Kolmogorov d'una distribució de cinquanta pedagogs, la diferència màxima dona un valor de $-0,2154$. Podem acceptar la distribució com a normal?

SOLUCIÓ:

$$|\text{di màx.}| = |-0,2154| \quad n = 50 \quad \alpha = 0,05$$

Taula: «Prova de Kolmogorov» (columna Lilliefors), trobem que, quan la mida de la mostra és superior a 35, hem d'aplicar

$$\frac{0,886}{\sqrt{n}} = \frac{0,886}{\sqrt{50}} = 0,1253$$

$$0,2154 > 0,1253$$

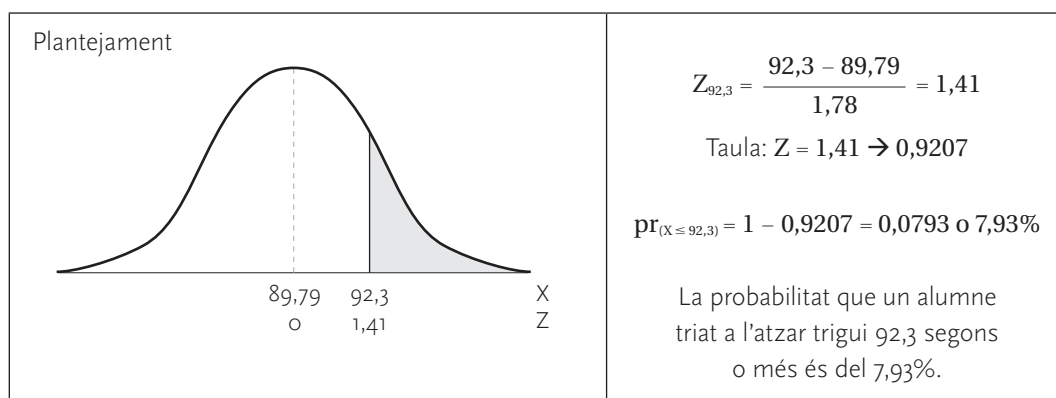
Per tant, direm que la distribució no es pot acceptar com a normal.

3. Una distribució que segueix la Llei Normal expressa el temps utilitzat per cent alumnes en la realització d'una prova (la mitjana és 89,79 seg. i la desviació típica 1,78).

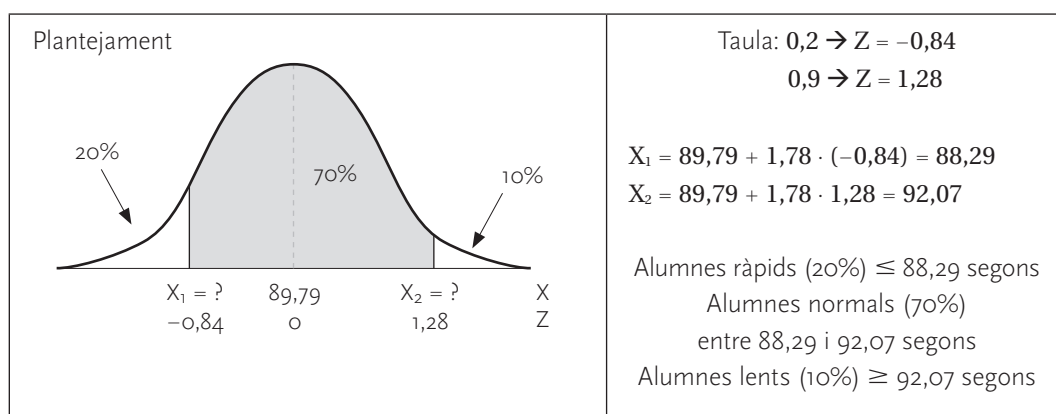
- Calcula la probabilitat de trobar a l'atzar un alumne que trigui 92,3 segons o més a realitzar la prova.
- Si volguéssim classificar els alumnes en tres classes (lents, normals i ràpids) de manera que els lents siguin el 10%, els normals el 70% i els ràpids la resta, quins valors (de temps) haurem d'agafar per establir aquesta classificació?

SOLUCIONS:

- a) Calcula la probabilitat de trobar a l'atzar un alumne que trigui 92,3 segons o més a realitzar la prova.

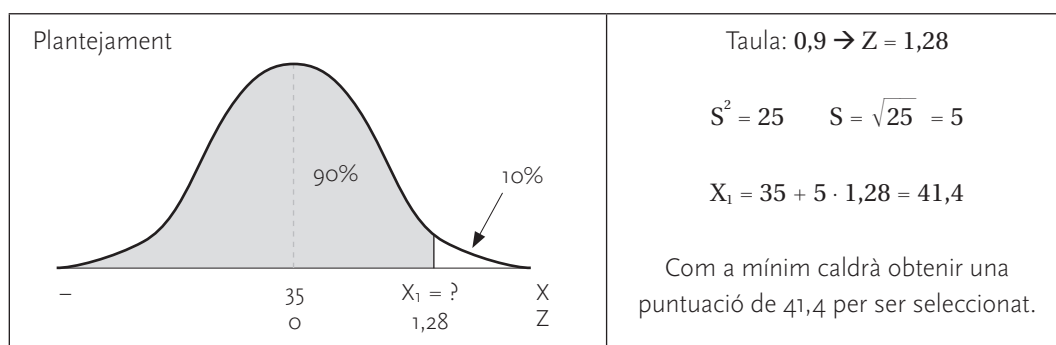


- b) Si volguéssim classificar els alumnes en tres classes (lents, normals i ràpids) de manera que els lents siguin el 10%, els normals el 70% i els ràpids la resta, quins valors (de temps) haurem d'agafar per establir aquesta classificació?



4. Es fa una prova per poder optar a una beca de formació a l'estranger. Aquesta prova ens dóna una mitjana de 35 punts i una variància de 25 (la variable segueix la Llei Normal). Si només podrà optar-hi un 10%, quina puntuació caldrà aconseguir, com a mínim, per ser seleccionat?

SOLUCIÓ:

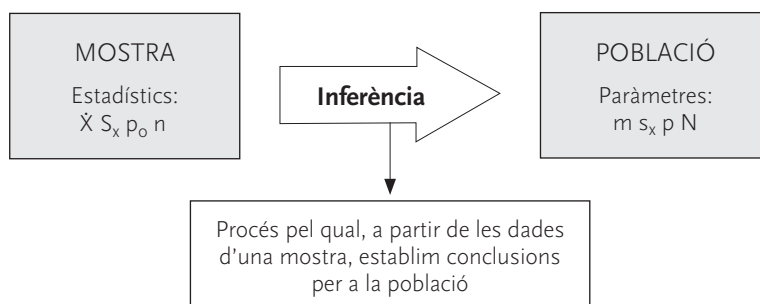


En aquest capítol, es desenvolupen els conceptes bàsics sobre la distribució mostral i es desenvolupen les teories de l'estimació estadística i de la decisió estadística. En la teoria de l'estimació estadística, es calculen els intervals de probabilitat i confiança en mitjanes i proporcions.

4.1. Conceptes bàsics

L'estadística inferencial es basa en la inferència estadística. No es limita a l'estudi purament descriptiu d'unes dades, sinó que a partir de les dades observades en una mostra (d'on obtenim els estadístics) s'infereix, o es generalitza, l'estudi a la població d'aquesta mostra (els paràmetres).

Normalment no es treballa amb totes les unitats o els individus de la població que es vol estudiar, sinó amb mostres extretes d'aquesta població: els elements seleccionats en la mostra han de ser suficientment representatius de la població total per garantir que els resultats obtinguts sobre ella siguin generalitzables i inferibles a tota la població; per això, és molt important realitzar una correcta extracció de la mostra.



Una **mostra** és representativa quan:

- És un subconjunt de la població.
- Té una mida suficient (depèn de la precisió que es desitgi i del risc d'error).
- La tècnica de selecció de la mostra és correcta (mostreig).

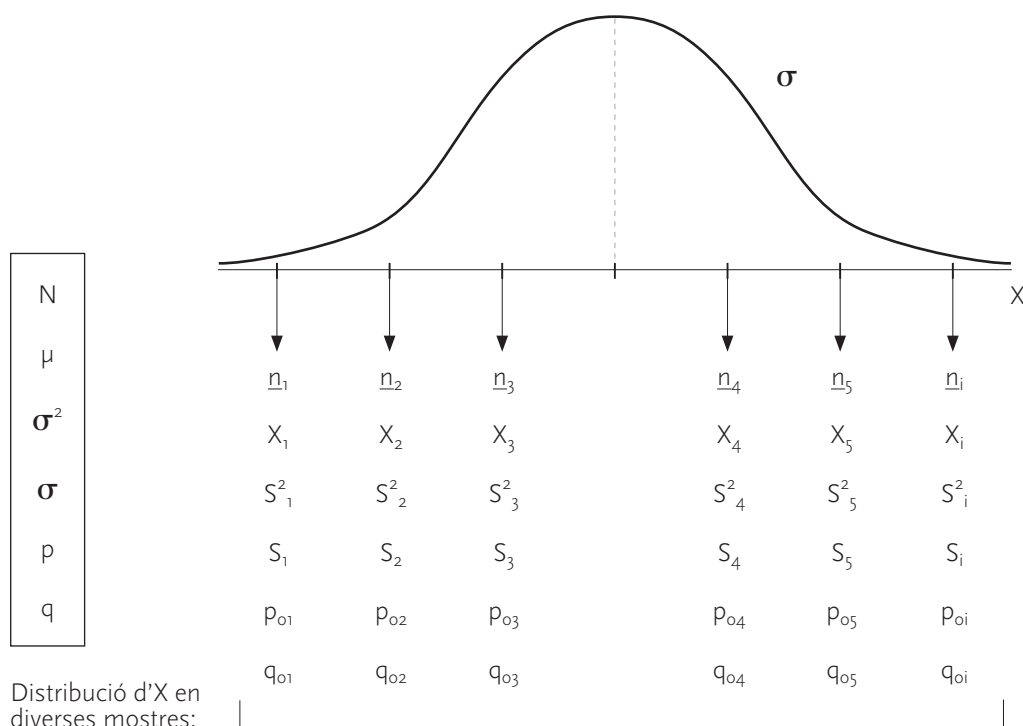
L'**error mostral** és la diferència entre un estadístic (mostra) i el seu paràmetre (població). Pot ser de dos tipus:

- **Aleatori**: és el degut a l'atzar. Es produeix quan fem un mostreig probabilístic o aleatori. Es pot calcular i controlar si es coneix quina llei de probabilitat segueix.
- **De biaix**: es deu a altres causes, a més de l'atzar, i es produeix, entre altres motius, quan utilitzem un mostreig no probabilístic. El problema no és que el biaix sigui gran o petit, sinó que no el podem calcular, ja que no coneixem quina llei segueix.

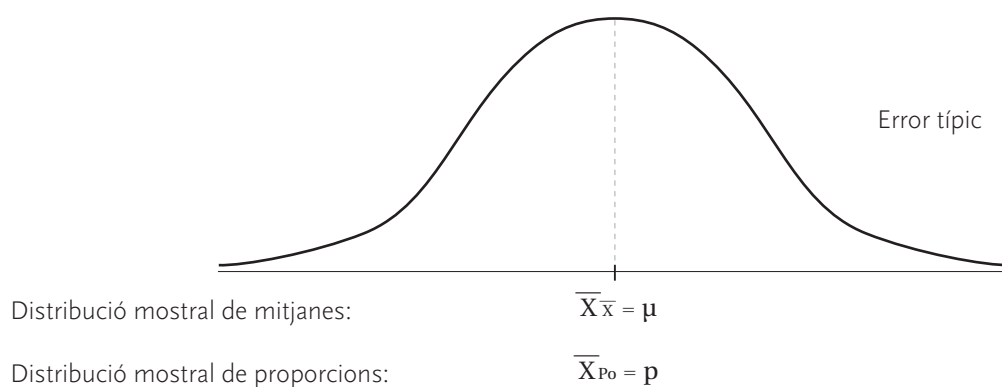
Per calcular l'error mostrat s'utilitza l'**error típic**, que mesura la desviació dels estadístics de la mostra al voltant dels paràmetres de la població (és la desviació típica de la distribució mostrat de l'estadístic que estima el paràmetre poblacional). Es distingeix entre distribució poblacional, distribució d'una mostra i distribució mostrat.

DISTRIBUCIÓ POBLACIONAL

(és la distribució d'una variable en la població)



DISTRIBUCIÓ MOSTRAL D'ESTADÍSTICS



L'**error típic** (o error estàndard) és la desviació típica de la distribució mostrat. I es calcula de la manera següent:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

L'error típic de **mitjanes**

$$\sigma_{p_0} = \sqrt{\frac{p \cdot q}{n}}$$

L'error típic de **proporcions**

Teorema del límit central:

- Els estadístics obtinguts en les mostres tendeixen a agrupar-se al voltant del paràmetre.
- Quan la «n» augmenta, els estadístics estan més agrupats.

Si les mostres són grans ($n \geq 30$) i la població d'origen segueix la Llei Normal, el gràfic de la distribució mostral és simètric i té forma de campana, és a dir, segueix la Llei Normal. També pot seguir altres models: t de Student, χ^2 (khi Quadrat), F de Snedecor.

DISTRIBUCIÓ MOSTRAL**Mitjanes (variables quantitatives):**

- Segueix la Llei Normal.
- La mitjana és el paràmetre (o sigui, la mitjana de la població: μ).
- L'error típic es calcula:

Si la població és infinita:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Si la població és finita:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

N = mida de la població
n = mida de la mostra

Proporcions (variables qualitatives):

- Segueix la Llei Normal.
- La mitjana de les proporcions observades (p_o) és el paràmetre ($\bar{X}_{p_o} = p$).
- L'error típic es calcula:

Si la població és infinita:

$$\sigma_{p_o} = \sqrt{\frac{p \cdot q}{n}}$$

Si la població és finita:

$$\sigma_{p_o} = \sqrt{\frac{p \cdot q}{n}} \sqrt{\frac{N-n}{N-1}}$$

N = mida de la població
n = mida de la mostra

EXEMPLES:

1. En un test de velocitat lectora, la mitjana de paraules en alumnat de segon curs d'educació primària és 10 i la desviació típica és 2. Quin error típic es pot calcular si assumim aquesta mitjana en una classe de 25 alumnes?

$$\mu = 10 \text{ i } \sigma = 2$$

$$\bar{X}_{\bar{x}} = \mu = 10$$

$$\sigma_{\bar{x}} = \frac{2}{\sqrt{25}} = 0,4$$

2. I en un grup de 400 alumnes?

$$\mu = 10 \text{ i } \sigma = 2$$

$$\overline{X_{\bar{x}}} = \mu = 10$$

$$\sigma_{\bar{x}} = \frac{2}{\sqrt{400}} = 0,1$$

3. El 60% dels infants no estan escolaritzats a l'escola bressol. Calcula l'error típic en un grup de 25 infants.

$$p = 0,6 \text{ i } q = 0,4$$

$$\overline{X_{p_0}} = p = 0,6$$

$$\sigma_{p_0} = \sqrt{\frac{0,6 \cdot 0,4}{25}} = 0,0979$$

4. I en un grup de 100 infants?

$$p = 0,6 \text{ i } q = 0,4$$

$$\overline{X_{p_0}} = p = 0,6$$

$$\sigma_{p_0} = \sqrt{\frac{0,6 \cdot 0,4}{100}} = 0,0489$$

5. Què passa amb l'error típic quan augmenta la mida de la mostra?

Observem com l'error típic disminueix a mesura que augmenta n , perquè l'atzar hi influeix menys.

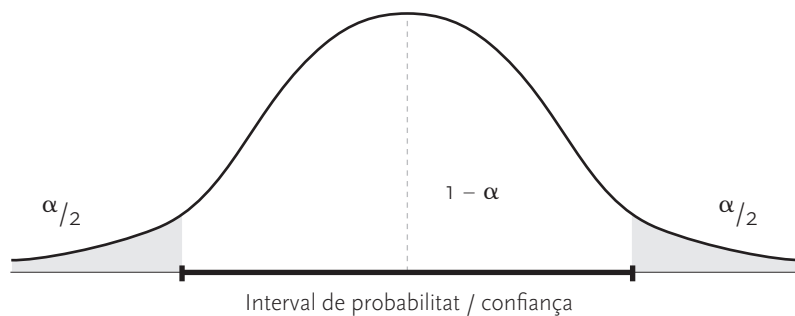
Els dos pilars bàsics de l'estadística inferencial són:

- La Teoria de l'estimació estadística.
- La Teoria de la decisió estadística.

4.2. Teoria de l'estimació estadística

La Teoria de l'estimació estadística tracta de l'estimació de paràmetres a partir dels estadístics. Es basa en els conceptes de distribució mostral i error típic.

Com que moltes de les distribucions mostrals segueixen una Llei Normal, de la qual coneixem la mitjana (el paràmetre) i la desviació típica (error típic o estàndard), podem resoldre qualsevol problema de probabilitat sense necessitat de resoldre integrals, simplement utilitzant la taula de la Llei Normal estàndard (centrada i reduïda: $\mu_Z = 0$, $\sigma_Z = 1$), com feiem en el tema anterior, per a qualsevol distribució X que segueix la Llei Normal. L'algorisme és el mateix, però utilitzant en cada cas la mitjana i la desviació típica corresponents.



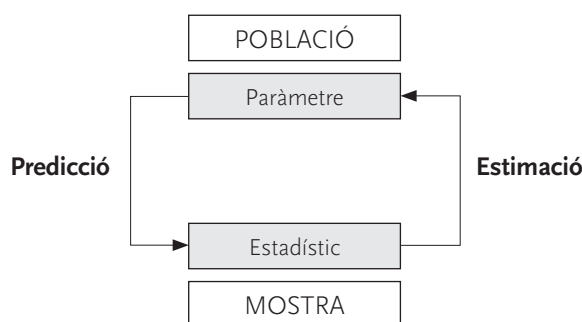
Per conveni, en ciències humanes i socials el risc α sol valer:

$$\alpha = 0,05 \text{ (5\%)} \text{ o } \alpha = 0,01 \text{ (1\%)}$$

Així, doncs,

$$1 - \alpha \begin{cases} \text{Si } \alpha = 0,05 & 1 - \alpha = 0,95 \text{ (95\%)} \\ \text{Si } \alpha = 0,01 & 1 - \alpha = 0,99 \text{ (99\%)} \end{cases}$$

- L'**interval de probabilitat** és un interval (simètric) al voltant d'un paràmetre que conté una proporció $1 - \alpha$ dels estadístics de la distribució mostral. L'interval de probabilitat serveix per resoldre una qüestió de simple **predicció** (dels estadístics a partir del paràmetre), però no per a l'**estimació estadística** pròpiament dita (estimar paràmetres a partir dels estadístics).
- L'**interval de confiança** és un interval (simètric) al voltant d'un estadístic que té una probabilitat $1 - \alpha$ (generalment, 0,95 o 0,99) de contenir el paràmetre. Un interval és un valor lineal que queda definit pels seus dos límits (superior i inferior). Els límits de l'interval de confiança van ser denominats per R. A. Fisher, límits de confiança o límits fiducials.



Dos tipus d'estimació estadística:

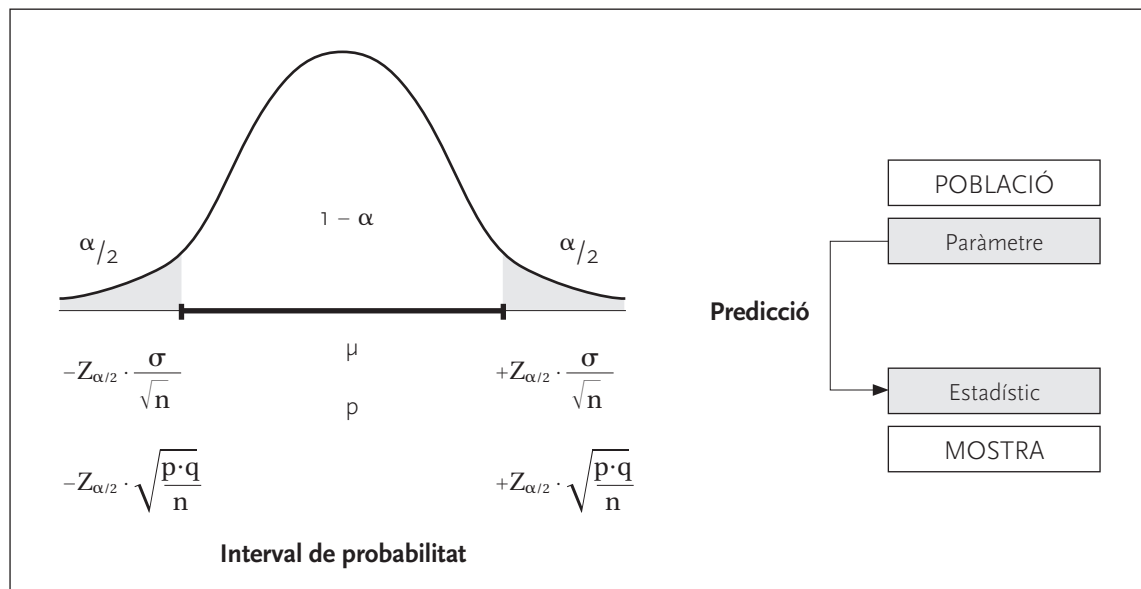
- **Estimació puntual:** a partir d'un estadístic, hem de donar un sol i únic valor per al paràmetre. Es demostra matemàticament que, en general, la millor estimació puntual d'un paràmetre (μ , σ^2 , σ , p) és el seu estadístic corresponent ($\bar{X}$, S^2 , S , p_o), especialment si treballem amb mostres grans ($n \geq 30$).
Aquest tipus d'estimació té l'inconvenient que no indica quina és la precisió d'aquesta estimació. És per aquest motiu que s'utilitza l'estimació per interval.
- **Estimació per interval:** permet estimar els paràmetres a partir del seu estadístic corresponent, amb un risc d'error conegut (i petit).

4.2.1. Interval de probabilitat

L'interval de probabilitat permet predir, amb un risc α (0,05 o 0,01) d'equivocar-se, els límits dins dels quals es trobaran, partint d'un paràmetre poblacional, els $1 - \alpha$ (0,95 o 0,99, 95% o 99%) dels estadístics observats en mostres de grandària n extretes a l'atzar de la població. Dit en altres paraules, l'interval de probabilitat conté dins dels seus límits $1 - \alpha$ (0,95 o 0,99, 95% o 99%) de la distribució mostral.

Amb **variables quantitatives**, és un interval simètric al voltant de la mitjana de la població (μ) que conté una proporció igual a $1 - \alpha$ de les mitjanes observades en mostres de grandària n extretes a l'atzar de la població.

Amb **variables qualitatives**, és un interval simètric al voltant de la proporció de la població (p) que conté una proporció $1 - \alpha$ de les proporcions observades (p_o) en mostres de grandària n extretes a l'atzar de la població origen.



Mitjanes (quantitatives):

$$\bar{X} \in \mu \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Proporcions (qualitatives):

$$p_o \in p \pm Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}}$$

EXEMPLES:

1. S'accepta que els coeficients intel·lectuals es distribueixen en la població segons una Llei Normal de mitjana 100 i desviació típica 15. Calculeu entre quins valors estaran els coeficients intel·lectuals mitjans, en el 95% i el 99% dels casos, en mostres a l'atzar de 100 subjectes i de 50 subjectes.

$$n = 100 \quad \alpha = 0,05$$

$$\bar{X} \in 100 \pm 1,96 \cdot \frac{15}{\sqrt{100}} = 100 \pm 2,94 \quad \begin{array}{l} \rightarrow 102,94 \text{ (màxim)} \\ \rightarrow 97,06 \text{ (mínim)} \end{array}$$

$$n = 100$$

$$\alpha = 0,01$$

$$\bar{X} \in 100 \pm 2,58 \cdot \frac{15}{\sqrt{100}} = 100 \pm 3,87 \quad \begin{array}{l} \rightarrow 103,87 \text{ (màxim)} \\ \rightarrow 96,13 \text{ (mínim)} \end{array}$$

$$n = 50$$

$$\alpha = 0,05$$

$$\bar{X} \in 100 \pm 1,96 \cdot \frac{15}{\sqrt{50}} = 100 \pm 4,16 \quad \begin{array}{l} \rightarrow 104,16 \text{ (màxim)} \\ \rightarrow 95,85 \text{ (mínim)} \end{array}$$

$$n = 50$$

$$\alpha = 0,05$$

$$\bar{X} \in 100 \pm 2,58 \cdot \frac{15}{\sqrt{50}} = 100 \pm 5,47 \quad \begin{array}{l} \rightarrow 105,47 \text{ (màxim)} \\ \rightarrow 94,53 \text{ (mínim)} \end{array}$$

Observem:

- En els dos casos, l'interval $1 - \alpha = 0,99$ és més ampli que $1 - \alpha = 0,95$, ja que ha de contenir una proporció més gran d'estadístics ($99\% > 95\%$).
 - A mesura que disminuïm **n** (mida de les mostres), l'interval de probabilitat es va ampliant o, el que és el mateix, la dispersió dels estadístics és més gran com menor és **n**. Això és així perquè, com més petita és **n**, més hi influeix l'atzar.
2. Sabem que una determinada assignatura és suspesa pel 30% de l'alumnat. Calculeu entre quins valors oscil·laran els percentatges de suspesos ($\alpha = 0,05$) en mostres a l'atzar de 100 i 400 alumnes.

$$n = 100$$

$$a_1 (\alpha = 0,05), 1 - \alpha = 0,95$$

$$p_0 \in 0,3 \pm 1,96 \cdot \sqrt{\frac{0,3 \cdot 0,7}{100}} = 0,3 \pm 0,0898 \quad \begin{array}{l} \rightarrow 0,3898 \rightarrow 38,98\% \\ \rightarrow 0,2102 \rightarrow 21,02\% \end{array}$$

$$n = 400$$

$$b_1 (\alpha = 0,05), 1 - \alpha = 0,95$$

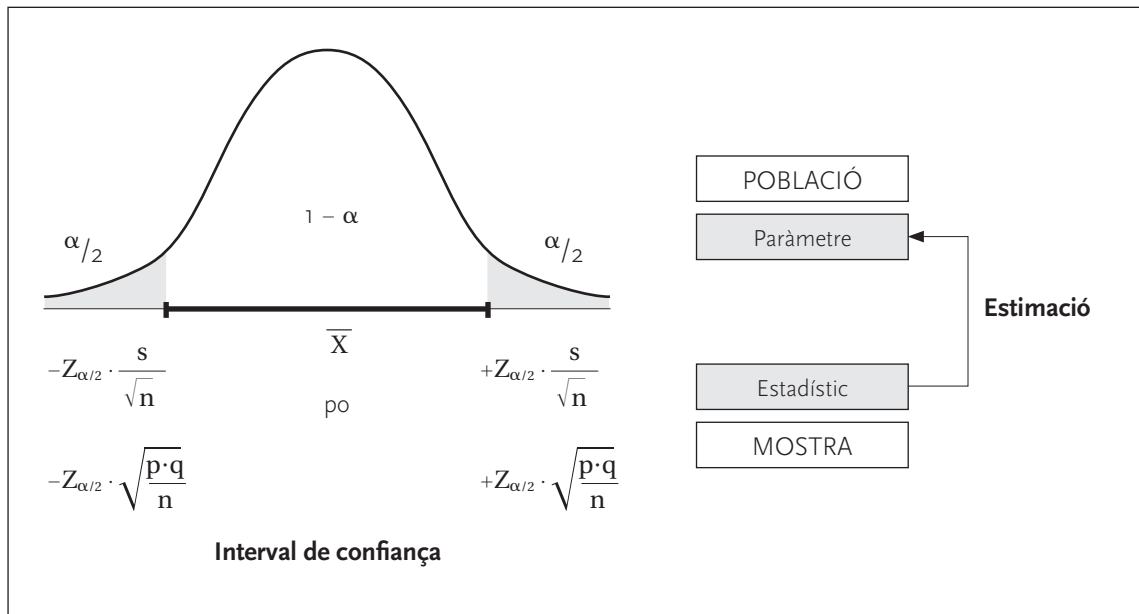
$$p_0 \in 0,3 \pm 1,96 \cdot \sqrt{\frac{0,3 \cdot 0,7}{400}} = 0,3 \pm 0,0449 \quad \begin{array}{l} \rightarrow 0,3449 \rightarrow 34,49\% \\ \rightarrow 0,2551 \rightarrow 25,51\% \end{array}$$

4.2.2. Interval de confiança

També es basa en els conceptes de distribució mostral i error típic. Com que aquesta distribució segueix la llei normal (per a mostres grans), el:

95% ($1 - \alpha = 0,95$) està comprès entre $\pm 1,96 Z$
 99% ($1 - \alpha = 0,99$) està comprès entre $\pm 2,58 Z$ } En general, $\pm Z_{\alpha/2}$

- Amb **variables quantitatives**, és un interval simètric al voltant de la mitjana observada que té una probabilitat $1 - \alpha$ de contenir la mitjana de la població (μ).
- Amb **variables qualitatives**, és un interval simètric al voltant de la proporció observada (p_o) que té una probabilitat $1 - \alpha$ de contenir la proporció de la població (p).



Mitjanes (quantitatives):

$$\mu \in \bar{X} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Proporcions (qualitatives):

$$p \in p_o \pm Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}}$$

Si es desconeixen p i q (que és lògic):

a) S'estimen a partir de dades obtingudes en estudis anteriors.

b) Es realitza un sondeig previ (estudi pilot).

c) S'agafa el cas més desfavorable o màxim ($p = q = 0,5$), perquè el producte $p \cdot q$ és màxim quan $p \cdot q = 0,25$.

d) S'utilitzen els valors observats (p_o i q_o):

$$p \in p_o \pm Z_{\alpha/2} \cdot \sqrt{\frac{p_o \cdot q_o}{n}}$$

Hi ha condicions d'aplicació: mostres grans, és a dir, tots els productes $n \cdot p$, $n \cdot q$, $n \cdot p_o$, $n \cdot q_o$ han de valer com a mínim 5.

EXEMPLES:

- Un grup de cent alumnes d'estadística ha obtingut en una prova una mitjana de 5,8 punts i una desviació típica de 1,2. Estimeu la mitjana de la població origen d'aquest grup. Quina és l'estimació puntual de la mitjana de la població? Quina és l'estimació de la mitjana de la població, amb un nivell de confiança del 95%? I amb un nivell de confiança del 99%?

$$\mu = \bar{X} = 5,8$$

$$\mu \in 5,8 \pm 1,96 \cdot \frac{1,2}{\sqrt{100}} = 5,8 \pm 0,252 \quad \begin{array}{l} \rightarrow 6,0352 \\ \rightarrow 5,5648 \end{array}$$

$$\mu \in 5,8 \pm 2,58 \cdot \frac{1,2}{\sqrt{100}} = 5,8 \pm 0,3096 \quad \begin{array}{l} \rightarrow 6,1096 \\ \rightarrow 5,4904 \end{array}$$

- Fes el mateix, però en un grup de 36 alumnes.

$$\mu = \bar{X} = 5,8$$

$$\mu \in 5,8 \pm 1,96 \cdot \frac{1,2}{\sqrt{36}} = 5,8 \pm 0,392 \quad \begin{array}{l} \rightarrow 6,192 \\ \rightarrow 5,408 \end{array}$$

$$\mu \in 5,8 \pm 2,58 \cdot \frac{1,2}{\sqrt{36}} = 5,8 \pm 0,516 \quad \begin{array}{l} \rightarrow 6,316 \\ \rightarrow 5,284 \end{array}$$

- En un grup de trenta alumnes d'estadística n'han aprovat divuit. Quina serà la proporció d'aprovat de l'assignatura? Fes una estimació puntual i una per interval (1%).

- Estimació puntual:

$$p_o \rightarrow p = \frac{18}{30} = 0,6 \quad p = p_o = 0,6$$

- Estimació interval ($\alpha = 0,01$):

$$p \in 0,6 \pm \sqrt{\frac{0,6 \cdot 0,4}{30}} \cdot 2,58 = 0,6 \pm 0,2308 \quad \begin{array}{l} \rightarrow 0,8308 \\ \rightarrow 0,3692 \end{array}$$

- Com a l'exercici anterior, amb una proporció d'aprovat de 0,6, però en un grup de cent alumnes.

- Estimació puntual: $p = p_o = 0,6$

- Estimació interval ($\alpha = 0,01$):

$$p \in 0,6 \pm \sqrt{\frac{0,6 \cdot 0,4}{100}} \cdot 2,58 = 0,6 \pm 0,1264 \quad \begin{array}{l} \rightarrow 0,7264 \\ \rightarrow 0,4736 \end{array}$$

4.2.3. Grandària de la mostra

Podem calcular la mida que la mostra ha de fer com a mínim per poder realitzar l'estimació del paràmetre. Aquesta mida dependrà de la precisió (error màxim = e) i aquest, per la seva banda, dependrà del risc α que estiguem disposats a admetre. A continuació apareixen les fórmules per calcular la n mínima tant per variables quantitatives (mitjanes) com per qualitatives (proporcions):

	Poblacions infinites	Poblacions finites
Mitjana aritmètica	$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{e} \right)^2$	$n = \frac{N \cdot Z_{\alpha/2}^2 \cdot \sigma^2}{Z_{\alpha/2}^2 \cdot \sigma^2 + e^2 (N - 1)}$
Proporció	$n = p \cdot q \left(\frac{Z_{\alpha/2}}{e} \right)^2$	$n = \frac{N \cdot Z_{\alpha/2}^2 \cdot p \cdot q}{e^2 (N - 1) + (Z_{\alpha/2}^2 \cdot p \cdot q)}$

EXEMPLES:

1. Es vol estimar la mitjana de l'alumnat d'una determinada assignatura. S'accepta que la seva desviació típica és 0,8. Calculeu la mida mínima necessària de la mostra amb un risc d'error α de 0,01 i un error màxim de 0,2.

$$\alpha = 0,01 \text{ i } e = 0,2$$

$$n = \left(\frac{2,58 \cdot 0,8}{0,2} \right)^2 = 106,5024 \rightarrow 107$$

2. I amb un risc α del 5%?

$$\alpha = 0,05 \text{ i } e = 0,2$$

$$n = \left(\frac{1,96 \cdot 0,8}{0,2} \right)^2 = 61,4656 \rightarrow 62$$

3. Es vol estimar la mitjana d'edat dels joves en consumir una determinada substància addictiva. S'accepta que la seva desviació típica és 0,8. Calculeu la mida mínima necessària de la mostra amb un error màxim de 0,1, sabent que la població juvenil en risc és de $N = 512$ ($\alpha = 0,01$).

$$n = \frac{512 \cdot 2,58^2 \cdot 0,8^2}{2,58^2 \cdot 0,8^2 + 0,1^2 (511)} = \frac{2.181,1692}{9,3701} = 232,7797 \rightarrow 233$$

4. En un estudi d'opinió en el qual només s'ha de contestar sí/no, es desconeix la tendència de la població. Es vol saber quants individus serà necessari enquestar per estimar la proporció de resposta amb un error màxim del 4% i una probabilitat de 0,95.

$$n = \frac{Z_{\alpha/2}^2 \cdot p \cdot q}{e^2} = \frac{1,96^2 \cdot 0,5 \cdot 0,5}{0,04^2} = 600,25 \rightarrow 601$$

5. Igual que en el problema anterior, però en el cas que s'hagi estimat que el percentatge de respostes a favor està comprès entre el 60% i el 70%.

$$n = \frac{1,96^2 \cdot 0,6 \cdot 0,4}{0,04^2} = 576,24 \rightarrow 577$$

4.3. Teoria de la decisió estadística

Tracta de la presa de decisions en estadística. També es basa en els conceptes de distribució mostral i error típic.

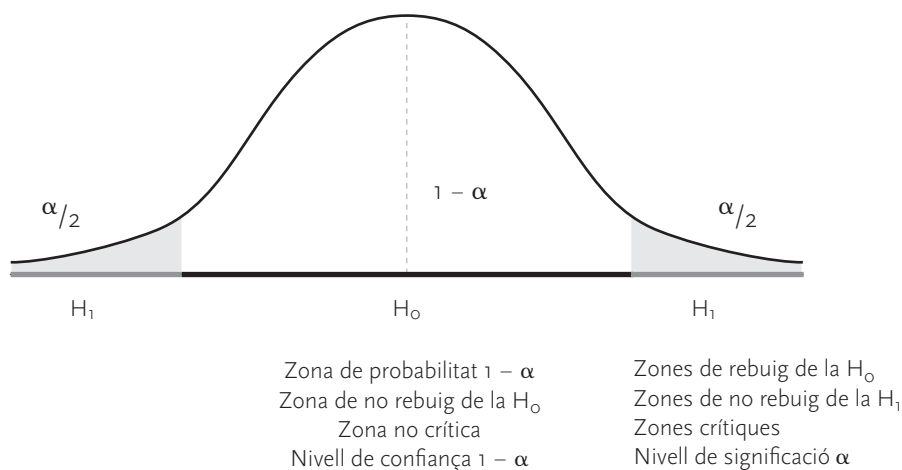
Es diu prova d'hipòtesi o **contrast d'hipòtesi** al procediment mitjançant el qual s'arriba a la decisió de rebutjar o no una determinada hipòtesi.

Hipòtesi: és una proposició o afirmació que relaciona de forma concreta variables, de la qual no coneixem la seva naturalesa amb absoluta certesa i que, justament per això, es planteja en termes de conjectura, de quelcom que possiblement és cert, de quelcom a verificar.

Dos tipus d'hipòtesis estadístiques:

- **Hipòtesi nul·la (H_0):** la que ens plantegem a priori a verificar i que «acceptem» o rebutgem després de l'anàlisi estadística de les dades. Ens indica que la diferència és estadísticament nul·la o no significativa, prou petita per a poder ser explicada per l'atzar.
- **Hipòtesi alternativa (H_1):** és la complementària (no només la contrària) de la H_0 . Ens indica que la diferència és estadísticament diferent de zero o significativa i que és suficientment gran o diferent de zero per a no poder ser explicada només per l'atzar.

Tal com s'ha dit abans, la teoria de la decisió estadística també es basa en els conceptes de distribució mostral i error típic:



Dos tipus de decisions:

- Rebutjar la H_0 : rebutgem la H_0 amb un risc α i acceptem la H_1 (diferència significativa).
- No rebutjar («acceptar») la H_0 : res s'oposa a acceptar la H_0 (diferència no significativa).

4.3.1. Errors i riscos

Dos tipus d'errors (o riscos d'error):

- α : de primera espècie o error típic I
- β : de segona espècie o error típic II

Decisió	Realitat (desconeguda)	
	H_0 verdadera	H_0 falsa
Rebutjar la H_0	α : error típic I o de 1a espècie	Decisió correcta (potència d'un contrast $1 - \beta$)
No rebutjar la H_0 (acceptar)	Decisió correcta	β : Error típic II o de 2a espècie

α : risc d'error conegut (fixat a priori) i petit, habitualment 0,05 (5%) o 0,01 (1%).

β : risc d'error desconegut perquè depèn de la diferència entre el paràmetre real i el paràmetre teòric i, generalment, desconeixem el paràmetre real.

4.3.2. Potència d'un contrast (funció de potència) ($1 - \beta$)

Capacitat que té un contrast per detectar que la H_0 és falsa i rebutjar-la. Com augmentar la potència ($1 - \beta$) d'un contrast? (disminuir β):

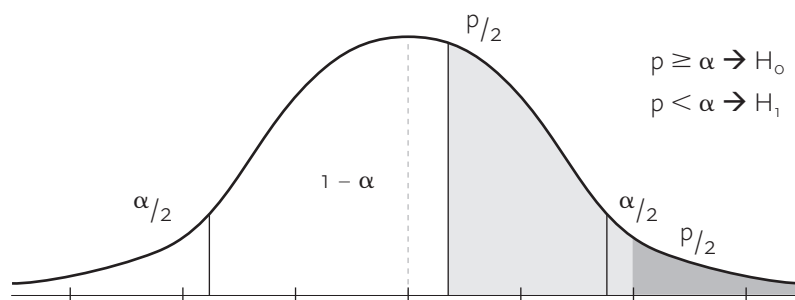
- Augmentar α (mala solució).
- Augmentar n : la distribució mostral té menys desviació i l'interval de probabilitat es redueix i també β (sense augmentar α).

En augmentar n , els intervals són més estrets i, per tant, el risc β (intersecció) disminueix.

4.3.3. Grau de significació (p)

És la probabilitat d'observar un valor igual o major a l'estadístic (quan la H_0 és verdadera) o la probabilitat associada a l'estadístic de contrast.

En els *outputs* de programes informàtics figura la dada p , i serveix per a la presa de decisió estadística. S'interpreta de la manera següent:

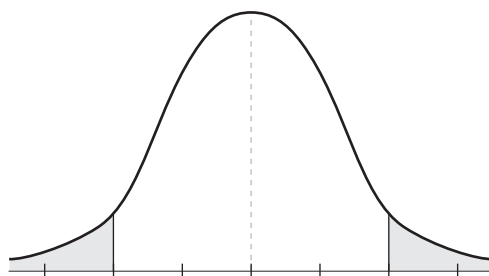


Si la X és a la zona crítica, el valor p ($p/2$) és menor que α ($\alpha/2$); per tant, la decisió és rebutjar la hipòtesi nul·la ($H_0 \rightarrow H_1$). La diferència és estadísticament significativa.

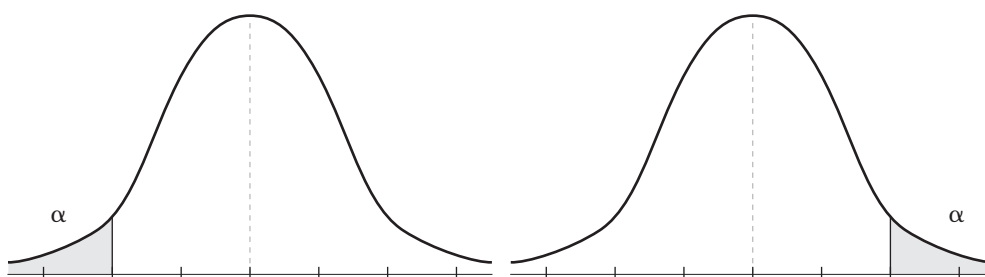
Si la X és a la zona no crítica, el valor p és més gran que α , per tant, la decisió és que res no s'oposa a acceptar la hipòtesi nul·la (H_0). La diferència no és estadísticament significativa.

4.3.4. Contrasts unilaterals i bilaterals

Un **contrast bilateral** (també anomenat de dues cues, no direccional o bidireccional) és aquell contrast en què l' α està situat a ambdós extrems de la distribució.



Un **contrast unilateral** (també anomenat d'una cua, direccional o unidireccional) és aquell contrast en què l' α està situat en un dels extrems (inferior o esquerre; superior o dret) de la distribució.



Avui en dia els contrastos unilaterals estan força qüestionats perquè:

- Són menys exigents que els bilaterals.
- Amb els bilaterals també podem saber si és inferior o superior.

En principi i sempre que no tinguem un criteri explícit (si és superior o inferior, ha millorat o empitjorat, etc.), els contrastos, els plantejarem com a bilaterals.

4.4. Exercicis resolts

1. Després dels resultats generals, el 75% de l'alumnat necessita reforç en unes determinades competències. Quants alumnes necessitaran reforç en una escola de 200 alumnes? ($\alpha = 0,05$)

SOLUCIÓ:

$$p_0 \in p \pm Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}} = 0,75 \pm 1,96 \cdot \sqrt{\frac{0,75 \cdot 0,25}{200}} = 0,75 \pm 0,06 \quad \begin{array}{l} \rightarrow 0,81 \cdot 200 = 162 \\ \rightarrow 0,69 \cdot 200 = 138 \end{array}$$

En aquesta escola de 200 alumnes necessitaran reforç entre 138 i 162 alumnes, amb un nivell de confiança del 95%.

2. Es vol analitzar la persistència acadèmica de l'alumnat d'origen estranger a Catalunya. La mitjana de rendiment acadèmic és de 5,5 amb una desviació típica de 1,80. Una mostra de 500 estudiants és suficient per tenir un error màxim de 0,08? ($\alpha = 0,05$)

SOLUCIÓ:

$$n = \left(\frac{Z_{\alpha/2} \cdot \sigma}{e} \right)^2 = \left(\frac{1,96 \cdot 1,80}{0,08} \right)^2 = 1.944,81 \approx 1945$$

No, 500 estudiants no són suficients. Per tenir un error màxim de 0,08, la mostra hauria de contenir com a mínim 1.945 estudiants.

3. En un estudi en què han participat 128 docents d'educació primària, s'ha obtingut una mitjana de 9 punts (i una $s = 0,98$) en una escala d'actitud davant de la tecnologia educativa. Quina es preveu que sigui la mitjana dels docents en aquesta actitud, amb un risc d'error de l'1%?

SOLUCIÓ:

$$\mu \in \bar{X} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 9 \pm 2,58 \cdot \frac{0,98}{\sqrt{128}} = 9 \pm 0,087 \quad \begin{array}{l} \rightarrow 9,087 \\ \rightarrow 8,913 \end{array}$$

L'actitud dels docents d'educació primària vers la tecnologia educativa es troba entre els 9,087 i els 8,913 punts de l'escala, amb un nivell de confiança del 99%.

4. En un estudi sobre violència 2.0, s'identifica que, en una mostra de 895 joves, el 39% ha patit *ciberbullying* en alguna ocasió. Quins percentatges de joves han patit *ciberbullying* entre els joves en general? ($\alpha = 0,05$)

SOLUCIÓ:

$$p \in p_0 \pm Z_{\alpha/2} \cdot \sqrt{\frac{p \cdot q}{n}} = 0,39 \pm 1,96 \cdot \sqrt{\frac{0,39 \cdot 0,61}{895}} = 0,39 \pm 0,0163 \quad \begin{array}{l} \rightarrow 0,4063 \\ \rightarrow 0,3737 \end{array}$$

Entre el 37% i el 41% dels joves han patit en alguna ocasió *ciberbullying* en el 95% dels casos.

5. La mitjana d'edat dels joves en iniciar el consum d'alcohol és de 13 anys ($s = 1,28$). Amb quines edats iniciaran el consum un grup de quaranta adolescents que participaran d'un programa preventiu? ($\alpha = 0,01$)

SOLUCIÓ:

$$\bar{X} \in \mu \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 13 \pm 2,58 \cdot \frac{1,28}{\sqrt{40}} = 13 \pm 0,2022 \quad \begin{array}{l} \rightarrow 13,2022 \\ \rightarrow 12,7978 \end{array}$$

En el grup de quaranta adolescents que participaran del programa, es preveu que iniciïn el consum d'alcohol entre els 12,8 i els 13,2 anys, amb un risc d'error de l'1%.

6. Si volem fer un estudi sobre els percentatges d'estudiants universitaris que finalitzen els seus estudis en el termini previst segons les diferents titulacions, quants estudiants haurem d'incloure en la mostra per mantenir un error màxim de 0,02 amb un nivell de significació del 99%?

SOLUCIÓ:

$$n = p \cdot q \left(\frac{Z_{\alpha/2}}{e} \right)^2 = 0,5 \cdot 0,5 \left(\frac{2,58}{0,02} \right)^2 = 4.160,25 \approx 4.161$$

Per mantenir un error màxim de 0,02 haurem d'incloure en la mostra un mínim de 4.161 estudiants universitaris.

CORRELACIÓ I REGRESSIÓ

En aquest capítol, s'exposen els conceptes bàsics sobre la correlació i la regressió estadística a fi d'entendre les anàlisis que es poden fer per estudiar la relació entre diversos fenòmens socioeducatius i la seva possible predicció. D'una banda, s'explica la correlació i la regressió lineal simple, per als casos en què es compleixen els supòsits paramètrics, i d'altra banda, s'esmenten altres tipus de correlació en funció de la naturalesa de les variables estudiades. Per facilitar la comprensió es presenten casos il·lustratius. Finalment, s'expliquen la correlació parcial i la regressió lineal múltiple, amb la definició de conceptes i la il·lustració de diversos casos.

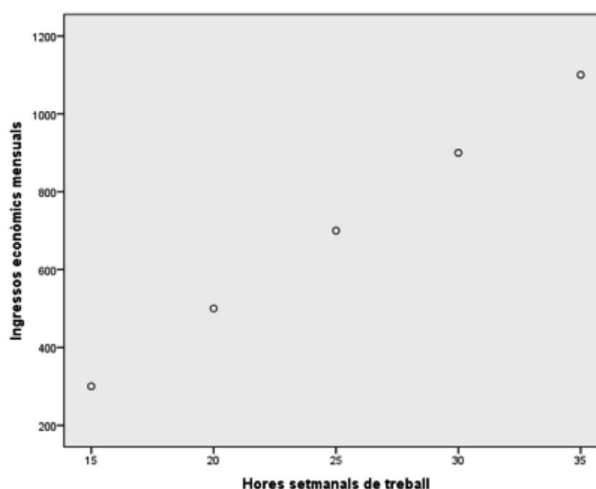
5.1. Conceptes bàsics

La **correlació** és la relació entre dues o més variables, és a dir, la variació conjunta (covariació) entre elles. Aquesta relació pot ser positiva, negativa o nul·la en funció de com varien les dades conjuntament:

- **Relació positiva:** també anomenada relació de dependència positiva, succeeix quan l'augment dels valors d'una de les variables coincideix amb l'augment dels valors de l'altra (o de les altres variables) i a l'inrevés (els valors inferiors d'una variable coincideixen amb els valors inferiors de l'altra variable).

Per exemple, hores setmanals de treball (X) i ingressos econòmics mensuals (Y). Quan augmenten les hores setmanals de treball augmenten també els ingressos econòmics mensuals.¹

X	Y
15	300
20	500
25	700
30	900
35	1100



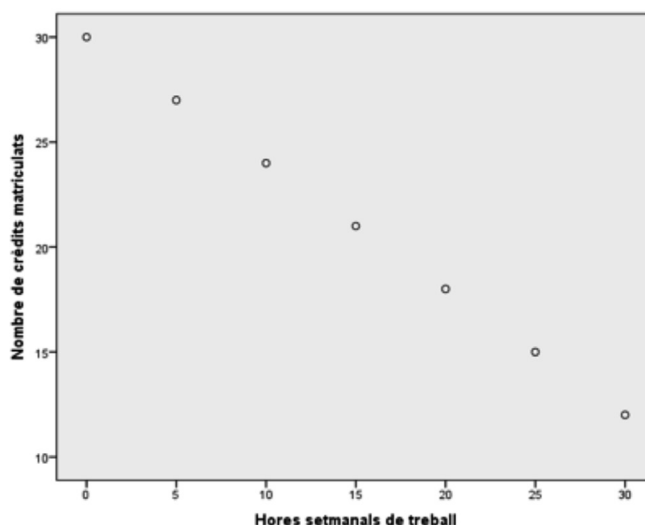
Gràfic de dispersió

¹ Opcions del menú de l'SPSS per obtenir el gràfic de dispersió: Gràfics > Cuadros de diálogo antiguos > Dispersión/Puntos... > Dispersión simple > Definir.

- **Negativa:** també anomenada relació de dependència negativa, succeeix quan augmenten els valors d'una de les variables i disminueixen els de l'altra (o altres variables) i a l'inrevés.

Per exemple, hores setmanals de treball (X) i nombre de crèdits matriculats (Y). Quan augmenten les hores setmanals de treball disminueix el nombre de crèdits matriculats.

X	Y
0	30
5	27
10	24
15	21
20	18
25	15
30	12



Gràfic de dispersió

- **Nul·la:** També anomenada relació d'independència, succeeix quan les variables no guarden cap relació entre elles, varien independentment l'una de l'altra.

El **coeficient de correlació** indica el grau de relació entre les variables i el sentit d'aquesta relació (positiva o negativa):

- Si el coeficient de correlació és 0 (correlació nul·la), indica que les variables són independents les unes de les altres.
- Si el coeficient és 1, significa que la correlació és perfecta positiva.
- Si el coeficient és -1, significa que la correlació és perfecta negativa.

El **coeficient de determinació** indica la proporció de variància que és comuna entre variables correlacionades. Equival al coeficient de correlació elevat al quadrat (r^2).

La **regressió** indica la forma com es produeix la variació conjunta i això permet predir els valors d'una variable en funció dels de l'altra. Només té sentit quan existeix correlació o dependència entre les variables.

5.2. Correlació i regressió lineal simple

Existeixen diferents tipus de proves estadístiques que permeten confirmar o rebutjar relacions entre variables. L'elecció d'una prova o d'una altra depèn de la natura i escala de mesura de les variables amb les quals treballem. Quan l'estudi es fa entre dues variables, es parla de coeficients simples, i si es vol estudiar la correlació de més de dues variables, el terme adequat és el múltiple.

Així, doncs, si ens preguntem ... aplicarem

Existeix relació entre la nota mitjana de batxillerat i la nota mitjana de primer any de carrera?

Treballem amb dues variables: nota mitjana de batxillerat i nota mitjana del primer any de carrera; totes dues variables són quantitatives i mesurades amb l'escala d'interval. El coeficient de correlació que s'aplicarà dependrà de si les variables segueixen o no la distribució de probabilitat de la Llei Normal com un dels criteris d'aplicació bàsic.

Si les dues variables segueixen la Llei Normal, s'utilitzarà el **coeficient de correlació de Pearson simple** i se simbolitzarà amb una «r» minúscula amb el subíndex «xy» representant les dues variables estudiades.

Si una o cap de les variables segueix la Llei Normal, la prova més adient és el **coeficient de correlació de Spearman**. Aquest coeficient també s'aplica si les variables estan mesurades amb una escala ordinal (per exemple: no amb nota numèrica, sinó amb qualificació —aprovat, notable...). Se simbolitza amb una «r» minúscula amb el subíndex «s» de Spearman o amb una lletra grega « ρ_{xy} ».

Existeix algun tipus d'associació entre el fet de treballar mentre estudies i aprovar o suspendre el curs?

Les dues variables amb les quals treballem: situació laboral (treballar o no) i superació del curs (aprovar o suspendre) són variables de natura qualitativa i categòriques.

El coeficient que s'utilitza habitualment independentment del tipus d'escala (ordinal-nominal) i de dicotomització (artificial *vs* natural) és el **Khi quadrat** (χ^2).

A la taula següent es presenta la diversitat de proves de correlació en funció del tipus de variable. Així mateix, es detalla el recorregut que s'ha de fer en l'aplicació de la prova quan s'utilitza l'SPSS.

Coeficient de correlació	Variable X	Variable Y	SPSS
Pearson (r_{xy})	Quantitativa contínua (o discreta)	Quantitativa contínua (o discreta)	Analitzar > Correlacions > Bivariades... > Pearson
Spearman (r_s o ρ_{xy})	Mesurada amb una escala ordinal	Mesurada amb una escala ordinal	Analitzar > Correlacions > Bivariades... > Spearman
Khi quadrat (χ^2)	Qualitativa	Qualitativa	Analitzar > Estadístics descriptius > Taules de contingència... > Estadístics... > Khi quadrat
Biserial (r_b)	Quantitativa contínua (o discreta)	Dicotomitzada (qualitativa artificial)	--
Biserial puntual (r_{bp})	Quantitativa contínua (o discreta)	Qualitativa dicotòmica	--
Tetracòrica (r_t)	Dicotomitzada (qualitativa artificial)	Dicotomitzada (qualitativa artificial)	--

5.2.1. Correlació de Pearson (r_{xy})

Abans d'aplicar la prova de correlació de Pearson s'ha de confirmar que les variables i les dades segueixen una sèrie de condicions d'aplicació:

- a) Les variables han de ser quantitatives (contínues o discretes) i s'han hagut de mesurar segons l'escala d'interval o raó (a l'SPSS anomenada «escala»).
- b) La mostra de subjectes ha de ser gran (superior o igual a 30).
- c) La distribució de les seves dades ha de seguir la Llei Normal. Es confirmarà aquesta condició mitjançant l'aplicació de la prova de Kolmogorov-Smirnov (K-S).
- d) S'ha de confirmar la homogeneïtat de les variàncies de les variables, anomenada *homocedasticitat*.

Els passos per aplicar aquesta prova són, en primer lloc, confirmar la verificació de les condicions d'aplicació i, en segon lloc, fer una primera aproximació del tipus de relació existent entre les dues variables. Aquesta aproximació es fa mitjançant un gràfic de dispersió simple. En funció de la dispersió de les dades resultant es calcularà el valor del coeficient de correlació de Pearson i, posteriorment, s'interpretarà.

La interpretació del coeficient de correlació considera dues hipòtesis estadístiques:

- **H₀ (hipòtesi nul·la):** existeix una relació d'independència o correlació no significativa entre els valors de les dues variables. La variació conjunta és suficientment petita com per ser explicada per la influència de l'atzar.
- **H₁ (hipòtesi alternativa):** la relació entre ambdues variables és significativa i de dependència, és a dir, existeix una variació conjunta suficientment gran com per no ser explicada només per la influència de l'atzar.

El coeficient de correlació de Pearson és l'únic coeficient que, a més de possibilitar l'acceptació o rebuig de la H₀, també informa sobre la intensitat de la relació existent entre les variables (vegeu l'apartat «Conceptes bàsics»).

En funció de la interpretació del coeficient de correlació resultant es poden continuar fent estudis de regressió o predicció. Per exemple, si es confirma que les notes de primer any de carrera tenen una relació de dependència amb les notes mitjanes de batxillerat podríem establir una predicció en situacions futures: un alumne amb una nota X al batxillerat obtindrà una nota Y en el primer any de carrera.

Per tant, **els estudis de regressió només tindran sentit si prèviament s'ha confirmat l'existència de relació significativa.**

La regressió que més s'aplica en l'àmbit educatiu és la lineal. Per tant, per fer un estudi predictiu es partirà del càlcul de l'equació d'una recta en què hi ha un punt de partida **a** i una pendent **b**, una variable dependent **X** i una variable independent **Y**.

- 1) Càlcul de la recta de regressió: predicció del valor més probable:

$$\bullet \quad X' = \bar{X} + b_{xy} \cdot (Y - \bar{Y}) \quad \text{o bé} \quad X' = a_{xy} + b_{xy} \cdot Y$$

$$\bullet \quad Y' = \bar{Y} + b_{yx} \cdot (X - \bar{X}) \quad \text{o bé} \quad Y' = a_{yx} + b_{yx} \cdot X$$

La predicció pot ser puntual (en funció del valor Y, s'obtindrà un valor en X) o establir un interval probable considerant un determinat marge d'error.

2) Càlcul de l'interval de regressió (amb un determinat marge d'error):

- $X \in X' \pm Z_{\alpha/2} \cdot S_{xy}$ on $S_{xy} = S_x \sqrt{1 - r_{xy}^2}$
- $Y \in Y' \pm Z_{\alpha/2} \cdot S_{yx}$ on $S_{yx} = S_y \sqrt{1 - r_{xy}^2}$

CAS PRÀCTIC DE CORRELACIÓ LINEAL SIMPLE

En un determinat centre educatiu, s'ha observat una certa relació entre els resultats de dues proves de llengua. Es vol confirmar si estadísticament la relació existent és significativa i si es pot fer un estudi de regressió. Disposem dels resultats obtinguts en un **test verbal** i un **examen de llengua** d'un grup d'alumnes de batxillerat escollits a l'atzar.

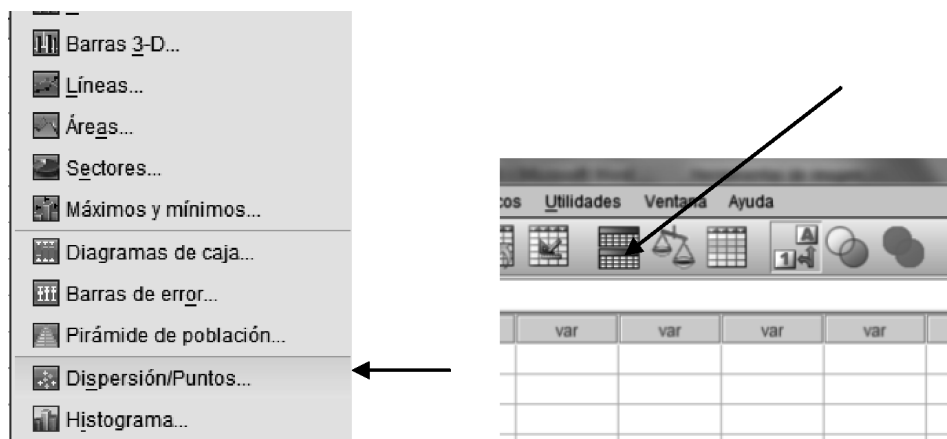
1) Introduir a la matriu les dades de les variables.

1: Test_verbal		21
	Test_verbal	Examen_Llengua
1	21	13
2	23	12
3	21	12
4	17	10
5	19	8
6	19	6
7	14	5
8	17	3
9	16	3
10	11	2

NOTA EXPLICATIVA:
Les dades d'aquest cas pràctic només han de servir per a exemplificar el procediment del càlcul de la prova i la seva comprensió.

2) Verificar els supòsits d'aplicació de la prova.

3) Fer el gràfic de dispersió seleccionant les dues variables de l'estudi.

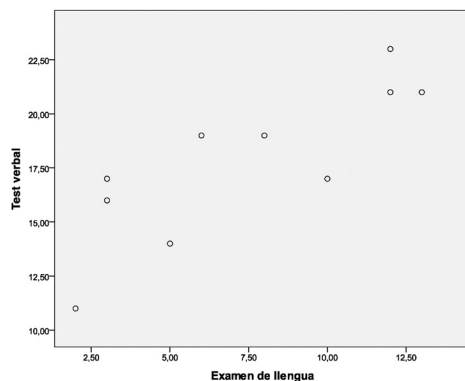
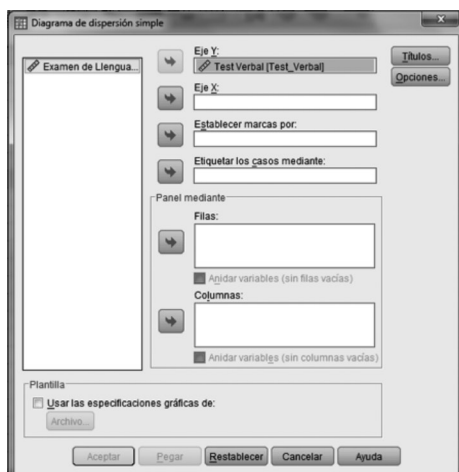


Se seleccionen les dues variables i se situen en cadascun dels eixos (X i Y). Tot seguit, s'accepta l'ordre. Aquest tipus de gràfic també permet escriure el títol dins de les seves possibilitats d'edició.

A continuació, es crea un gràfic de dispersió. Si es dibuixa una línia imaginària que envolti els punts d'intersecció, es podrà intuir descriptivament quin tipus de relació es pot establir entre les dues variables.

S'observa una tendència ascendent en la relació:

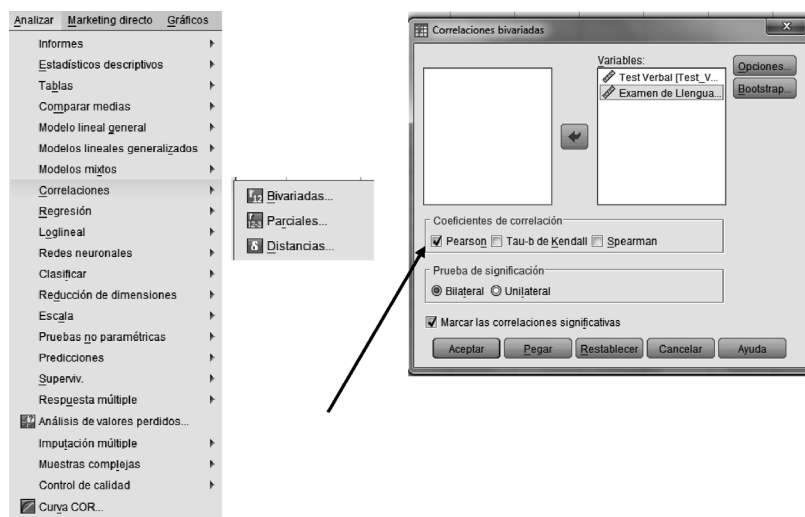
A més puntuació a l'examen de llengua més puntuació al test verbal.



- 4) Es calcula el valor del coeficient de correlació de Pearson.

En la finestra d'Analitzar, seleccionem l'opció de «Correlació» i optarem per la correlació bivariada. Apareix un menú en el qual hem de seleccionar les variables marcar l'opció corresponent al coeficient de correlació de Pearson.

Aquesta prova dóna l'opció de marcar d'entrada aquells coeficients que són significatius estadísticament amb dos marges d'error.



- 5) Interpretació del coeficient de correlació.

El programa calcula una matriu de correlació entre les variables seleccionades. La matriu de correlació és simètrica i els valors de les diagonals tenen el valor d'1, ja que és la relació de la variable amb si mateixa.

Es pot observar que el coeficient de correlació de Pearson entre ambdues variables és de $r_{xy} = ,824$ amb un grau de significació del ,003. Com que la significació (,003) és menor que el marge d'error 0,01, rebutgem la H_0 amb un marge d'error de 0,01 i acceptem la H_1 . Es pot concloure que la correlació (,824) és significativa i que la relació és de dependència positiva entre els resultats de les dues proves; les variacions no es poden explicar només per la influència de l'atzar.

CORRELACIONS

		Text verbal (x)	Examen de llengua (y)
Text verbal (x)	Correlació de Pearson	1	,824*
	Sig. (bilateral)	,	,003
	N	10	10
Examen de llengua	Correlació de Pearson	,824*	1
	Sig. (bilateral)	,003	,
	N	10	10

* La correlació és significativa a nivell 0,01 (bilateral).

6) Càlcul de la recta de regressió.

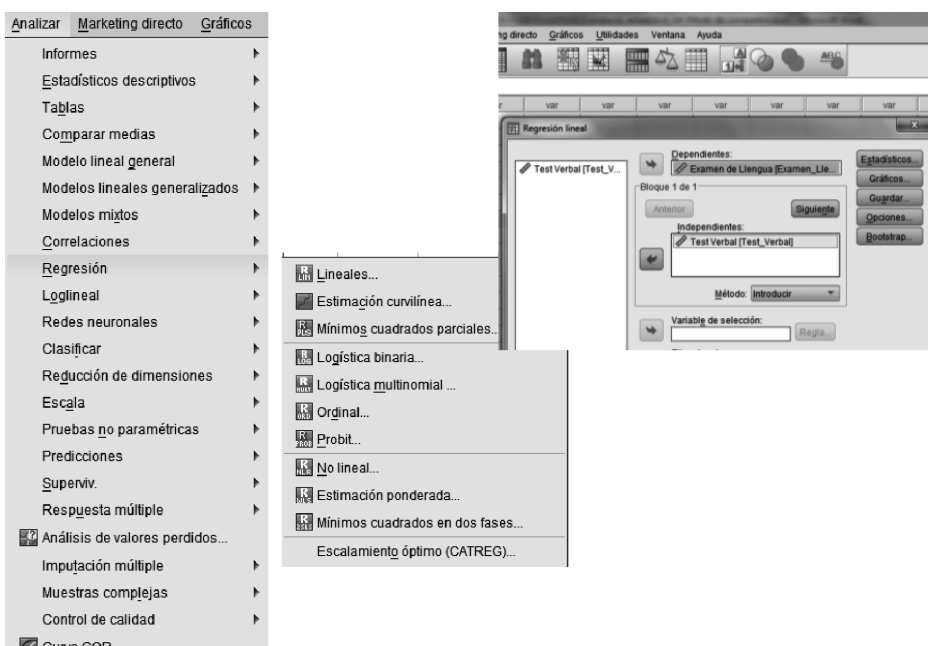
En haver confirmat que existeix una relació significativa, es pot procedir a fer l'estudi de regressió. El programa SPSS ens calcula totes les dades necessàries per al càlcul de la recta de regressió lineal.

En la correlació lineal simple, teòricament, hi ha dues rectes de regressió:

- Recta de regressió d'Y en funció d'X ($R_{Y/X}$). Expressa la variació de la variable Y en funció de la variable X. Permet predir el valor més probable (y') de la variable Y en funció d'un determinat valor d'X.
- Recta de regressió d'X en funció d'Y ($R_{X/Y}$). Expressa la variació de la variable X en funció de la variable Y. Permet predir el valor més probable (x') de la variable X en funció d'un determinat valor d'Y.

A la pràctica, utilitzem solament la primera (Y en funció d'X) perquè a la variable dependent l'anomenem X i a la variable independent l'anomenem Y.

Suposem que per al nostre estudi ens interessa poder fer una estimació de la nota que es podrà treure en l'examen de llengua sabent la puntuació que té una persona en el test verbal; per tant, necessitem calcular la recta de regressió d'Y en funció d'X.



En l'*output* apareix tot un seguit d'informació. Primer, el càlcul del coeficient de correlació múltiple (R), que coincideix amb el valor de coeficient de correlació bivariada, ja que només són dues variables (,824). En el segon requadre, apareix la significació del coeficient a través de l'aplicació de la prova ANOVA. I, finalment, apareixen les dades corresponents amb la recta de regressió, la constant **ayx**, la pendent **byx** de la variable independent.

RESUM DEL MODEL

Model	R	R quadrat	R quadrat corregida	Error típic de l'estimació
1	,824 ^a	,679	,639	2,50501

a. Variables predictores: (constant), Test_Verbal Test Verbal.

ANOVA^a

Model		Suma de quadrats	gl	Mitjana quadràtica	F	Significació
1	Regressió	106,199	1	106,199	16,924	,003 ^b
	Residual	50,201	8	6,275		
	Total	1.156,400	9			

a. Variable dependent: Examen_Llengua Examen de Llengua.

b. Variables predictor: (constant), Test_Verbal Test Verbal.

COEFICIENTS^a

Model	Coeficients no estandarditzats		Coeficients tipificats		t	Significació
	B	Error típic	Beta			
1 (Constant)	-9,661	4,222			-2,288	,051
Test_Verbal Test Verbal	,958	,233	,824		4,114	,003

a. Variables dependents: Examen_Llengua Examen de Llengua.

Els coeficients no estandarditzats ens donen els valors a_{yx} (29,661) i b_{yx} (0,958) de la recta de regressió que necessitem. Per tant, tenim que la recta de regressió és:

$$Y' = -9,66 + 0,958 \cdot X$$

Així, si una persona té una puntuació de 20 punts en el test verbal, el valor més probable en l'examen de llengua és $Y' = -9,66 + 0,958 \cdot 20 = 9,51$.

7) Càlcul de l'interval de regressió (amb un determinat marge d'error).

La predicció és més encertada en el nostre context si la fem considerant un interval probable. Per calcular-lo, considerarem el coeficient de determinació del coeficient de correlació (rx^2) i la dispersió de la variable dependent S_y .

$$S_{yx} = S_y \sqrt{1 - r_{xy}^2} = 3,9547 \sqrt{1 - 0,824^2} = 2,24 \text{ (vegeu la nota 2)}^2$$

$$Y \in Y' \pm Z_{\alpha/2} \cdot S_{yx} = 9,51 \pm 1,96 \cdot 2,24 = 9,51 \pm 4,39 \rightarrow 13,90 \\ \rightarrow 5,12$$

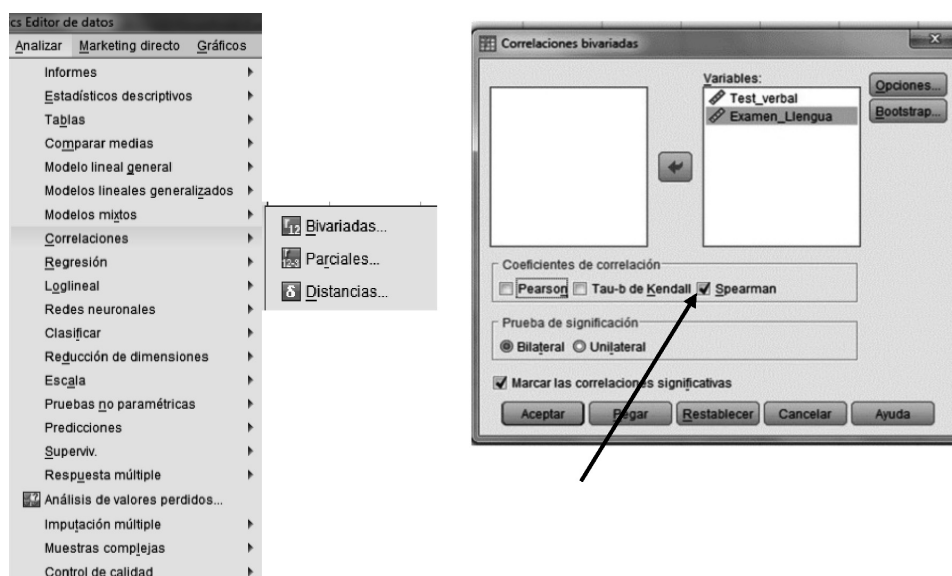
Aquest resultat ens indica que, per a un individu que ha obtingut 20 punts a X (test verbal), el més probable (Y') és que obtingui 9,51 a Y, però en el 95% dels casos (amb un nivell de confiança del 95% o un risc d'error del 5%) la seva puntuació a Y oscil·larà entre un màxim de 13,90 i un mínim de 5,12.

De la mateixa manera, es podria estudiar la regressió d'X en funció d'Y, utilitzant l'equació de la recta $X = f(y)$ per a calcular el valor més probable d'X (X') per a un valor determinat d'Y, i després calcular l'interval de regressió al voltant d'X'.

5.2.2. Altres coeficients de correlació simple

COEFICIENT DE CORRELACIÓ DE SPEARMAN

Si volem estudiar la relació entre variables mesurades amb l'escala de mesura ordinal o bé les variables quantitatives contínues que no segueixen la Llei Normal, s'utilitzarà la prova de correlació de Spearman. La interpretació d'aquest coeficient és igual que en el cas del coeficient de Pearson i el seu càlcul amb el programa SPSS és similar.



Apareixerà una matriu de «Correlació no paramètrica». El coeficient resultant només s'interpreta en funció de si existeix una relació significativa o no.

A partir de les dades del cas pràctic de l'apartat anterior, es demana el càlcul del coeficient de Spearman. La matriu de correlacions presenta el valor del coeficient, el grau de significació i el nombre de subjectes. En el nostre cas, el coeficient és de ,868 i la seva inter-

² El valor d' S_{yx} apareix en l'*output* de l'SPSS (Resumen del modelo) com a «Error típ. de la estimación» (en aquest cas, dóna 2,505), tot i que en calcular-ho manualment dóna 2,24, ja que s'utilitza la quasidesviació típica.

pretació es farà a partir de contrastar el grau de significació amb el marge d'error escollit; així, doncs, com que la significació és de ,001 i és superior al nostre marge d'error del ,01, rebutgem la H_0 .

CORRELACIONS

			Test_verbal	Examen_Llengua
Rho de Spearman	Test_verbal	Coefficient de correlació	1,000	,868*
		Sig. (bilateral)		,001
		N	10	10
	Examen_Llengua	Coefficient de correlació	,868*	1
		Sig. (bilateral)	,001	
		N	10	10

* La correlació és significativa a nivell 0,01 (bilateral).

KHI QUADRAT χ^2

Es pot aplicar amb qualsevol tipus de variables (qualitatives). No té cap mena de restriccions, no cal una n mínima ni que les distribucions s'ajustin a la corba normal. Les variables han de ser categòriques (ordinals o nominals).

A diferència dels coeficients de correlació de Pearson i de Spearman, la de Khi quadrat indica si hi ha o no correlació, però no el sentit ni en la intensitat de la correlació.

No obstant això, a partir del valor de **Khi quadrat** es poden calcular dos coeficients (Phi i el coeficient de contingència) que quantifiquen la intensitat de la relació. El valor resultant s'interpreta igual que Pearson.

Coefficients que indiquen la intensitat de la correlació a partir de Khi quadrat:

$$2 \times 2 \rightarrow \Phi = \sqrt{\frac{\chi^2}{n}} \quad k \times l \rightarrow C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

El càlcul de Khi quadrat té una condició d'aplicació que consisteix en el fet que, com a màxim, un 20% de les freqüències teòriques (o esperades) pot ser inferior a 5. Si aquesta condició no es compleix, aleshores s'ha de reagrupar alguna categoria i tornar a fer el càlcul de la Khi quadrat des del principi.³

CAS PRÀCTIC DE KHI QUADRAT

Un grup d'estudiants universitaris ha facilitat voluntàriament les dades sobre la seva situació laboral i sobre si ha superat o no l'assignatura de Sociologia. Les variables són: **situació laboral** amb les categories de treballa i no treballa i la **qualificació** de l'assignatura quant a si l'ha superat o aprovat.

L'objectiu és confirmar fins a quin punt les dues variables tenen algun tipus de relació/associació. Així, doncs,

³ Si la taula de contingència és de 2×2 i hi ha alguna freqüència esperada inferior a 5, aleshores, com que ja no es pot reagrupar cap categoria, en comptes de consultar la «Chi-cuadrado de Pearson» que apareix a l'*output*, s'ha de consultar la «Corrección por continuidad».

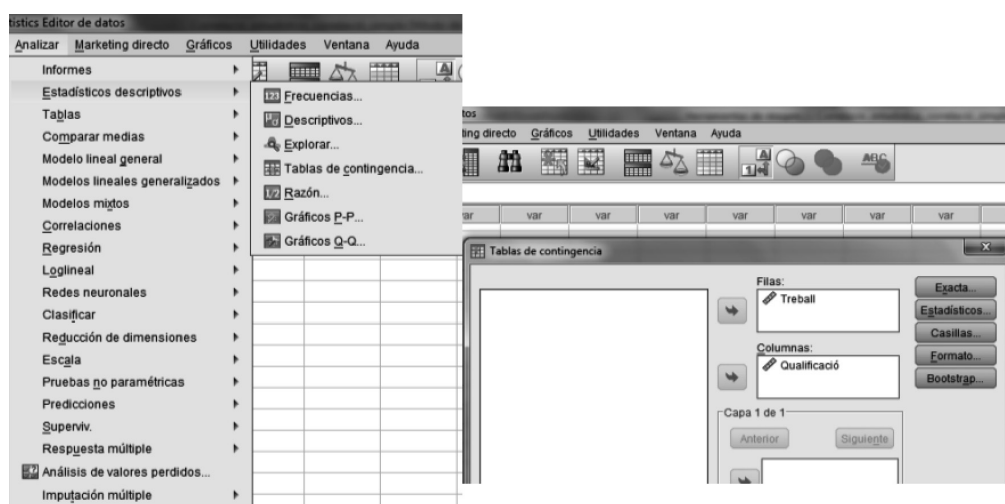
1) Distribució de les dades en una taula de doble entrada

A partir de la distribució dels subjectes de la mostra en les diferents caselles que apareixen a la taula de contingència i en funció dels valors en ambdues variables categòriques, es podrà estimar si existeix alguna tendència de relació o no.

El cas A exemplifica una distribució en la qual s'observa una major concentració de subjectes en les caselles d'Homes treballadors *versus* Dones no treballadores. En canvi, en el cas B no s'observa cap tendència; és indistint el fet de ser home o dona respecte a la situació laboral.

Cas A	Treballa	No treballa	Cas B	Treballa	No treballa
Home	20	2	Home	11	11
Dona	5	15	Dona	10	10

Amb el programa SPSS, es demanarà dins de l'opció d'«Estadísticos descriptivos» «Tablas de contingencia». En el nostre cas, serà una taula de 2×2 (dues categories per a cadascuna de les variables).



A la taula apareixen com estan distribuïts els 32 estudiants. A nivell descriptiu, s'observa una tendència: els alumnes que treballen majoritàriament suspensen, enfront dels que no treballen, que solen aprovar.

El programa permet, a partir de l'opció de «Casillas», calcular els percentatges de les freqüències absolutes en funció del total de la fila, de la columna o del total. També permet mostrar la distribució gràficament.

TAULA DE CONTINGÈNCIA TREBALL * QUALIFICACIÓ

		Qualificació		
		1 Aprovat	2 Suspès	Total
Treball	1 Treballa	4	14	18
	2 No treballa	13	1	14
Total		17	15	32

2) Càlcul del valor de Khi quadrat

En primer lloc, es comprova el percentatge de freqüències esperades (o teòriques) que hi ha amb valors inferiors a 5; en aquest exemple, n'hi ha un 0%. Per tant, es compleix la condició d'aplicació.

En el quadre de diàleg de «Tablas de Contingencia» hi ha l'opció de demanar «Estadísticos», el valor de Khi quadrat, el valor dels coeficients de contingència i el de phi, entre d'altres...

El valor de Khi quadrat és 15,778 i la seva significació de ,000; com que aquesta significació és menor al marge d'error 0,01, podem concloure que el fet de treballar o no mentre s'estudia té una relació significativa amb la qualificació de l'assignatura.

PROVES DE KHI QUADRAT

Prova	Valor	gl	Sig. asintòtica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Khi quadrat de Pearson	15,778 ^a	1	,000		
Correcció per continuïtat ^b	13,069	1	,000		
Raó de versemblances	17,962	1	,000		
Estadístics exactes de Fischer				,000	,000
Associació lineal per lineal	15,285	1	,000		
N de casos vàlids	32				

a. 0 caselles (0%) tenen una freqüència esperada inferior a 5. La freqüència mínima esperada és 6,56.

b. Calculat només per a una taula de 2×2 .

Si volem saber com és d'intensa aquesta relació, dins de l'opció d'«Estadísticos» haurem de marcar el coeficient Phi en ser una taula de 2×2 .

En el nostre cas, el valor resultant dona -,702 significativa.

MIDES SIMÈTRIQUES

		Valor	Sig. aproximada
Nominal per nominal	Phi	-,702	,000
	V de Cramer	,702	,000
	Coeficient de contingència	,575	,000
N de casos vàlids		32	

5.3. Correlació parcial i regressió lineal múltiple

En la recerca social i educativa, donada la complexitat dels fenòmens que s'estudien, és molt probable que es considerin més de dues variables al mateix temps per tal d'intentar explicar la realitat.

Suposem, per exemple, que en una mostra de quinze estudiants hem mesurat les variables intel·ligència general (G), alçada (A) i edat (E). Hem calculat els estadístics de les tres variables i els coeficients de correlació lineal simple entre elles. Aquestes dades estan recollides a les taules següents:

	G	A	E
1	85	144	8
2	70	130	9
3	80	147	9
4	95	145	12
5	93	143	12
6	97	145	12
7	90	149	12
8	94	154	12
9	95	150	13
10	100	151	13
11	120	152	13
12	145	156	15
13	102	157	15
14	107	175	15
15	133	160	17

ESTADÍSTICS DESCRIPTIUS

	N	Mitjana	Desviació típica
Intel·ligència general (G)	15	100,40	19,518
Alçada (A)	15	150,53	9,899
Edat (E)	15	12,47	2,475

CORRELACIONS

	G	A	E
Correlació de Pearson	1	,610*	,812**
G Significació (bilateral)		,016	,000
N	15	15	15
Correlació de Pearson	,610*	1	,747**
A Significació (bilateral)	,016		,001
N	15	15	15
Correlació de Pearson	,812**	,747**	1
E Significació (bilateral)	,000	,001	
N	15	15	15

* La correlació és significativa a nivell 0,05 (bilateral).

** La correlació és significativa a nivell 0,01 (bilateral).

Les correlacions simples trobades en realitat no són exactament així perquè en qualsevol cas cadascuna de les variables que forma una parella està relacionada també amb l'altra. Les tres estan correlacionades.

La **correlació parcial** permet estudiar la relació entre dues variables, eliminant la influència de les restants: en general, en un conjunt de m variables, la relació entre dues d'elles, eliminant la influència de les restants. També es podria aconseguir el mateix amb un disseny en què es controlés el possible efecte de les variables «restants», és a dir, les que no es plantegen ni com a independents ni com a dependents.

La **correlació múltiple (R)** permet estudiar la relació entre una variable dependent i un conjunt de variables independents. La regressió múltiple estableix l'equació que permet predir la variable independent en funció d'una o més variables independents.

Y_0 variable dependent
 X_1 variable independent 1
 X_2 variable independent 2

Regressió múltiple:

$$Y' = b_0 + b_1X_1 + b_2X_2$$

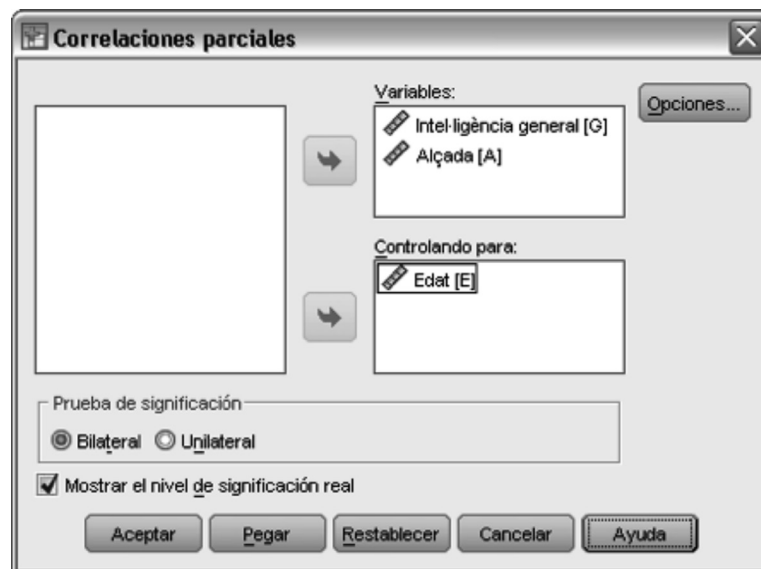
interval:

$$Y \in Y' \pm Z_{\alpha/2} S_e$$

CAS PRÀCTIC DE CORRELACIÓ PARCIAL

Es vol confirmar la relació que hi ha entre la intel·ligència general (G), l'alçada (A) i l'edat (E) d'un grup d'estudiants.

- 1) S'introdueixen les dades a una matriu de l'SPSS. Les dades corresponen a les que s'han exposat en l'exemple anterior.
- 2) Es confirmen els supòsits que han de complir les dades per al càlcul de la correlació parcial: a) les variables han de ser quantitatives i amb una distribució simètrica, i b) cada parell de variables ha d'ajustar-se a la normalitat bivariada.
- 3) Es calculen les correlacions parcials:
 - a) Entre les variables G i A, eliminant l'efecte de la variable E (menú principal de l'SPSS, «Anali-
zar, Correlaciones, parciales»).



- b) Entre les variables A i E, eliminant G.
- c) Entre les variables G i E, eliminant A.

- 4) S'interpreten els resultats:

CORRELACIONS

Variables de control			Intel·ligència general	Alçada
Edat	Intel·ligència general	Correlació	1,000	,009
		Significació (bilateral)	,	,976
		gl	0	12
	Alçada	Correlació	,009	1,000
		Significació (bilateral)	,976	,
		gl	12	0

CORRELACIONS

Variables de control		Alçada	Edat
Intel·ligència general	Alçada	Correlació	1,000
		Significació (bilateral)	,044
		gl	12
	Edat	Correlació	,544
		Significació (bilateral)	,044
		gl	12

CORRELACIONS

Variables de control		Edat	Intel·ligència general
Alçada	Edat	Correlació	1,000
		Significació (bilateral)	,008
		gl	12
	Intel·ligència general	Correlació	,676
		Significació (bilateral)	,008
		gl	12

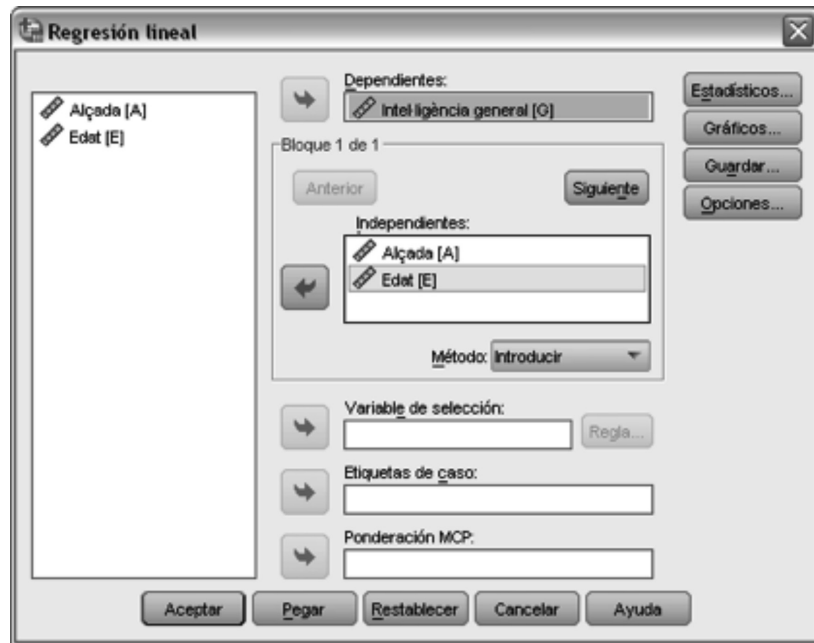
Es pot observar com, en eliminar l'efecte de l'edat (que és la que de veritat correlaciona amb la intel·ligència en períodes evolutius), la correlació entre l'alçada i la intel·ligència general queda reduïda a nivells no significatius, mentre que les altres dues (intel·ligència-edat i edat-alçada) es mantenen.

Limitacions: només es pot utilitzar quan les variables de les quals s'elimina la influència siguin molt precises i objectives (alçada, pes, edat, pressió sanguínia..., h) en general variables biomètriques o fisiològiques). En educació, com en ciències humanes i socials, la majoria de les variables no compleixen aquesta condició, per la qual cosa la correlació parcial té poca aplicació.

CAS PRÀCTIC DE CORRELACIÓ I REGRESSIÓ LINEAL MÚLTIPLE

Es vol confirmar la relació que hi ha entre la intel·ligència general (G), l'alçada (A) i l'edat (E) d'un grup d'estudiants i calcular l'equació de regressió per predir la intel·ligència general (com a variable dependent) en funció de l'alçada i l'edat (com a variables independents).

- 1) S'introdueixen les dades a una matriu de l'SPSS. Les dades corresponen a les que s'han exposat en l'exemple anterior.
- 2) Es confirmen els supòsits que han de complir les dades per al càlcul de la correlació parcial: a) les variables han de ser quantitatives, i b) per a cada valor de la variable independent, la distribució de la variable dependent ha de ser normal. La variància de distribució de la variable dependent ha de ser constant per a tots els valors de la variable independent. La relació entre la variable dependent i cada variable independent ha de ser lineal i totes les observacions han de ser independents.
- 3) Es calcula la regressió lineal. Al menú principal de l'SPSS, seleccioneu les opcions Analitzar > Regresión > Lineales...



4) S'interpreten els resultats:

RESUM DEL MODEL

Model	R	R quadrat	R quadrat corregida	Error típic de l'estimació
1	,812 ^a	,660	,603	12,300

a. Variables predictores: (constant), edat, alçada.

ANOVA^a

Model	Suma de quadrats	gl	Mitjana quadràtica	F	Significació
1 Regressió	3,518,265	2	1,759,132	11,628	,002 ^b
Residual	1,815,335	12	151,278		
Total	5,333,600	14			

a. Variable dependent: intel·ligència general.

b. Variables predictores: (constant), edat, alçada.

COEFICIENTS^a

Model	Coeficients no estandarditzats		Coeficients tipificats		t	Significació
	B	Error típic	Beta			
1 (Constant)	18,823	59,064			,319	,755
Alçada	,015	,500	,008		,030	,976
Edat	6,361	1,999	,806		3,182	,008

a. Variable dependent: intel·ligència general.

A la primera taula de resultats (resum del model), es pot observar que la correlació múltiple (R) és 0,812. A més a més, el 66% ($R^2 = 0,660$, denominat coeficient de determinació) de la variància de la variable dependent és explicat pel model de regressió trobat, o dit d'una altra manera, les dues variables predictores d'aquest cas expliquen el 66% de la variabilitat de la variable dependent. El valor de R^2 corregida té en consideració el nombre de subjectes i de variables independents de tal manera que, quan hi ha pocs casos i moltes variables independents, l' R^2 pot ser artificialment alta i, per tant, és més oportú tenir en compte l' R^2 corregida. Finalment, aquesta taula també informa de l'error típic de l'estimació ($S_e = 12,3$).

La segona taula (ANOVA), amb una significació 0,002 inferior a 0,05, indica que es pot rebutjar la hipòtesi nul·la que sosté que $R^2 = 0$, i, per tant, es pot concloure que el model de regressió trobat permet fer prediccions amb més seguretat que si la predicció es fes a l'atzar.

A la tercera taula (coeficients), s'observa que el coeficient de la variable Alçada no és significatiu ($0,976 > 0,05$) i, per tant, no contribueix gaire a la regressió múltiple. En canvi, el coeficient de l'Edat sí que és significatiu ($0,008 < 0,05$) i, per tant, sí que hi contribueix. Per determinar la importància dels predictors, podem observar la columna de coeficients tipificats (s'han de considerar en valors absoluts, sense tenir en compte el signe); en aquest cas, l'edat és molt més important (0,806) com a predictor que l'alçada (0,008).

A efectes il·lustratius, tot i que l'alçada es podria eliminar del càlcul de la regressió, considerarem les dues variables independents en l'equació de predicció:

$$\text{Intel·ligència general} = \text{constant} + (\text{alçada} \times \text{coeficient de l'alçada}) + (\text{edat} \times \text{coeficient de l'edat})$$

Suposem que predim la intel·ligència d'un subjecte de 155 cm d'alçada i 15 anys d'edat. Per tant,

$$\text{Intel·ligència general} = 18,823 + 155 \times 0,015 + 15 \times 6,361$$

$$\text{Intel·ligència general} = 18,823 + 2,325 + 95,415$$

$$\text{Intel·ligència general} = 116,563$$

L'interval de puntuacions d'intel·ligència general amb un marge de confiança del 95% és el següent:

$$116,563 \pm 12,3 \cdot Z_{\alpha/2}$$

$$116,563 \pm 12,3 \cdot 1,96$$

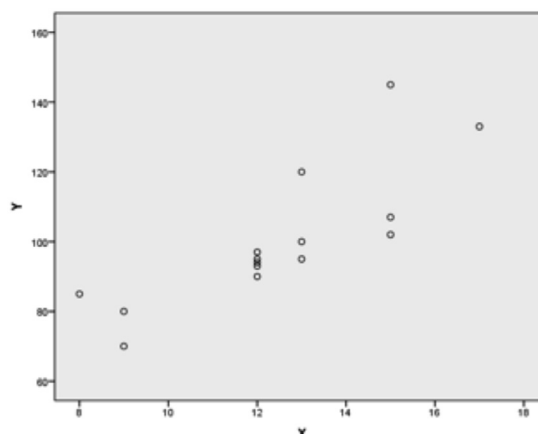
$$116,563 \pm 24,108 \rightarrow 140,67$$

$$\rightarrow 92,46$$

5.4. Exercicis resolts

1. Quin dels valors següents correspon a la correlació del gràfic de dispersió tenint en compte com es distribueix el núvol de punts?

- a) $-0,65$
- b) $+0,09$
- c) $+0,81$



SOLUCIÓ:

S'ha de considerar si el sentit del núvol que es dibuixa és ascendent i l'amplada del núvol. La resposta correcta és la c), el valor és positiu i mostra una intensitat alta.

2. Si volem estudiar si existeix relació entre dues variables quantitatives (sou mensual i edat) i només una segueix la distribució de la Llei Normal, quin coeficient utilitzaríem?

SOLUCIÓ:

El coeficient de correlació de Spearman.

3. A partir de les correlacions que apareixen en la matriu adjunta, identifica entre quins aspectes de l'avaluació d'una recerca (projecte, eines de recollida d'informació i anàlisi realitzada) existeix una relació.

CORRELACIONS

		Utilitat del projecte	Utilitat de l'anàlisi	Utilitat de les eines
Utilitat del projecte	Correlació de Pearson	1	,195	,197
	Significació (bilateral)		,070	,064
	N	103	87	89
Utilitat de l'anàlisi	Correlació de Pearson	,195	1	,464
	Significació (bilateral)	,070		,000
	N	87	101	91
Utilitat de les eines	Correlació de Pearson	,197	,464	1
	Significació (bilateral)	,064	,000	
	N	89	91	103

SOLUCIÓ:

Existeix relació significativa entre la utilitat de l'anàlisi i la de les eines de recollida d'informació amb un coeficient de 0,464 i un grau de significació del 0,000.

4. Es vol estudiar la relació entre el nivell instructiu-educatiu i els ingressos de 400 pares.

TAULA DE CONTINGÈNCIA NIVELL D'INGRESSOS * NIVELL INSTRUCTIU

			Nivell instructiu			
			Baix	Mitjà	Alt	Total
Nivell d'ingressos	Baix	Recompte	63	78	29	170
		Freqüència esperada	51,0	76,5	42,5	170,0
	Mitjà	Recompte	48	82	45	175
		Freqüència esperada	52,5	78,8	43,8	175,0
	Alt	Recompte	9	20	26	55
		Freqüència esperada	16,5	24,8	13,8	55,0
Total	Recompte		120	180	100	400
	Freqüència esperada		120,0	180,0	100,0	400,0

PROVES DE KHI QUADRAT

Prova	Valor	gl	Significació asintòtica (bilateral)
Khi quadrat de Pearson	22,931 ^a	4	,000
Raó de versemblances	21,744	4	,000
Associació lineal per lineal	19,436	1	,000
N de casos vàlids	400		

a. 0 caselles (,0%) tenen una freqüència esperada inferior a 5. La freqüència mínima esperada és 13,75.

MIDES SIMÈTRIQUES

		Valor	Significació aproximada
Nominal per nominal	Coefficient de contingència	,233	,000
N de casos vàlids		400	

a. Assumint la hipòtesi alternativa.

b. Utilitzant l'error típic asintòtic basat en la hipòtesi nul·la.

SOLUCIÓ:

Es comprova que no hi ha més d'un 20% de freqüències esperades inferiors a 5. El valor de Khi quadrat és 22,931 amb una significació de 0,000, la qual, en ser menor que 0,01, permet concloure que hi ha una relació significativa entre el nivell instructiu i el nivell econòmic dels subjectes estudiats. La intensitat d'aquesta relació és de 0,233.

5. A partir de les dades que es mostren a la matriu següent: grau de satisfacció amb la formació rebuda (variable y), hores de formació dedicades a activitats pràctiques (variable x1) i hores de formació teòrica rebudes (variable x2), calcula:

- Correlació múltiple considerant la variable y com a dependent, i x_1 i x_2 com a independents.
- Percentatge de la variància de la variable y explicada per les variables dependents.
- Significació del model de regressió.
- Importància de cada variable dependent en la regressió múltiple.
- Recta de regressió.
- Interval de puntuacions amb un marge de confiança del 95% que es poden predir pel que fa al grau de satisfacció en la formació rebuda per a un cas que tingui un valor de seixanta hores de formació dedicades a activitats pràctiques i cinquanta hores de formació teòrica.

	y	x1	x2
1	9,1	31	99
2	6,9	72	56
3	6,8	83	60
4	6,9	40	44
5	6,8	63	55
6	7,9	69	53
7	8,7	31	42
8	8,5	45	48
9	8,5	28	41
10	8,7	38	47

SOLUCIÓ:

- Tal com es mostra a la taula següent (Resum del model), el valor de la correlació múltiple (R) és 0,786. Es pot interpretar com una correlació moderada-alta, ja que s'aproxima a 1.

RESUM DEL MODEL

Model	R	R quadrat	R quadrat corregida	Error típic de l'estimació
1	,786 ^a	,618	,509	,6546

a. Variables predictores: (constant), hores teòriques, hores pràctiques.

- A la taula anterior es pot observar que el coeficient de determinació (R^2) és 0,618; per tant, el percentatge de variància de la variable dependent (grau de satisfacció amb la formació rebuda) explicat per les variables predictores és del 61,8%.
- A la taula següent (ANOVA) es mostra la significació (0,034) del model de regressió. Com que aquest valor és inferior a 0,05, es pot interpretar que és millor predir la variable dependent amb les variables predictores que merament a l'atzar.

ANOVA^a

	Model	Suma de quadrats	gl	Mitjana quadràtica	F	Significació
1	Regressió	4,857	2	2,428	5,668	,034 ^b
	Residual	2,999	7	,428		
	Total	7,856	9			

a. Variable dependent: grau de satisfacció.

b. Variables predictor: (constant), hores teòriques, hores pràctiques.

- La importància de cada variable predictor en el model de regressió s'observa a la taula següent (coeficients) en la columna de coeficients tipificats: en valors absoluts, té més importància la variable Hores pràctiques (0,762) que la variable Hores teòriques (0,223). A més a més, a la darrera columna

de la taula s'observa com la variable Hores pràctiques és significativa ($0,014 < 0,05$), mentre que la variable Hores teòriques no és significativa ($0,371 > 0,05$) per a la predicció de la variable dependent.

COEFICIENTS^a

	Model	Coeficients no estandarditzats		Coeficients tipificats	t	Significació
		B	Error típic	Beta		
1	(Constant)	8,991	,901		9,977	,000
	Hores pràctiques	-,036	,011	-,762	-3,261	,014
	Hores teòriques	,012	,012	,223	,956	,371

a. Variable dependent: grau de satisfacció.

e) La recta de regressió queda expressada de la manera següent:

$$\text{Grau de satisfacció} = 8,991 - 0,036 \times h \text{ pràctiques} + 0,012 \times h \text{ teòriques}$$

f) Per calcular l'interval de puntuacions, cal aplicar la recta de regressió amb els valors donats per a cadascuna de les variables predictores i posteriorment sumar i restar l'error típic d'estimació (extret de la primera taula Resum del model) amb el valor de Z que correspon al marge de confiança de 95%.

$$\text{Grau de satisfacció} = 8,991 - 0,036 \cdot 60 + 0,012 \cdot 50 \pm Z_{\alpha/2} \cdot S_e$$

$$\text{Grau de satisfacció} = 8,991 - 2,16 + 0,60 \pm 1,96 \cdot 0,6546$$

$$\text{Grau de satisfacció} = 7,431 \pm 1,2830$$

Per tant, el grau de satisfacció amb la formació rebuda per a una persona que ha dedicat seixanta hores de formació a activitats pràctiques i ha rebut cinquanta hores de formació teòrica estarà entre 6,148 i 8,714 (amb un marge de confiança del 95%).

PROVES PARAMÈTRIQUES I NO PARAMÈTRIQUES

Una de les aplicacions de l'estadística és extreure inferències en poblacions a partir de l'estudi de mostres. Aquest procés rep el nom d'**estadística inferencial** i els seus estudis pretenen deduir (inferir) propietats o característiques d'una població a partir d'una mostra representativa.

Un dels aspectes que permet fer la inferència és determinar si existeix associació entre diferents variables o no. Per realitzar aquest procediment se sol partir d'hipòtesis, és a dir, d'unes suposicions la validesa de les quals cal confirmar o rebutjar. I per dur a terme aquesta comprovació apliquem unes proves estadístiques o tests que permeten contrastar la veracitat o falsedat de les hipòtesis enunciades des del punt de vista estadístic. Aquest tipus de proves es classifiquen en **proves paramètriques** i **proves no paramètriques**. Mentre que les primeres són proves molt robustes i subjectes a determinades condicions (variables dependents quantitatives contínues, distribució normal de les mostres, variàncies similars i grandària de les mostres major a trenta casos), les segones són proves de distribució lliure, l'aplicació de les quals està indicada davant l'absència d'alguna d'aquestes condicions.

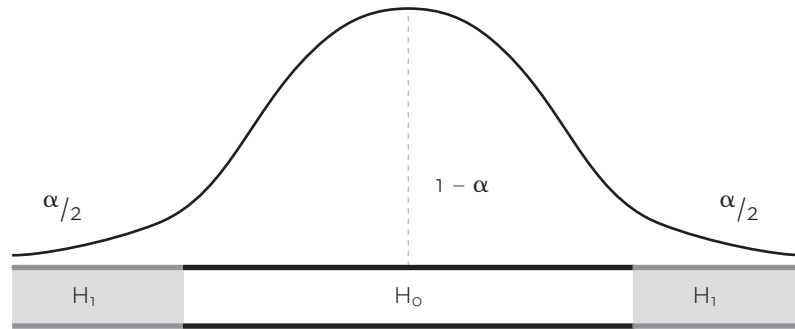
6.1. Condicions de parametricitat

Quan volem avaluar el grau d'associació o independència entre una variable quantitativa i una variable categòrica (cal recordar que aquesta classifica o diferencia els individus en grups, tants com categories té aquesta variable), el procediment estadístic inferencial recorre a comparar les mitjanes de les distribucions de la variable quantitativa en els diferents grups establerts per la variable categòrica. Si aquesta té només dues categories (és dicotòmica), la comparació de mitjanes entre dos grups independents es porta a terme pel **test t de Student**; si té tres o més categories, la comparació de mitjanes entre tres o més grups independents es realitza a través d'un model matemàtic més general, l'**anàlisi de la variància** (ANOVA). En ambdós casos, les proves estadístiques exigeixen certs requisits previs: la distribució normal de la variable quantitativa en els grups que es comparen i l'homogeneïtat de variàncies en les poblacions de les quals procedeixen els grups; el seu no compliment comporta la necessitat de recórrer a proves estadístiques no paramètriques.

Vegem amb més detall aquests requisits de parametricitat:

- **Variable numèrica:** la variable d'estudi (la dependent) ha d'estar mesurada en una escala que sigui, almenys, d'interval i, idealment, de raó.
- **Normalitat:** els valors de la variable dependent han de seguir una distribució normal, si més no, en la població a la qual pertany la mostra. La distribució normal o gaussiana (perquè la seva forma representa la campana de Gauss) és la distribució teòrica millor estudiada i deu la seva importància fonamentalment a la freqüència amb què diferents variables associades a fenòmens naturals i quotidians segueixen, aproximadament, aquesta distribució. Caràcters morfològics com pot ser la talla o el pes, o psicològics

com el quocient intel·lectual són exemples de variables de les quals s'assumeix sovint que segueixen una distribució normal.



No obstant això, i encara que alguns autors han assenyalat que el comportament de molts paràmetres en el camp de les ciències socials pot ser descrit mitjançant una distribució normal, pot resultar fins i tot poc freqüent trobar variables que s'ajustin a aquest tipus de comportament. A causa d'això cal comprovar si la distribució del nostre estudi segueix aquesta estructura teòrica. Per això, una simple exploració visual de les dades pot suggerir la forma. Hi ha, però, altres mesures com són els gràfics de normalitat i els contrastos d'hipòtesis (prova de Kolmogorov-Smirnov) que poden ajudar-nos a decidir, d'una manera més rigorosa, si la mostra de la qual es disposa procedeix o no d'una distribució normal. Quan les dades no siguin normals, es faran servir altres mètodes estadístics que no exigeixin aquest tipus de restriccions (els anomenats mètodes no paramètrics).

- **Homocedasticitat (homogeneïtat de variàncies) entre els grups a comparar:** les variàncies de la variable dependent en els grups que es comparen han de ser aproximadament iguals. Per això, un dels passos previs a la comprovació de l'existència de diferències entre les mitjanes de diverses mostres és determinar si les variàncies en tals mostres són iguals, és a dir, comprovar si es compleix la condició d'homogeneïtat de variàncies, ja que del compliment d'aquesta condició dependrà la formulació que emprem en el contrast de mitjanes. Hi ha diverses proves que permeten comprovar la igualtat de variàncies (F de Fisher, fmax de Hartley, prova de Bartlett, etc.). En aquest estudi, desenvoluparem la prova de Levene, és a dir, aquella que emprà SPSS per a comprovar que les variàncies de la variable dependent en els grups que es comparen siguin aproximadament iguals.
- **n mostral:** no ha de ser inferior a 30, i com més s'acosti a la N poblacional millor. Atès que les proves paramètriques realitzen estimació de paràmetres de la població a partir de mostres estadístiques, és lògic pensar que com més gran sigui la mostra més exacta serà l'estimació; en canvi, com més petita, més distorsionada serà la mitjana de les mostres pels valors rars extrems.

6.2. Selecció de la mostra

Les mostres estudiades poden ser independents i dependents (o relacionades).

- **Mostres independents.** Són aquelles l'univers de població de les quals resulta diferent. És el cas, per exemple, de la variable sexe, que es divideix en dos universos diferents: homes i dones, o el de la variable estat civil, definida per solters, casats, divorciats,

vidus, etc. D'aquesta manera, es poden enumerar múltiples variables que donen lloc a una infinitat d'universos mostrals, entre els quals és factible triar mostres independents.

- **Mostres dependents o relacionades.** Es refereixen a les provinents d'un univers mostral, a les quals s'aplicarà un pla experimental mitjançant el qual s'espera un canvi, de manera que en l'anàlisi de les observacions hi ha dos períodes: abans i després del tractament.

Per acabar, i com ja hem comentat en paràgrafs anteriors, quan per qualsevol raó no es pugui complir algun dels requisits de les proves paramètriques, l'investigador podrà recórrer a les anomenades proves no paramètriques. No s'hi exigeix homogeneïtat de variàncies i, auxiliades per un model matemàtic d'ajust, es pot obtenir una eficàcia molt propera a la de les proves t de Student i d'anàlisi de variància.

6.3. Classificació de proves paramètriques i no paramètriques

6.3.1. Proves paramètriques

A continuació, resumim el tipus de prova que s'ha d'utilitzar segons el contrast plantejat:

Tipus de contrastos	Proves
Una mostra	Prova t
Dues mostres independents	Prova t per a dades independents
Dues mostres relacionades	Prova t per a dades relacionades
Més de dues mostres independents	ANOVA

PROVA T PER A UNA MOSTRA

Contrasta si la mitjana d'una població difereix significativament d'un valor donat conegut o hipotetitzat. La prova calcula estadístics descriptius per a les variables de contrast juntament amb la prova t. Per defecte, l'SPSS inclou l'interval de confiança del 95% per a la diferència entre la mitjana de la variable de contrast i el valor hipotetitzat de la prova.

PROVA T PER A DUES MOSTRES INDEPENDENTS

Aquesta opció s'ha d'utilitzar quan la comparació es realitzi entre les mitjanes de dues poblacions independents (els individus d'una de les poblacions són diferents als individus de l'altra) com, per exemple, en el cas de la comparació de les poblacions d'homes i dones. Per tant, compara les mitjanes d'una variable per a dos grups de casos.

La matriu de dades ha d'estar configurada com és habitual, és a dir, hi ha una columna per a les dades de la variable d'interès i una segona columna amb els codis que defineixen les poblacions objecte de comparació. La prova calcula estadístics descriptius per a cada grup a més de la prova de Levene per a la igualtat de variàncies, així com els valors de t per a variàncies iguals i desiguals i l'interval de confiança del 95% per a la diferència de mitjanes.

Els grups de la *variable d'agrupació* es poden definir de dues formes: a) mitjançant valors específicats (s'escriu un valor per al grup 1 i un altre per al grup 2, i els casos amb altres

valors queden exclosos de l'anàlisi), o *b*) amb un punt de tall (s'estableix un nombre que divideixi els valors de la variable d'agrupació en dues parts. Els casos amb valors menors que el punt de tall formen un grup i els casos amb valors majors o iguals que el punt de tall formen l'altre grup).

PROVA T PER A DUES MOSTRES RELACIONADES

Hi ha una segona alternativa per contrastar dues mitjanes. Es refereix al suposat cas en què les dues poblacions no siguin independents, és a dir, el cas en què es tracti de poblacions relacionades.

La prova *t* per a dues mostres relacionades compara les mitjanes de dues variables d'un sol grup. Calcula les diferències entre els valors de les dues variables i contrasta si la mitjana difereix de zero. És a dir, aquest disseny s'aplica quan les dades estan aparellades (provenen de subjectes amb variables mesurades abans i després d'un tractament, o de parells de subjectes aparellats a partir de les seves característiques similars en variables de control o que volen controlar de manera específica). Per fer un contrast d'hipòtesis de mostres relacionades es requereixen, almenys, dues variables que representin valors per als dos membres del parell, per exemple, mesures pretest i posttest.

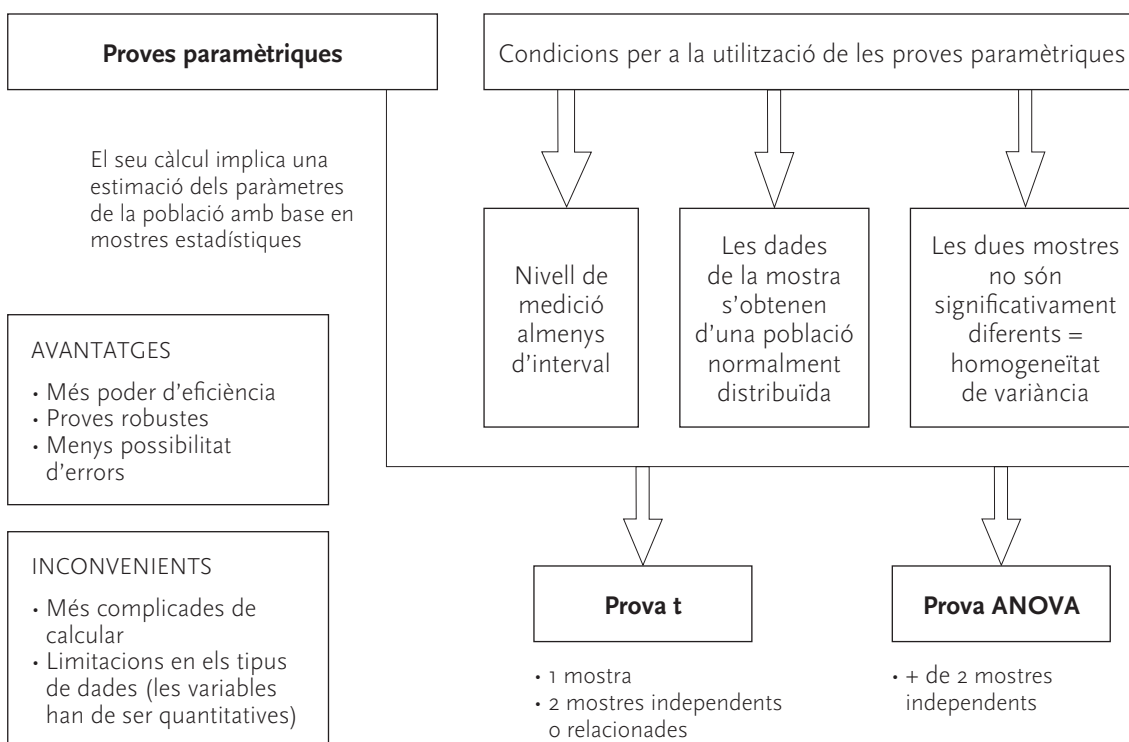
Només es poden utilitzar variables numèriques. Per a la realització d'aquesta anàlisi, les dues mostres han d'estar en dues variables diferents de la matriu de dades i han de formar la parella de mostres abans de poder afegir la comparació a la llista de variables relacionades. Per tant, compara les mitjanes de dues variables en un sol grup. El resultat inclou estadístics descriptius de les variables que es contrastaran, la correlació entre elles, estadístics descriptius de les diferències aparellades, la prova *t* i l'interval de confiança del 95%.

PROVA ANOVA PER A MÉS DE DUES MOSTRES INDEPENDENTS

ANOVA és l'acrònim d'anàlisi de la variància. És una prova estadística desenvolupada per realitzar simultàniament la comparació de les mitjanes de més de dues poblacions. A l'assumpció de normalitat ha d'afegir la de l'homogeneïtat de les variàncies de les poblacions a comparar. Aquesta condició prèvia d'aplicació es verificarà estadísticament mitjançant una de les opcions que es troben dins de la configuració de l'ANOVA.

S'haurà d'introduir la variable que es vol analitzar (variable dependent), així com la variable que defineix els grups objecte de comparació (factor). Si l'ANOVA resultés ser el rebuig de la hipòtesi nul·la d'igualtat de mitjanes, s'ha de prosseguir l'anàlisi amb la realització dels contrastos a posteriori (*post hoc*).

A manera de resum, presentem el mapa següent en el qual recollim els diferents **contrastos d'hipòtesis** vistos:



Font: Berlanga i Rubio (2012).

6.3.2. Proves no paramètriques

A continuació, oferim una classificació completa de les proves no paramètriques (disset en total). En la seva classificació s'han assumit dimensions relacionades amb el nombre de mostres i la relació o independència entre aquestes mostres tal com mostren les taules següents. No s'han inclòs les correlacions no paramètriques, incloses en el capítol anterior d'aquest text docent.

Variable dependent	Una mostra (bondat d'ajust)	Mostres relacionades		Mostres independents	
		2 mostres	> 2 mostres	2 mostres	> 2 mostres
Nominal	Binomial Khi quadrat Ratxes	McNemar	Cochran	-	-
Ordinal/ Interval	Kolmogorov-Smirnov	Signes Wilcoxon	Friedman Kendall	Ratxes de Wald-Wolfowitz U de Mann-Whitney Moses Kolmogorov-Smirnov	Mitjana Kruskal-Wallis Jonckheere-Terpstra

Font: Berlanga i Rubio (2012).

Es diu que les proves no paramètriques són alternatives a les paramètriques; per això, considerem apropiat establir l'equivalència entre ambdues tal com es mostra a la taula.

Mostra	Prova paramètrica	Prova no paramètrica
Mostres relacionades		
2 mostres	t-Student	Wilcoxon
> 2 mostres	ANOVA	Friedman
Mostres independents		
2 mostres	t-Student	U de Mann-Whitney
> 2 mostres	ANOVA	Kruskal-Wallis

A la taula següent es recullen les principals característiques de les disset proves no paramètriques trobades:

PROVES NO PARAMÈTRIQUES

Proves	Una mostra	Variables
Prova de Khi quadrat de Pearson	És una prova de bondat d'ajust que permet esbrinar si la distribució empírica d'una variable categòrica s'ajusta o no (s'assembla o no) a una determinada distribució teòrica (uniforme, binomial, multinomial, etc.).	VD: nominal
Prova binomial	És una prova de bondat d'ajust que permet esbrinar si una variable dicotòmica segueix o no un determinat model de probabilitat. Permet contrastar la hipòtesi que la proporció observada d'encerts s'ajusta a la proporció teòrica d'una distribució binomial (la qual cosa es tradueix en la possibilitat de contrastar hipòtesis sobre proporcions i sobre quantils).	VD: nominal
Prova de ratxes	- És una prova de bondat d'ajust que permet esbrinar si una variable dicotòmica segueix o no un determinat model de probabilitat. Permet contrastar la hipòtesi que la proporció observada d'encerts s'ajusta a la proporció teòrica d'una distribució binomial (la qual cosa es tradueix en la possibilitat de contrastar hipòtesis sobre proporcions i sobre quantils). - Una ratxa és una seqüència d'observacions d'un mateix atribut o qualitat. Una sèrie de dades en què hi ha moltes o poques ratxes permet concloure que aquestes no han ocorregut per atzar.	VD: nominal
Prova de Kolmogorov-Smirnov (K-S)	És una prova de bondat d'ajust que serveix per contrastar la hipòtesi nul·la que la distribució d'una variable s'ajusta a una determinada distribució teòrica de probabilitat que pot ser amb tendència a la normal, a la de Poisson o exponencial.	VD: ordinal/ interval
Proves	Dues mostres relacionades	Variables
Prova de McNemar	- Serveix per contrastar hipòtesis sobre igualtat de proporcions. - S'usa quan hi ha una situació en què les mesures de cada subjecte es repeteixen, de manera que la resposta de cada un s'obté dues vegades: una vegada abans i una altra després que tingui lloc un esdeveniment específic.	VI: dicotòmica VD: nominal
Prova dels signes	- Permet contrastar la hipòtesi d'igualtat entre dues mitjanes poblacionals. - Pot ser usada per saber si una variable tendeix a ser més gran que una altra. També és útil per provar la tendència que segueix una sèrie de variables ordinals positives, o per a una valoració ràpida d'un estudi exploratori.	VI: dicotòmica VD: ordinal/ interval

Prova de Wilcoxon	- Permet contrastar la hipòtesi d'igualtat entre dues mitjanes poblacionals. - Paral·lela a la prova paramètrica de contrast t per a mostres relacionades.	VI: dicotòmica VD: ordinal/ interval
-------------------	---	--

Proves	K-mostres relacionades	Variables
Prova de Friedman	- És una extensió de la prova de Wilcoxon per incloure dades registrades en més de dos períodes de temps o grups de tres o més subjectes rodolins, amb un subjecte de cada grup que ha estat assignat aleatòriament a una de les tres o més condicions. - La prova examina els rangs de les dades generades en cada període de temps per determinar si les variables comparteixen la mateixa distribució contínua del seu origen.	VI: politòmica VD: ordinal/ interval
Prova de Cochran	- Aquesta prova és idèntica a la prova de Friedman, però s'aplica quan totes les respostes són binàries. - La Q de Cochran prova la hipòtesi que diverses variables dicotòmiques que estan relacionades entre si tenen la mateixa mitjana. En observacions múltiples, les variables són mesures en el mateix individu o en individus rodolins. Té l'avantatge d'examinar canvis en les variables categòriques.	VI: dicotòmica VD: nominal
Coefficient de concordància de W de Kendall	Té les mateixes indicacions que la prova de Friedman, encara que el seu ús en investigació ha estat, principalment, per conèixer la concordança entre rangs més que per provar que existeix una diferència entre les mitjanes.	VI: politòmica VD: ordinal/ interval

Proves	Dues mostres independents	Variables
Prova U de Mann-Whitney	És equivalent a la prova de suma de rangs de Wilcoxon i la prova de dos grups de Kruskal-Wallis. És l'alternativa no paramètrica a la comparació de dues mitjanes independents a través de la t de Student.	VI: dicotòmica VD: ordinal/ interval
Prova de Kolmogorov-Smirnov	Serveix per contrastar la hipòtesi que dues mostres procedeixen de la mateixa població. Per a això, compara les funcions de distribució (funcions de probabilitat acumulades) de les dues mostres.	VI: dicotòmica VD: ordinal/ interval
Prova de ratxes de Wald-Wolfowitz	Contrasta si dues mostres amb dades independents procedeixen de poblacions amb la mateixa distribució.	VI: dicotòmica VD: ordinal/ interval
Prova de reaccions extremes de Moses	- Serveix per estudiar si existeix diferència en el grau de dispersió o variabilitat de dues distribucions. - Aquesta prova se centra en la distribució del grup de control i és una mesura per saber quants valors extrems del grup experimental influeixen en la distribució quan es combinen amb el grup de control.	VI: dicotòmica VD: ordinal/ interval

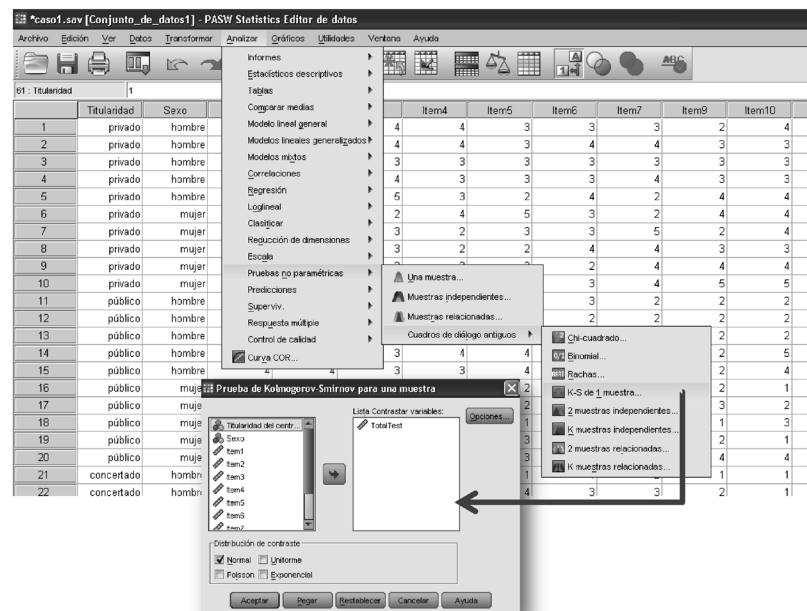
Proves	K-mostres independents	Variables
Prova de la Mediana	- Contrasta diferències entre dos o més grups en relació amb la seva mitjana, bé perquè no compleixen les condicions de normalitat per utilitzar la mitjana com a mesura de tendència central, bé perquè la variable és quantitativa discreta. - És similar a la prova Khi quadrat.	VI: politòmica VD: ordinal/ interval

Prova de Jonckheere-Terpstra	• És més potent que les seves homònimes, la prova de Kruskal-Wallis i la de la mitjana, quan hi ha una ordenació a priori (ascendent o descendent) de les K poblacions de les quals s'extreuen les mostres.	VI: politòmica VD: ordinal/ interval
Prova H de Kruskal-Wallis	• És una extensió de la d'U de Mann-Whitney i representa una excel·lent alternativa a l'ANOVA d'un factor completament aleatoritzat.	VI: politòmica VD: ordinal/ interval

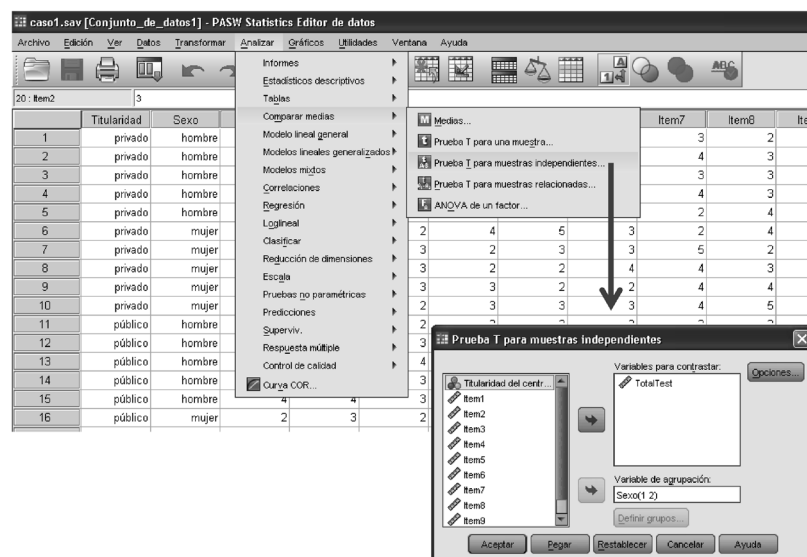
Font: Berlanga i Rubio (2012).

Vegem a continuació com obtenir aquestes proves (paramètriques i no paramètriques) amb el programa SPSS.

El primer pas sempre serà comprovar la normalitat de la distribució de dades de la variable quantitativa, com veiem a la figura següent:



Si segueix la llei normal ($p > 0,00$), el tipus de proves a emprar seran paramètriques, les quals es localitzen en SPSS a la finestra que mostra la figura següent:



En demanar alguna de les proves paramètriques, també hem de comprovar l'homoce-
dasticitat (o igualtat de variàncies), la qual s'estima en SPSS a través de la prova de Levene
que ha de llançar una significació major de 0,05 perquè es compleixi el requisit de homoce-
dasticitat (expressat a la taula següent com «s'han assumit variàncies iguals» a través de
l'estadístic F).

SPSS Statistics Visor

Archivo Edición Ver Datos Transformar Insertar Formato Analizar Gráficos Utilidades Ventana Ayuda

Resultado

- Frecuencias
 - Título
 - Notas
 - Estadísticos
- Prueba T
 - Título
 - Notas
 - Estadísticos de grupo
 - Prueba de muestras inde

Frecuencias

Estadísticos

TotalTest

N	Valídos	90
	Perdidos	0
Media		27,3667

Prueba T

Estadísticos de grupo

Sexo	N	Media	Desviación tip.	Error tip. de la media
TotalTest hombre	45	28,0000	5,56368	,82938
TotalTest mujer	45	26,7333	6,16220	,91861

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas				Prueba T para la igualdad de medias		95% Intervalo de confianza para la diferencia		
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error tip. de la diferencia	Inferior	Superior
TotalTest	Se han asumido varianzas iguales	,875	,352	1,023	88	,309	1,26667	1,23763	-1,19285	3,72619
	No se han asumido varianzas iguales			1,023	87,097	,309	1,26667	1,23763	-1,19321	3,72654

Si no segueix la llei normal ($p > 0,00$), haurem d'optar per les proves no paramètriques, tal com mostra la figura següent:



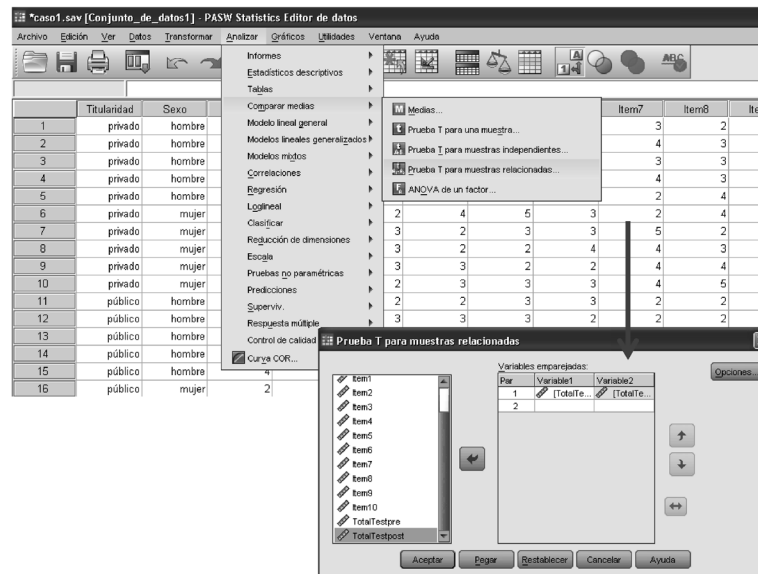
6.4. Exercicis resoltos

1. Suposem que volem comprovar si una intervenció educativa orientada a la millora de competències informacionals millora aquestes competències en l'alumnat d'un determinat centre de Batxillerat. Per a això, vam passar un test abans d'iniciar la formació per comprovar el nivell de l'alumnat (pretest) i, en finalitzar el programa, tornem a aplicar el mateix test al

mateix alumnat (posttest). Quin tipus de prova estadística haurem d'aplicar un cop comprovada la normalitat de les puntuacions obtingudes en el pretest i el posttest?

SOLUCIÓ:

Ja que es tracta d'un únic grup de subjectes mesurats en dos moments diferents, i es compleix la llei normal, s'ha d'aplicar la prova paramètrica t per a mostres relacionades.



2. Si l'*output* que apareix al programa SPSS després d'aplicar la prova corresponent és el següent, podem afirmar que el programa de formació competencial ha funcionat?

Par	Medio	N	Desviación típ.	Error típ. de la media
Par 1. TotalTestpre	27,367	50	6,57214	,91598
TotalTestpost	31,722	50	6,88014	,97863

Par	N	Correlación	Sig.
Par 1. TotalTestpre y TotalTestpost	50	,193	,082

Par	Medio	Desviación típ.	Error típ. de la media	Diferencia de medias		t	Sig.	Sig. bilateral
				inferior	superior			
Par 1. TotalTestpre - TotalTestpost	-4,35587	7,89720	,92286	-4,86187	-3,72140	-4,308	,000	,000

SOLUCIÓ:

Vam comprovar que en els resultats sí que hi ha diferències significatives entre abans i després de la formació, ja que la significació bilateral és menor que 0,05, de manera que vam concloure que el programa de formació rebut pels estudiants ha millorat les seves competències. S'ha passat d'una puntuació de 27,36 en el test inicial a una puntuació de 31,7 en el test final, i aquesta diferència és significativa.

3. Seguint amb el cas anterior, suposem la hipòtesi en relació amb les competències informacionals de l'alumnat de Batxillerat, que afirma que hi ha diferències entre els estudiants de centres educatius privats, públics i concertats, sent millor el nivell competencial en els alumnes dels centres privats. Per comprovar aquesta suposició, s'ha passat un test de competències informacionals a una mostra aleatòria d'estudiants de Batxillerat (trenta d'un institut privat, trenta d'un institut públic i trenta d'un de concertat).

Quins són els passos que hem de seguir amb el programa SPSS per comprovar la hipòtesi?

SOLUCIÓ:

Si es compleixen els supòsits de normalitat i homocedasticitat (disponible a la pestanya «Opcions»), llavors podem aplicar el test F de la prova paramètrica ANOVA, ja que es tracta de comprovar les diferències en les puntuacions de diverses mostres independents (les puntuacions dels estudiants de tres tipus diferents de centres en un test de competències).

S'ha d'introduir la variable objecte d'anàlisi (variable dependent; en el nostre cas, la puntuació total del test de competències) així com la variable que defineix els grups que es volen comparar (variable factor; en el nostre cas, la variable anomenada titularitat del centre).

Si l'ANOVA resultés ser el rebuig de la hipòtesi nul·la d'igualtat de mitjanes, s'ha de prosseguir l'anàlisi amb la realització dels contrastos a posteriori. A la pestanya anomenada «Post hoc» (contrastos o comparacions múltiples a posteriori) seleccionarem algun dels procediments que s'ofereixen. Els més habituals són el de Bonferroni i el de Scheffé. Aquests contrastos tenen sentit només si l'ANOVA surt significativa o propera a la significació estadística, ja que el que realitzen són comparacions de les mitjanes en les múltiples parelles de grups que puguin contrastar, per intentar esbrinar on és la diferència (o diferències) que ha causat que es rebutgi la hipòtesi nul·la en la primera part de l'ANOVA.



4. Si els *outputs* són els següents, podem dir que es compleix la nostra hipòtesi?

DESCRIPTIUS

	N	Mitjana	Desviació típica	Error típic	Interval de confiança per a la mitjana al 95%		Mínim	Màxim
					Límit inferior	Límit superior		
Privat	30	33,3000	2,13590	,38996	32,5024	34,0976	30,00	36,00
Públic	30	26,5333	5,61852	1,02580	24,4353	28,6313	19,00	35,00
Concertat	30	23,9333	3,10654	,56717	22,7733	25,0933	18,00	31,00
Total	90	27,9222	5,53889	,58385	26,7621	29,0823	18,00	36,00

TotalTestpre

ANOVA

	Suma de quadrats	gl	Mitjana quadràtica	F	Significació
Intergrups	1.402,822	2	701,411	45,964	,000
Intragrups	1.327,633	87	15,260		
Total	2.730,456	89			

TotalTestpre

PROVES POST HOC

	(I) Titularitat del centre educatiu	(J) Titularitat del centre educatiu	Diferència de mitjanes (I - J)	Error típic	Significació	Interval de confiança al 95%	
						Límit inferior	Límit superior
Scheffe	Privat	Públic	6,76667*	1,00863	,000	4,2547	9,2787
		Concertat	9,36667*	1,00863	,000	6,8547	11,8787
	Públic	Privat	-6,76667*	1,00863	,000	-9,2787	-4,2547
		Concertat	2,60000*	1,00863	,041	,0880	5,1120
	Concertat	Privat	-9,36667*	1,00863	,000	-11,8787	-6,8547
		Públic	-2,60000*	1,00863	,041	-5,1120	-,0880
Bonferroni	Privat	Públic	6,76667*	1,00863	,000	4,3044	9,2289
		Concertat	9,36667*	1,00863	,000	6,9044	11,8289
	Públic	Privat	-6,76667*	1,00863	,000	-9,2289	-4,3044
		Concertat	2,60000*	1,00863	,035	,1378	5,0622
	Concertat	Privat	-9,36667*	1,00863	,000	-11,8289	-6,9044
		Públic	-2,60000*	1,00863	,035	-5,0622	-,1378

Variable dependent: TotalTestpre

* La diferència de mitjanes és significativa al nivell 0,05.

SOLUCIÓ:

Els descriptius ens informen de les puntuacions obtingudes per l'alumnat de cada centre. Mitjana del privat: 33,3; mitjana del públic: 26,5; mitjana del concertat: 26,9.

A continuació apareix la sortida de l'ANOVA pròpiament dita, amb l'estadístic F de Snedecor, que en el nostre exemple equival a 45,96 i té un valor «p» associat de 0,000 (no significatiu). Amb això podríem concloure la nostra avaluació dient que «les variables Titularitat del Centre educatiu i Totaltestpre sí que mostren associació», o que «s'accepta la hipòtesi alternativa que les mitjanes del Totaltestpre són diferents en els diferents centres educatius».

En el quadre de comparacions múltiples veiem que cada centre educatiu es compara amb els altres dos, amb la qual cosa s'obté en cada contrast la diferència de mitjanes, l'interval de confiança al 95%, l'error estàndard i el valor «p» associat, que en tots els casos és significatiu en ser menor de 0,05.

En ser diferents les mitjanes de cada centre educatiu, conclouríem que les competències informacionals estan més desenvolupades en primer lloc al centre privat, seguit del públic i, finalment, el concertat.

Per acabar, vam concloure que res s'oposa a acceptar la hipòtesi que hi ha relació o associació entre les competències i el centre de procedència (prova ANOVA).

5. En un estudi orientat a conèixer l'actitud del professorat universitari cap a les TIC en la docència, mesurada a través d'una escala tipus Likert (quantitativa), volem comprovar si hi

ha relació entre el sexe i els anys d'experiència docent (portar més o menys cinc anys com a docent).

H_0 : el sexe és independent dels anys d'experiència.

H_1 : el sexe i els anys d'experiència estan relacionats.

Quina prova hem d'aplicar?

SOLUCIÓ:

Com que tenim dues variables de tipus categòric: sexe i experiència docent, la prova adequada és Khi quadrat, tal com es reflecteix en la figura següent.



6. Si l'output amb SPSS és el següent, què podem concloure?

TAULA DE CONTINGÈNCIA SEXE $\times$ ANYS D'EXPERIÈNCIA

Recompte		Anys d'experiència		Total
		Menys de 5	5 o més anys	
Sexe	Home	18	22	40
	Dona	14	26	40
Total		32	48	80

PROVES DE KHI QUADRAT

Prova	Valor	gl	Significació asintòtica (bilateral)	Significació exacta (bilateral)	Significació exacta (unilateral)
Khi quadrat de Pearson	,833 ^a	1	,361		
Correcció per continuïtat ^b	,469	1	,494		
Raó de versemblances	,835	1	,361		
Estadístics exacte de Fischer				,494	,247
Associació lineal per lineal	,823	1	,364		
N de casos vàlids	80				

a. o caselles (o%) tenen una freqüència esperada inferior a 5. La freqüència mínima esperada és 16,00.

b. Calculat només per a una taula de 2×2 .

SOLUCIÓ:

No es rebutja la H_0 , el que significa que les petites diferències trobades en la taula de contingència són degudes a l'atzar i, per tant, no hi ha relació entre el sexe i els anys d'experiència docent ($\text{sig.o}, 361 > 0,05$).

7. Seguint amb l'exemple anterior, ara ens proposem conèixer la influència de l'experiència docent (< 5 anys o > 5 anys) en la utilitat que el professorat atribueix a les TIC en l'ensenyament.

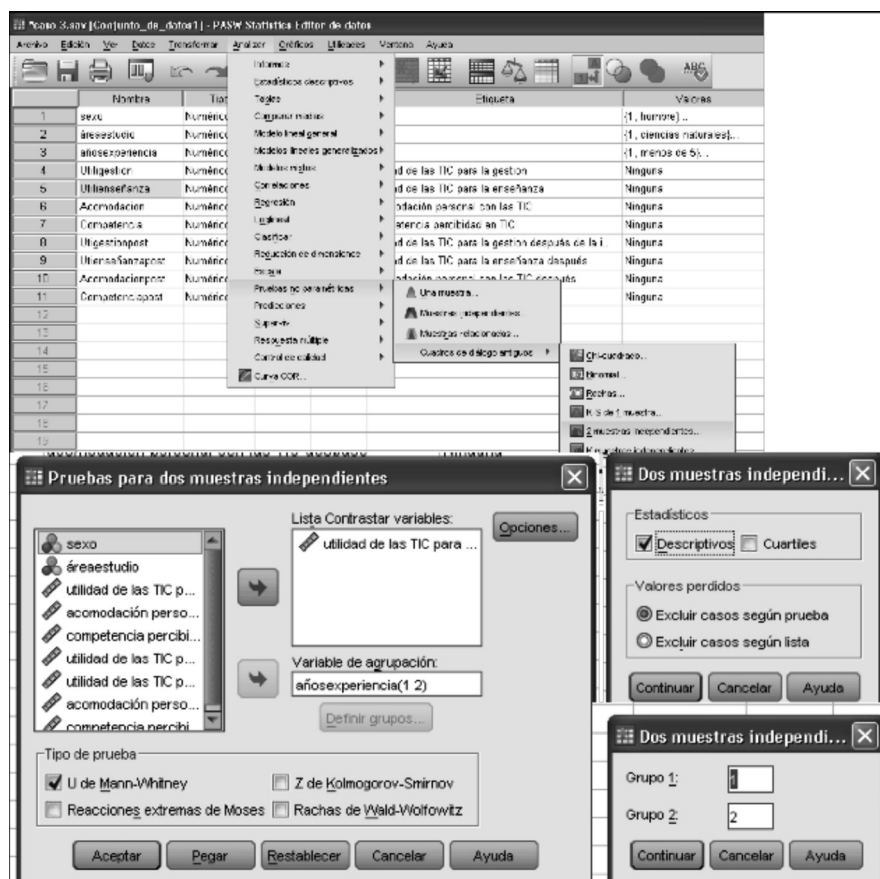
H_0 : l'experiència docent no influeix en la utilitat que el professorat atribueix a les TIC en l'ensenyament.

H_1 : l'experiència docent influeix en la utilitat que el professorat atribueix a les TIC en l'ensenyament.

Quina és la prova que hem d'aplicar, un cop comprovada la NO normalitat de la variable quantitativa: utilitat percebuda pel professorat? ($p < 0,00$)

SOLUCIÓ:

En tractar-se de dues mostres diferents: professorat amb més de cinc anys d'experiència i professorat amb menys de cinc anys d'experiència, i havent de triar una prova no paramètrica (per la NO normalitat de la variable dependent), l'adequada és l'U de Mann-Whitney, que en l'SPSS trobem com a prova per a dues mostres independents. Hi seleccionem la variable dependent «utilitat de les TIC» i la variable independent o factor «experiència docent», amb les seves dues categories: 1 (més de cinc anys) i 2 (menys de cinc anys).



8. Si l'*output* de l'SPSS ofereix la informació següent, podem concloure que hi ha relació entre l'experiència docent i la percepció d'utilitat de les TIC per part del professorat?

PROVA DE MANN-WHITNEY

Rangs				
	Anys d'experiència	N	Rang promig	Suma de rangs
Utilitat de les TIC per a l'ensenyament	Menys de 5 anys	32	49,03	1.569,00
	5 o més anys	48	34,81	1.671,00
	Total	80		

ESTADÍSTICS DE CONTRAST^a

Utilitat de les TIC per a l'ensenyament	
U de Mann-Whitney	495,000
W de Wilcoxon	1.671,00
Z	-2,809
Significació asintòtica (bilateral)	,005

a. Variable d'agrupació: anys d'experiència

SOLUCIÓ:

S'accepta la H_1 , que significa que l'experiència docent influeix en la utilitat que el professorat atribueix a les TIC en l'ensenyament (sig. 0,005 < 0,05). El professorat que té menys de cinc anys d'experiència les considera més útils (el seu rang mitjana és superior).

9. En el mateix estudi del professorat universitari i les TIC, l'objectiu ara és conèixer si l'àrea d'estudi (quatre àrees: ciències naturals, ciències socials, lletres, tècniques) a la qual pertany el professorat influeix en la utilitat que atribueix a les TIC en l'ensenyament.

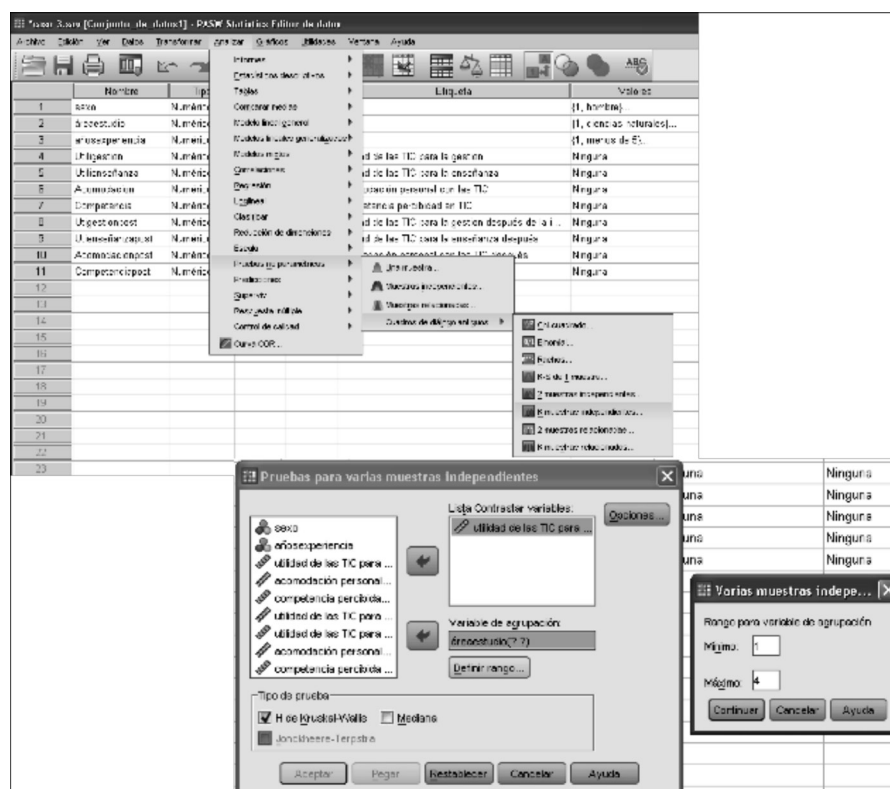
H_0 : l'àrea d'estudi a la qual pertany el professorat no influeix en la utilitat que atribueix a les TIC en l'ensenyament.

H_1 : l'àrea d'estudi a la qual pertany el professorat influeix en la utilitat que atribueix a les TIC en l'ensenyament.

Quina prova haurem d'aplicar per saber la resposta, un cop hem comprovat la NO normalitat de la variable dependent quantitativa «utilitat que el professorat atribueix a les TIC en l'ensenyament»?

SOLUCIÓ:

Ja que haurem d'optar per una prova no paramètrica (per la NO normalitat de la variable dependent) i tenim diverses mostres: el professorat procedent de quatre àrees científiques, la prova adequada és Kruskal-Wallis, que s'obté en SPSS en el quadre de proves no paramètriques, per k mostres independents, com es pot veure a la figura següent.



10. Si l'*output* que s'obté després d'aplicar la prova adequada és el següent, podem concloure que els professors d'alguna àrea d'estudi en concret pensen que les TIC són més útils que la resta de professors?

PROVA DE MANN-WHITNEY

Rangs			
Àrea d'estudi		N	Rang mitjana
Utilitat de les TIC per a l'ensenyament	Ciències naturals	20	52,98
	Ciències socials i educació	20	32,23
	Lletres	20	20,30
	Tècniques	20	56,50
Total		80	

ESTADÍSTICS DE CONTRAST^{ab}

Utilitat de les TIC per a l'ensenyament	
Khi quadrat	36,099
gl	3
Significació asintòtica (bilateral)	,000

a. Prova de Kruskal-Wallis.

b. Variable d'agrupació: àrea d'estudi

SOLUCIÓ:

S'accepta la H_1 i es conclou que l'àrea d'estudi a la qual pertany el professorat influeix en la utilitat que atribueixi a les TIC en l'ensenyament (sig. 0,00 < 0,05). El professorat de les àrees de ciències naturals (52,98) i de les àrees tècniques (56,50) considera que les TIC són més útils que la resta de professorat.

Referències bibliogràfiques

- ÁLVAREZ, M. (2000). *Análisis estadístico con SPSS: procedimiento básico*. Bilbao: Deusto.
- BERLANGA, V.; RUBIO, M. J. (2012). «Classificació de proves no paramètriques. Com aplicar-les en SPSS». *Revista d'innovació i recerca en educació*, vol. 5, 2.
- BISQUERRA, R. (1987). *Introducción a la estadística aplicada a la investigación educativa. Un enfoque informático con los paquetes BMDP y SPSS-X*. Barcelona: PPU.
- (2004). *Metodología de la investigación educativa*. Madrid: La Muralla.
- ECHEVERRÍA, B. (1982). *Estadística aplicada a las Ciencias Humanas*. Barcelona: Daimon.
- FERRÁN ARANAZ, M. (1996). *SPSS para Windows. Programación y análisis estadístico*. Madrid: Mc Graw-Hill.
- (2002). *Curso de SPSS para Windows*. Madrid: McGraw-Hill.
- FERRÁN ARANAZ, M.; GONZÁLEZ ARANGÜENA, E.; MANUEL GARCÍA, C. M.; ORTEGA CASTELLÓ, E. (2001). *SPSS para Windows: Análisis estadístico*. Madrid: McGraw-Hill Interamericana de España.
- FILGUEIRA, E. (2001). *Análisis de datos con SPSSWIN*. Madrid: Alianza.
- FREIXA, M.; SALAFRANCA, L.; GUARDIA, J.; FERRER, R.; TURBANY, J. (1992). *Análisis exploratorio de datos: nuevas técnicas estadísticas*. Barcelona: PPU.
- GIL, X.; ORMAZÁBAL, J.; TORRADO, M. (1993). *Estadística aplicada a las ciencias humanas II: Bloque: Análisis de la variancia y pruebas no paramétricas*. Barcelona: PPU.
- GÓMEZ, A.; DÍEZ-PALOMAR, J.; ORMAZÁBAL, J.; FLECHA, R.; VILÀ, R. (2012). *Estadística básica para educadores*. Madrid: Síntesis.
- HOPKINS, K. D.; HOPKINS, B. R.; COLASS, G. V. (1997). *Estadística básica para las ciencias sociales y del comportamiento*. Mèxic: Prentice Hall.
- NAVARRO, A.; MARTÍN, M. (2002). *Uso profesional del SPSS: autoaprendizaje a través de un estudio real*. Bellaterra: Servei de Publicacions, UAB.
- PEÑA, D. (1999). *Estadística. Modelos y Métodos. Tomo I: Fundamentos*. Madrid: Alianza Universidad Textos.
- (2001). *Fundamentos de Estadística*. Madrid: Alianza.
- PÉREZ JUSTE, R.; GARCÍA LLAMAS, J. L.; GIL PASCUAL, J. A.; GALÁN GONZÁLEZ, A. (2009). *Estadística aplicada a la Educación*. Madrid: UNED – Pearson.
- PÉREZ-LÓPEZ, C. (2001). *Técnicas estadísticas y SPSS*. Madrid: Prentice Hall.
- RUBIO, M. J.; BERLANGA, V. (2012). «Com aplicar les proves paramètriques bivariades t de Student i ANOVA en SPSS. Cas pràctic». *Revista d'innovació i recerca en educació*, vol. 5, 2.
- TORRADO, M. (2006). «Com utilitzar l'estadística en la recerca educativa?». *Butlletí La Recerca*, núm. 10.
- TORRADO, M. (coord.). (2001). *Estadística descriptiva básica* (CD-ROM). Barcelona: Edicions UB.
- VILÀ, R. (2006). «Com fer una anàlisi quantitativa de dades de tipus descriptiu amb el paquet estadístic de l'SPSS?». *Butlletí La Recerca*, núm. 6.
- (2008). «Com optimitzar una matriu de dades? Utilitats bàsiques de l'SPSS». *Butlletí La Recerca*, núm. 11.
- VISAUTA VINACUA, B. (2007). *Análisis estadístico con SPSS 14: Estadística básica* (3a ed.). Madrid: McGraw-Hill Interamericana.
- VISAUTA, B.; BATALLÉ, P. (1991). *Métodos estadísticos aplicados. Estadística descriptiva*. Barcelona: PPU.
- WELKOWITZ, J. et al. (1981). *Estadística aplicada a las ciencias de la educación*. Madrid: Santillana.
- YEPES, M. (2006). «Com puc passar una base de dades d'Excel a l'SPSS?». *Butlletí La Recerca*, núm. 5.