Mètodes d'investigació quantitativa

PID_00258287

Neus Calaf Gozalo

Temps mínim de dedicació recomanat: 4 hores

Universitat Oberta de Catalunya

Neus Calaf Gozalo

La revisió d'aquest recurs d'aprenentatge UOC ha estat coordinada pel professor: Sergi Fàbregues Feijóo (2019)

Primera edició: febrer 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Disseny: Manel Andreu
Realització editorial: Oberta UOC Publishing, SL

Índex

Introducció

Quan una persona es planteja iniciar els estudis de Logopèdia, en molts casos no s'imagina que haurà de cursar matèries tan diverses, ni encara menys que en algun moment haurà d'estudiar estadística per a obtenir la titulació. Tanmateix, l'estadística és una eina imprescindible per a resoldre problemes i prendre decisions en qualsevol context científic. Caldrà, doncs, que aparqueu la possible «por dels números» que pugueu tenir i que considereu que les dades que aconseguireu en les vostres investigacions, o al llarg de la vida laboral, són les úniques que us podran assegurar que aconseguiu els objectius que us heu proposat.

Però les dades «brutes» o «crues», sense ordenar ni reduir, no permeten mostrar-ne el significat, sinó que conformen una quantitat més o menys ingent d'informació caòtica que cal estructurar i sintetitzar mitjançant les diverses tècniques d'anàlisi estadística. De fet, aquesta és la finalitat última del mòdul: que l'alumnat aprengui a **donar sentit a les dades** i a **fer-les interpretables**.

Atès el caràcter instrumental d'aquest mòdul, a més de proporcionar les eines conceptuals que permetran a l'alumnat analitzar les dades obtingudes en les seves investigacions, també proporcionarem els elements necessaris per a poder analitzar de manera crítica els resultats i els procediments estadístics utilitzant els diferents informes d'investigacions (articles, informes, llibres, tesis, etc.) que puguin ser-los d'interès.

En finalitzar aquest mòdul, l'alumnat tindrà coneixements elementals d'estadística descriptiva, probabilitat i inferència estadística, de manera que podrà planificar, analitzar i interpretar, rigorosament i amb les tècniques estadístiques apropiades, les dades obtingudes amb els dissenys d'investigació en qualsevol de les modalitats en un nivell elemental.

Lectura recomanada

A. Cosculluela; A. Fornieles; J. Turbany (2014). *Tècniques d'anàlisi de dades quantitatives*. Material docent de la UOC. Universitat Oberta de Catalunya.

Probablement, aquest és el primer contacte que molts de vosaltres deveu tenir amb l'estadística. Esperem que sigui profitós i motivador per a tothom.

Objectius

En finalitzar el mòdul, l'alumnat podrà:

1. Comprendre i formular problemes substantius d'investigació i identificar les variables que hi intervenen.

2. Conèixer els conceptes bàsics per a analitzar les dades de les investigacions necessàries per a solucionar els problemes de la manera més rigorosa possible, d'acord amb els criteris científics.

3. Saber escollir els subjectes mitjançant un mostreig acurat que permeti obtenir mostres representatives de la seva població.

4. Conèixer les característiques principals de les distribucions de les dades, la descripció de variables, tant quantitatives com categòriques, i la presentació dels resultats mitjançant la utilització de taules, índexs estadístics i gràfiques.

5. Fer inferències i estudiar associacions entre variables quantitatives, tenint en consideració, però, el concepte de probabilitat que hi ha darrere d'aquestes operacions.

6. Conèixer el full de càlcul de l'aplicació gratuïta Google Sheets per a fer les operacions estadístiques i obtenir els índexs necessaris.

1. Conceptes estadístics bàsics

1.1. Què és l'estadística?

El mot *estadística* deriva de la paraula *estat*. Durant el segle XIX, l'estadística es considerava la ciència de l'estat. Després va anar més enllà d'aquest límit i va adquirir una aplicació més universal. De fet, el cert és que l'estadística penetra en gairebé tots els aspectes de la nostra vida, i es pot utilitzar per a aconseguir una millor interpretació de tots els fenòmens que observem.

L'estadística es basa en la recopilació i l'anàlisi de dades. En l'apartat següent veureu que la distinció entre els dos tipus de dades possibles (quantitatives o numèriques, i qualitatives o categòriques) és crucial, ja que l'anàlisi de dades que es pot fer depèn del tipus de variable. L'hora en què cau el primer llamp en una tempesta, per exemple, és una variable numèrica en l'estudi meteorològic; la presència o l'absència d'un organisme marí és una variable categòrica en l'estudi ambiental, i l'assignació de feines és una variable categòrica en l'estudi de les lleis discriminatòries.

En cada situació hi ha un objectiu específic en la recopilació de dades. Per exemple, en la recopilació de dades sobre el primer llamp que cau, el meteoròleg o la meteoròloga vol esbrinar a quina hora del dia és més probable que caigui un llamp, i l'estudi proposa preparar-se millor per als perills que comporta. En canvi, en reunir dades sobre l'alçada d'un nen, el personal mèdic vol determinar-ne el ritme de creixement i comprovar que és normal.

També **podem observar dues maneres diferents de recopilar dades**. D'una banda, simplement s'observen les dades tal com s'esdevenen naturalment; per exemple, cau un llamp i nosaltres observem l'hora o el lloc on cau. D'altra banda, es poden reunir dades mitjançant l'experimentació. Per exemple, en un estudi sobre un fàrmac no s'estudien 20.000 persones i s'observa simplement quines tenen un atac de cor i quines han pres el fàrmac per a veure si hi ha una connexió. En aquest cas, es divideix la gent en dos grups aleatòriament (com a cara o creu) i després es determina que un grup prengui el fàrmac i l'altre no. No sempre podem (o no sempre té sentit) dur a terme experiments d'aquest tipus, però són més potents a l'hora de demostrar resultats vertaders o causals.

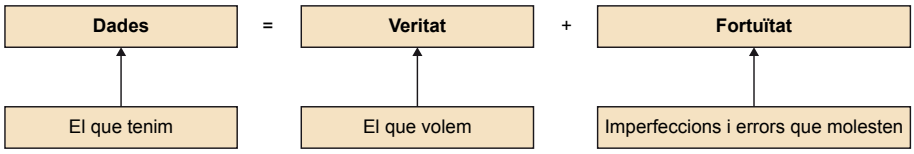
L'estadística s'utilitza per a descriure i analitzar les dades. Per exemple, en l'estudi del creixement d'una nena s'observen dues variables, l'alçada i l'edat, i es representen les dades de l'alçada contra les de l'edat en el que anomenem **diagrama de dispersió**, que consisteix en una descripció de les dades. No obstant això, en estudis previs els metges han establert el ritme de creixement normal per als infants. Per mitjà d'aquests gràfics, el personal mèdic pot de-

duir si hi ha una probabilitat alta que un nen o una nena no creixi prou de pressa. Aquesta anàlisi visual de les dades porta a una conclusió (instaurar el tractament).

Un altre aspecte que cal considerar és que **les dades que observem no són perfectes**. Pot haver-hi tot tipus d'errors, tant en l'observació com en la categorització, o en el registre de la informació. En les investigacions també cal tenir en compte quants subjectes s'han triat per a l'estudi i com s'han triat. Si poguéssim preguntar a totes les persones de Catalunya, d'una en una, si treballen o no, llavors tindríem una mesura perfecta del grau d'ocupació de la població (**població** és el total d'elements sobre els quals volem extrapolar el nostre estudi). No obstant això, habitualment hem de recórrer a fer la pregunta a una mostra de la població (**mostra** és un subconjunt de la població sobre el qual fem la nostra anàlisi de dades), la qual cosa significa que les nostres dades no són perfectes.

Totes les dades consten d'una part vertadera i d'una d'errònia, que nosaltres anomenem «fortuïtat» (figura 1), és a dir, un element que és imprevisible i que és fora del nostre control (encara que esperem i fem tot el possible perquè sigui molt petit). **L'anàlisi estadística té el propòsit de separar la veritat de la fortuïtat**, de manera que puguem treure conclusions fermes d'allò que observem. Es tracta d'un tema recurrent en aquesta assignatura i del qual parlarem amb freqüència.

Figura 1



Hi ha una seqüència d'esdeveniments comuna en qualsevol investigació que concerneixi l'estadística:

- En primer lloc, es troba la definició d'un problema i els seus objectius.
- En segon lloc, es reuneixen dades de les variables rellevants.
- En tercer lloc, es descriuen i possiblement s'analitzen les dades, la qual cosa porta a una conclusió amb relació a l'objectiu de l'estudi.

Aquest mòdul se centra principalment en la tercera part: la descripció i l'anàlisi de les dades dirigides a prendre decisions.

1.2. Variables i matrius de dades

En planificar una investigació, cal delimitar els aspectes de la realitat que es volen investigar. Quan operem amb dimensions (característiques, fenòmens, etc.) que poden prendre diferents valors que són mesurables, parlem de **vari-**

ables. Recordeu que en el mòdul «Introducció a la recerca en logopèdia» hem estudiat les diferents maneres de classificar les variables. En aquest mòdul estudiarem com analitzar-les.

En el pla de notació, quan considerem un conjunt de *n* observacions numèriques d'una variable *X*, denotem els valors genèrics amb els símbols x_1, x_2, x_3, etc., fins a x_n. Denotem aquest conjunt d'observacions amb x_1, x_2, x_3... x_n; o amb x_i, $i = 1... n$, en el qual el símbol *i* utilitzat en els subíndexs es denomina **índex**. Així, per a les dades de la taula 1: $x_1 = 9$, $x_2 = 5$, $x_3 = 6$... $x_{27} = 12$.

Taula 1

9	5	6	8	8	9	12	3	7
3	11	8	4	5	2	6	4	8
17	3	13	11	7	7	4	8	12

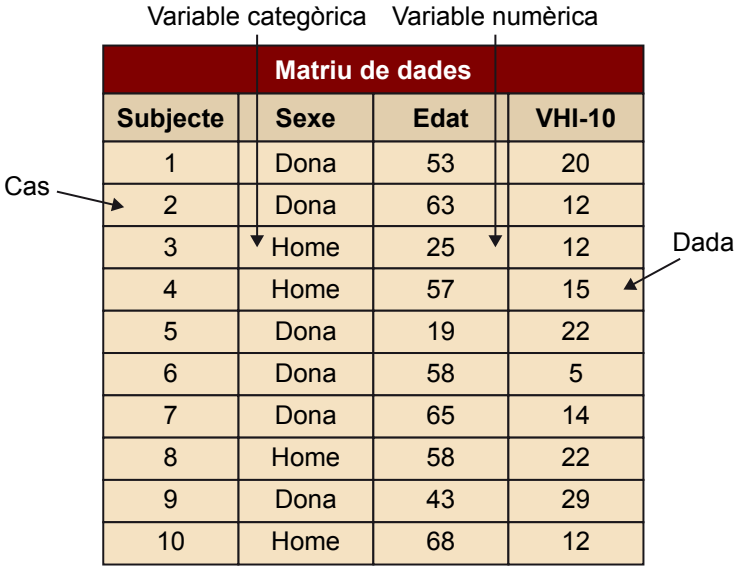
A l'hora d'ordenar les observacions, de més petita a més gran, denotem el nou conjunt de quantitats amb els símbols $x_{(1)}$, $x_{(2)}$, $x_{(3)}$, etc., fins a $x_{(n)}$. Per tant, $x_{(1)}$ és el valor més baix i $x_{(n)}$ és el més alt. En el nostre exemple: $x_{(1)} = 2$, i $x_{(27)} = 17$.

Un cop recollides les dades, el primer pas és tabular-les, és a dir, introduir-les en una matriu de subjectes (files) per a variables (columnes). Aquesta és precisament l'estructura que tenen els fulls de càlcul, com ara l'Excel. En la figura 2 es mostra un exemple de matriu de dades que recullen els valors de les variables sexe, edat i hàndicap vocal, mesurat amb el VHI-10 (Rosen i altres, 2004), de 10 subjectes amb alteracions de la veu que estan seguint tractament logopèdic.

Referència bibliogràfica

C. A. Rosen; A. S. Lee; J. Osborne; T. Zullo; T. Murry (2004). «Development and validation of the voice handicap index-10». *The Laryngoscope* (núm. 114, vol. 9, pàg. 1549-1556).

Figura 2



Com veurem més endavant, en l'ús de determinades eines estadístiques en l'estudi de variables categòriques dicotòmiques (com per exemple, el sexe), ens caldrà recodificar les variables amb 0 i 1, assignant l'1 al valor de l'atribut estudiat.

2. Estadística descriptiva

Les diverses tècniques d'anàlisi estadística tenen l'objectiu de donar sentit a les dades i fer-les interpretables. D'una banda, els conjunts de dades es poden organitzar, simplificar i resumir mitjançant procediments d'**estadística descriptiva**. D'altra banda, es poden inferir o deduir possibles resultats d'una població sotmesa a estudi a partir de l'anàlisi de mostres de la mateixa població, utilitzant procediments d'**estadística inferencial**.

En aquest apartat del mòdul presentarem diversos procediments d'estadística descriptiva i els ordenarem en dos grans subapartats: procediments per a variables quantitatives i procediments per a variables qualitatives.

2.1. Variables quantitatives

2.1.1. Histograma

Un **histograma** és una manera de representar gràficament una distribució de freqüències de dades quantitatives. És un gràfic molt útil quan volem veure l'aspecte del conjunt de la distribució d'un gran nombre d'observacions.

Per a construir-lo, es dibuixa una barra vertical que mostra el nombre de valors de les nostres dades que són dins de cada classe, interval o segment de l'histograma. Les classes, intervals o segments d'un histograma cobreixen tota l'escala de valors de la variable. Hi ha molta llibertat a l'hora de decidir les classes d'un histograma. No obstant això, hi ha dues consideracions importants que cal tenir en compte:

- Totes les classes han de tenir la mateixa amplitud.
- El nombre de classes depèn de la quantitat de dades i del detall amb què interessi veure la distribució. No hi ha cap regla estricta per a fer-ho, únicament és una qüestió de sentit comú.

Es pot dibuixar i personalitzar un histograma amb les opcions «Insereix > Gràfic» i «Tipus de gràfic: Gràfic» d'histogrames de l'aplicació Google Sheets.

Exemple

Vegeu en la figura 3 un exemple d'histograma de les qualificacions que un grup de setanta estudiants han obtingut en un examen del grau de Logopèdia:

Figura 3

Per a interpretar un histograma, cal examinar-ne els patrons generals i després buscar-hi les desviacions. En el nostre histograma, per exemple, veiem com el centre de la distribució està entre els 6 i els 8 punts. Les quantitats petites de valors que se separen de la distribució es denominen **valors allunyats** o **insòlits**. En el nostre exemple veiem un d'aquests valors insòlits, corresponent a una qualificació entre 1 i 2.

Si l'histograma no és simètric, la part llarga i arrossegada de la distribució asimètrica es denomina **cua**. Una distribució pot ser asimètrica per l'esquerra o asimètrica per la dreta. En la pràctica, és més freqüent trobar l'asimetria per la dreta.

2.1.2. Mesures de tendència central

En aquest apartat veurem tres maneres de mesurar en un sol valor el centre d'una distribució: la mediana, la mitjana aritmètica i la moda.

Mediana o observació central

La **mediana** és el valor que **divideix la distribució de les dades en dues parts iguals** (deixa un 50% de valors per sobre i un altre 50% per sota). Es tracta, doncs, d'un **índex de posició**.

Per a trobar la mediana, hem d'ordenar les dades de més petites a més grans i trobar l'observació que queda exactament al mig, cosa que implica que la meitat de les observacions queden per sota d'aquest valor i l'altra meitat per sobre. Atès que és un índex de posició, la mediana no queda afectada per la presència de valors extrems. Per això diem que és un índex resistent o robust.

El valor de la mediana d'un conjunt de dades numèriques es pot obtenir amb la funció «MEDIAN» de l'aplicació Google Sheets.

Mitjana aritmètica o valor mitjà

La **mitjana aritmètica**, per contra, és un **índex de pes basat en el moment de la distribució** (en realitat, la podem definir com el centre de gravetat de la distribució) i es calcula sumant tots els valors de les dades i dividint aquest sumatori pel nombre d'observacions (n).

El valor mitjà numèric d'un conjunt de dades es pot obtenir amb la funció «AVERAGE» de l'aplicació Google Sheets.

Tant la mediana com la mitjana aritmètica mesuren el centre de la distribució, però ho fan de manera diferent. Només quan la distribució és simètrica, les dues mesures coincideixen. La principal diferència entre ambdues és com estan afectades per les asimetries o per les dades allunyades.

Exemple

Imaginem que al llarg d'un període de vint-i-set dies anoteu l'estona que heu d'esperar fins que l'autobús arriba al matí. Les dades, en minuts, es mostren a la taula 2.

Taula 2: Temps d'espera fins que arriba l'autobús, en minuts

9	5	6	8	8	9	12	3	7
3	11	8	4	5	2	6	4	8
17	3	13	11	7	7	4	8	12

La mediana d'aquests valors és 7 (funció «MEDIAN»), mentre que la mitjana aritmètica, en canvi, és 7,41 (funció «AVERAGE»). En el cas de la mitjana aritmètica, cal prendre precaucions amb les dades allunyades o insòlites. Quan la distribució és asimètrica, com en aquest cas, la mitjana aritmètica sempre es desplaça cap a la cua de la distribució.

La presència d'un valor molt elevat no afecta la mediana, però influeix molt en la mitjana aritmètica. Diem que la mediana «resisteix» les dades allunyades. Per exemple, imaginem que, en lloc de 17 minuts, el valor més alt en les dades de l'exemple fos 45 minuts, que és una espera molt llarga per a un sol dia. Aquest canvi no afecta la mediana, de fet es manté igual, fins i tot si ho canviéssim per un valor molt més elevat. La mitjana aritmètica, en canvi, quedaria afectada, ja que la suma de totes les observacions seria 228, que, dividit per 27, dona un valor de 8,44 minuts. Aquest increment d'una observació fa pujar la mitjana aritmètica del temps d'espera en un minut, malgrat que els altres vint-i sis valors es mantinguin intactes. En una situació com aquesta, la mitjana aritmètica perd la condició de ser un valor representatiu.

Moda

La **moda** és el valor que es repeteix més vegades en un conjunt de dades. Es pot aplicar tant a variables quantitatives com qualitatives.

El valor que apareix amb més freqüència en un conjunt de dades es pot obtenir amb la funció «MODE» de l'aplicació Google Sheets. En el nostre exemple dels minuts d'espera de l'autobús, el valor que es repeteix més és 8.

2.1.3. Mesures de dispersió

En l'apartat anterior hem definit tres maneres de calcular els índexs del centre d'una distribució. Tanmateix, per a completar la descripció de les variables quantitatives cal afegir-hi els índexs de dispersió, que indiquen fins a quin punt les observacions es dispersen al voltant del centre.

Variància

La **variància** es pot definir com la mitjana aritmètica dels quadrats de les diferències que hi ha entre cada valor i la mitjana aritmètica. Això fa que, com més grans siguin aquestes diferències o distàncies (més dispersa o heterogènia sigui la variable), més gran serà el valor de la variància.

El fet que les diferències s'elevin al quadrat:

- Evita la presència de valors negatius (si no s'elevessin les diferències al quadrat, en haver-hi alguns valors per sobre i altres per sota de la mitjana, el sumatori seria 0).

- Fa que les diferències més grans pesin més en el valor de l'índex.

- Implica que la variància sigui sempre de signe positiu i estigui en la unitat de mesura de la variable elevada al quadrat (per exemple, el quocient d'intel·ligència, QI, té en la població una mitjana $\mu = 100$ punts de QI, i una variància $\sigma^2 = 225$ punts^2 de QI).

La variància, basant-se en una mostra, es pot obtenir utilitzant la funció «VAR» de l'aplicació Google Sheets. En el nostre exemple dels minuts d'espera de l'autobús, la variància és de 12,94 minuts.

Per a facilitar-ne la interpretació, en lloc de la variància s'acostuma a presentar l'arrel quadrada que, per tant, ja és en les mateixes unitats de mesura que la variable. Aquest índex es denomina **desviació estàndard**, **tipus** o **típica**.

Desviació estàndard, tipus o típica

La **desviació estàndard**, **tipus** o **típica** és l'arrel quadrada de la variància. Es tracta d'un valor únic que resumeix la dispersió de les dades, en concret, la dispersió al voltant de la mitjana aritmètica. És un dels índexs de dispersió més utilitzats. Per exemple, la desviació tipus del QI en la població és de σ = 15 punts de QI.

La desviació estàndard, basant-se en una mostra, es pot obtenir utilitzant la funció «STDEV» de l'aplicació Google Sheets. En el nostre exemple dels minuts d'espera de l'autobús, la desviació estàndard és de 3,6 minuts.

Les extensions de diferents distribucions es poden comparar simplement comparant-ne les respectives desviacions estàndard. Ara bé, quan es tracta de comparar variables mesurades en unitats diferents, hem d'utilitzar el **coeficient de variació** (CV), que elimina les unitats de les variables, la qual cosa en facilita la comparació. El càlcul d'aquest coeficient és molt senzill: s'obté dividint la desviació estàndard entre la mitjana aritmètica. El resultat és típicament menor que 1, però per a la seva millor interpretació s'expressa com a percentatge multiplicant el resultat per 100.

2.1.4. Mesures de posició

En l'apartat anterior ja hem vist una mesura de posició: la mediana. Recordem que la mediana és el valor que divideix la distribució de les dades en dues parts iguals, tot deixant un 50% de valors per sobre i un altre 50% per sota. Altres mesures de posició són els **percentils** i els **quartils**.

Percentils

Els **percentils** (P_1, P_2, P_3, ... P_{99}) són els 99 valors que divideixen la distribució en 100 parts iguals. En cada part hi haurà l'1% de la distribució. El percentil *k* (P_k) és el valor que deixa per sota el *k* per cent de les puntuacions d'una distribució. Per exemple, el percentil 90 (P_{90}) és el valor que deixa per sota el 90% de les dades d'una distribució i és superat pel 10% restant de les dades.

El valor corresponent a un percentil determinat d'un conjunt de dades es pot obtenir amb la funció «PERCENTILE» de l'aplicació Google Sheets. En el nostre exemple del temps d'espera de l'autobús, el percentil P_{95} és 12,7 minuts.

El **rang percentil**, en canvi, és la mesura inversa del percentil. El rang percentil d'un valor determinat es defineix com el percentatge de dades amb valors inferiors a aquest valor, i permet avaluar la posició relativa d'un valor en un conjunt de dades.

La classificació percentual (percentil) d'un valor especificat d'un conjunt de dades es pot obtenir amb la funció «PERCENTRANK» de l'aplicació Google Sheets. En el nostre exemple del temps d'espera de l'autobús, el rang percentil de 8 minuts és 0,54 (això vol dir que el valor 8 minuts té el 54% dels valors per sota).

Exemple

En l'avaluació logopèdica dels trastorns del llenguatge, els percentils s'utilitzen per a situar un usuari respecte a la mitjana poblacional i així poder valorar la severitat del trastorn que presenta. Això és possible gràcies a l'estandardització i a la normalització dels instruments d'avaluació, que permeten comparar les puntuacions obtingudes per l'usuari amb les puntuacions típiques del mateix grup d'edat.

Quartils

Els **quartils** (Q_1, Q_2 i Q_3) són els tres valors que divideixen la distribució en quatre parts iguals. En cada part hi haurà el 25% de la distribució. Equivalen als percentils P_{25}, P_{50} i P_{75}, respectivament (o, dit d'una altra manera, al 25è., 50è., i 75è. percentils). El primer quartil és el valor que deixa el 25% de les observacions per sota, el segon coincideix amb la mediana i, per tant, és el valor que divideix la distribució en dues parts iguals, i el tercer quartil correspon al valor que deixa el 75% dels valors per sota (i, lògicament, en queda el 25% per sobre).

El càlcul dels quartils és molt senzill, ja que podem dir que els quartils 1r. i 3r. són la mediana de les dues meitats de la distribució que queden definides per la mediana. Una vegada calculats els quartils, restant el 3r. del 1r. ($Q_3 - Q_1$) podem obtenir el **rang interquartílic**, que ens indica quina és la dispersió del 50% central de les observacions.

El valor més pròxim a un quartil especificat d'un conjunt de dades es pot obtenir amb la funció «QUARTILE» de l'aplicació Google Sheets. En el nostre exemple del temps d'espera de l'autobús, el valor del tercer quartil Q_3 és de 9 minuts, i el valor del primer quartil Q_1 és de 4,5 minuts. El rang interquartílic ($Q_3 - Q_1$) és de 4,5 minuts.

Exemple

En les publicacions científiques s'utilitzen els quartils per a avaluar el posicionament relatiu d'una revista dins del total de revistes de la mateixa matèria. Així, doncs, les revistes del Q_1 són les que estan entre el 25% de revistes amb un factor d'impacte més elevat (entenent el factor d'impacte d'una revista com la mitjana de vegades que, en un any en concret, van citar-se els articles publicats per la revista els dos anys anteriors).

2.1.5. Resum dels cinc nombres

El **resum dels cinc nombres** de les dades d'una distribució (o sumari de Tukey) dona una idea ràpida de la tendència central i de la dispersió d'un conjunt de dades. Aquests cinc números són els següents:

- **Mínim**: valor més petit de la mostra.
- **Q_1**: primer quartil o percentil P_{25}.
- **Mediana**: Q_2, segon quartil o percentil P_{50}.
- **Q_3**: tercer quartil o percentil P_{75}.
- **Màxim**: valor més gran de la mostra.

El valor mínim d'un conjunt de dades numèriques es pot obtenir amb la funció «MIN» de l'aplicació Google Sheets, i el valor màxim amb la funció «MAX». Com hem vist abans, el valor dels quartils s'obté amb la funció «QUARTILE», i el de la mediana amb la funció «MEDIAN».

A partir d'aquests nombres també es pot obtenir el **rang interquartílic** ($Q_3 - Q_1$) i el **rang** (o **recorregut** o **amplitud**) d'una variable, que s'obté calculant la diferència entre el valor màxim i el mínim. El rang interquartílic ens indica quina és la dispersió del 50% central de les observacions. El rang (o recorregut) indica la dispersió del 100% de les mostres i, per desgràcia, es tracta d'un índex d'escassa utilitat, ja que un únic valor extrem o insòlit pot fer que perdi gran part del seu sentit informatiu.

2.1.6. Diagrama de caixa o *box plot*

El resum dels cinc nombres es pot representar gràficament mitjançant un **diagrama de caixa** o ***box plot***. Aquest gràfic és de gran utilitat perquè, a més de ser una representació gràfica de la variable, permet comparar distribucions de la mateixa variable provinents de diferents mostres o subgrups.

Una manera molt senzilla de generar diagrames de caixa és per mitjà de l'eina web BoxPlotR[1].

(1)Eina web disponible a: http://shiny.chemgrid.org/boxplotr/.

Exemple

En la figura 4 es mostra el *box plot* de les dades d'espera de l'autobús en minuts generat amb BoxPlotR. A l'esquerra hi ha el gràfic i, a la dreta, les dades estadístiques a partir de les quals es genera el diagrama. Fixeu-vos que l'eina genera el diagrama amb 26 punts en comptes de 27. Això és així perquè considera que la dada màxima (17 minuts) és un valor allunyat o insòlit, i el marca amb un punt fora del diagrama de caixa.

Figura 4

	X9
Upper whisker	13.00
3rd quartile	9.00
Median	7.00
1st quartile	4.00
Lower whisker	2.00
Nr. of data points	26.00

2.2. Variables qualitatives

2.2.1. Taula de freqüències

La **taula de freqüències** o distribució de freqüències és una taula de les dades estadístiques on s'assigna a cada dada la seva corresponent freqüència.

Exemple

Considereu, per exemple, la taula de freqüències (taula 3) en la qual hi ha la nacionalitat dels assistents a un congrés de logopèdia que s'ha fet a París. Les dades ja estan recollides en forma de freqüència, amb les proporcions i els percentatges calculats. La variable categòrica és **país**, amb disset països europeus com a categories.

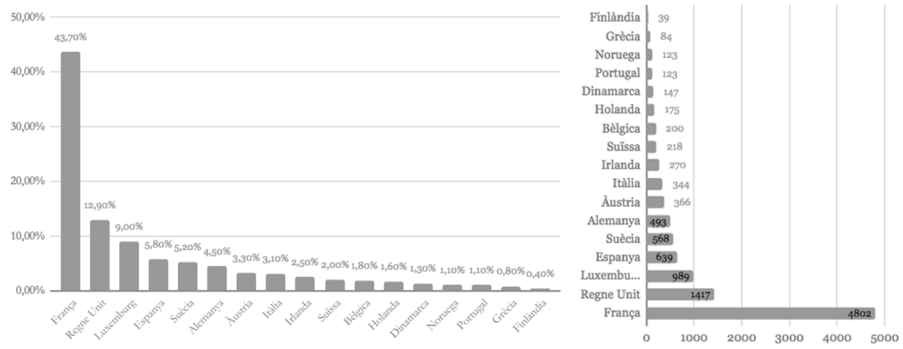
Taula 3

País	Nombre d'assistents	Proporció	Percentatge
França	4.802	0,437	43,7%
Regne Unit	1.417	0,129	12,9%
Luxemburg	989	0,090	9,0%
Espanya	639	0,058	5,8%
Suècia	568	0,052	5,2%
Alemanya	493	0,045	4,5%
Àustria	366	0,033	3,3%
Itàlia	344	0,031	3,1%
Irlanda	270	0,025	2,5%
Suïssa	218	0,020	2,0%
Bèlgica	200	0,018	1,8%
Holanda	175	0,016	1,6%
Dinamarca	147	0,013	1,3%
Noruega	123	0,011	1,1%
Portugal	123	0,011	1,1%
Grècia	84	0,008	0,8%
Finlàndia	39	0,004	0,4%
Sumatoris	10.997	1	100%

2.2.2. Diagrama de barres i de columnes

Podem representar les dades de l'apartat anterior tant amb un gràfic de columnes com amb un gràfic de barres (vegeu figura 5). En el primer cas, hem optat per a fer un gràfic de columnes amb percentatges i, en el segon cas, hem fet un gràfic de barres amb les freqüències.

Figura 5



Es pot dibuixar i personalitzar un diagrama de barres o de columnes amb les opcions «Insereix > Gràfic» i «Tipus de gràfic: Gràfic de barres/de columnes» de l'aplicació Google Sheets.

L'histograma i el diagrama de barres són dos sistemes de representació molt similars que permeten visualitzar la distribució d'una variable. Una diferència entre ambdós és que l'histograma es construeix per a una variable quantitativa després de decidir un conjunt de classes adequades, mentre que el diagrama de barres es construeix per a una variable categòrica en la qual les classes ja estan fetes. Una altra diferència és que a l'histograma les barres es toquen, mentre que en un diagrama de barres estan separades per espais. Encara una altra diferència és que les categories no estan en cap ordre específic i, per tant, podem ordenar-les perquè el diagrama de barres sigui més fàcil d'interpretar.

2.2.3. Gràfic de sectors

Una altra possible representació és amb **gràfics de sectors** (els popularment denominats «formatgets» o «pastissets»). Aquests gràfics poden expressar-se tant amb les freqüències (nombre d'elements en cada categoria) com amb els percentatges.

Es pot dibuixar i personalitzar un gràfic de sectors amb les opcions «Insereix > Gràfic» i «Tipus de gràfic: Gràfic circular» de l'aplicació Google Sheets.

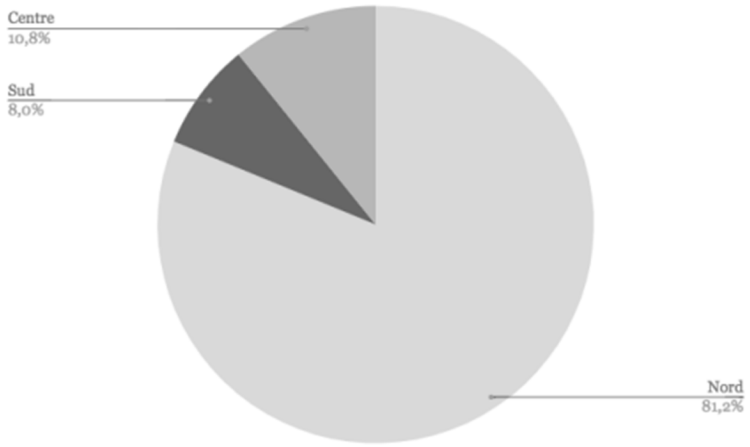
Exemple

Per tal d'exemplificar aquest tipus de gràfics (vegeu la taula 4 i la figura 6), agruparem els països en «Nord» (Suècia, Dinamarca, Noruega i Finlàndia), «Centre» (França, Regne Unit, Luxemburg, Alemanya, Àustria, Irlanda, Suïssa, Bèlgica i Holanda) i «Sud» (Espanya, Itàlia, Portugal i Grècia), i en farem la taula de freqüències corresponent i la gràfica de sectors.

Taula 4

Zona	Nombre d'assistents	Proporció	Percentatge
Nord	8.930	0,81	81,2%
Sud	877	0,08	8,0%
Centre	1.190	0,11	10,8%
Sumes	10.997	1	100%

Figura 6



2.2.4. Taula de contingència

La **taula de contingència** és una taula de doble entrada en la qual es representen de manera conjunta dues variables categòriques, una a les files i una altra a les columnes. S'identifica pel seu ordre, que és igual al nombre de categories de la variable disposada en files (k) i pel nombre de categories de la variable disposada en columnes (l).

Exemple

La següent taula de contingència 3 × 2 representa les variables «estat civil» i «estudis universitaris», i mostra les proporcions de cadascuna de les categories.

Taula 5

	Est. univ. No	Est. univ. Sí	Total
Solter/a	0,18	0,12	0,30
Casat/ada	0,20	0,23	0,43
D'altres	0,11	0,16	0,27
Total	0,49	0,51	1

Dins de la taula, és a dir, en els encreuaments de les categories d'una variable amb les de l'altra variable, hi ha les proporcions conjuntes. A les files i a les columnes «Total» hi ha les proporcions que corresponen a cada una de les categories de les dues variables, denominades **marginals de la taula**. Així, doncs, veiem que, per exemple, la proporció de solters amb estudis universitaris és P(S i S) = 0,12. També veiem que la proporció de subjectes casats és de 0,43, o que la proporció de persones sense estudis universitaris és de 0,49.

De la mateixa manera que hem utilitzat les proporcions, podem utilitzar els percentatges (multiplicant les proporcions per 100), o les freqüències.

3. Probabilitat

La importància de l'estudi de la probabilitat en l'àmbit de l'estadística es deriva del fet que és un dels pilars teòrics fonamentals sobre els quals s'assenta el desenvolupament i l'aplicació de l'**estadística inferencial** (vegeu apartat següent).

3.1. Experiment aleatori, espai mostral i esdeveniment

Un **experiment aleatori** és aquell en què no es pot predir el resultat que sortirà, però del qual sí que es coneixen tots els resultats possibles (per exemple, llançar un dau o una moneda a l'aire).

Es denomina **espai mostral** tot el conjunt de resultats possibles en una determinada situació. L'espai mostral, que també rep el nom de **conjunt mostral**, se simbolitza amb la lletra grega Ω.

Un exemple d'espai mostral són els sis resultats possibles quan es llança un dau. En aquest cas:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Un altre exemple és llançar dues monedes a l'aire. En aquest cas:

$$\Omega = \{CC, CX, XC, XX\}$$

Un **succés** o **esdeveniment** d'un experiment aleatori és un subconjunt del conjunt de possibles resultats d'Ω; per exemple, treure un nombre parell quan es llança un dau. Quan l'esdeveniment conté un sol punt mostral es denomina **esdeveniment elemental**; per exemple, treure un quatre quan es llança un dau.

3.2. Concepte de probabilitat

La **probabilitat** mesura la possibilitat (probabilitat) que pugui ocórrer un determinat esdeveniment quan es duu a terme un experiment aleatori. Segons la definició clàssica, la probabilitat és el quocient entre el nombre de casos favorables i el nombre total de casos possibles:

$$p(A) = \frac{casos\ \ favorables}{casos\ \ possibles}$$

L'inconvenient d'aquesta definició és que només és vàlida o aplicable en situacions en les quals tots els casos possibles són equiprobables, és a dir, tenen la mateixa probabilitat d'aparèixer.

La definició axiomàtica de probabilitat intenta resoldre el problema de l'equiprobabilitat i, a partir d'un espai mostral determinat Ω, assigna a cada esdeveniment A un nombre real, simbolitzat per $p(A)$, perquè compleixi els axiomes següents:

La probabilitat d'un esdeveniment A qualsevol oscil·la entre 0 i 1, és a dir:

$$0 \leq p(A) \leq 1$$

La probabilitat de l'esdeveniment segur és igual a 1 i la de l'esdeveniment impossible és igual a 0, és a dir:

$$p(\Omega) = 1 \quad \text{i} \quad p(\varnothing) = 0$$

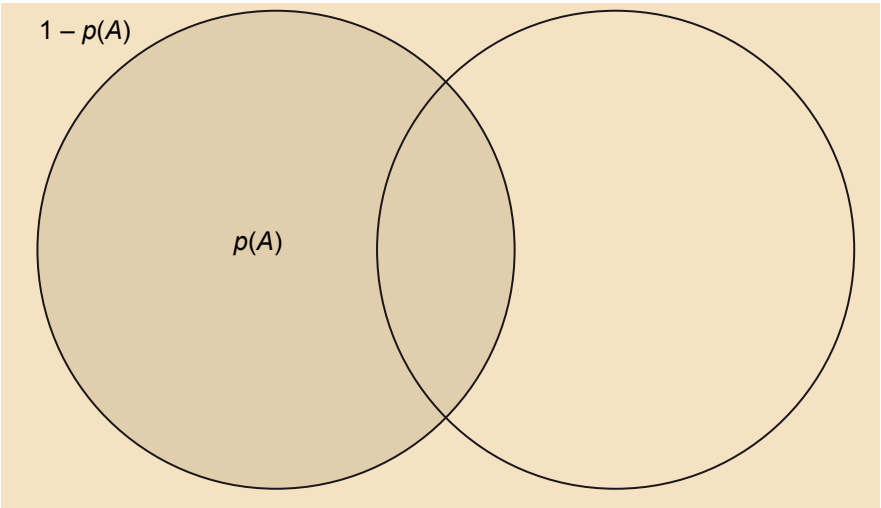
Si A_1, A_2... A_n són esdeveniments mútuament excloents, aleshores:

$$p(A_1 \cup A_2 \cup \ldots A_n) \;=\; p(A_1) \;+\; p(A_2) \;+ \ldots + p(A_n)$$

La probabilitat del complementari de l'esdeveniment A és igual a 1 menys la probabilitat de l'esdeveniment A, és a dir:

$$p(A^C) = 1 - p(A).$$

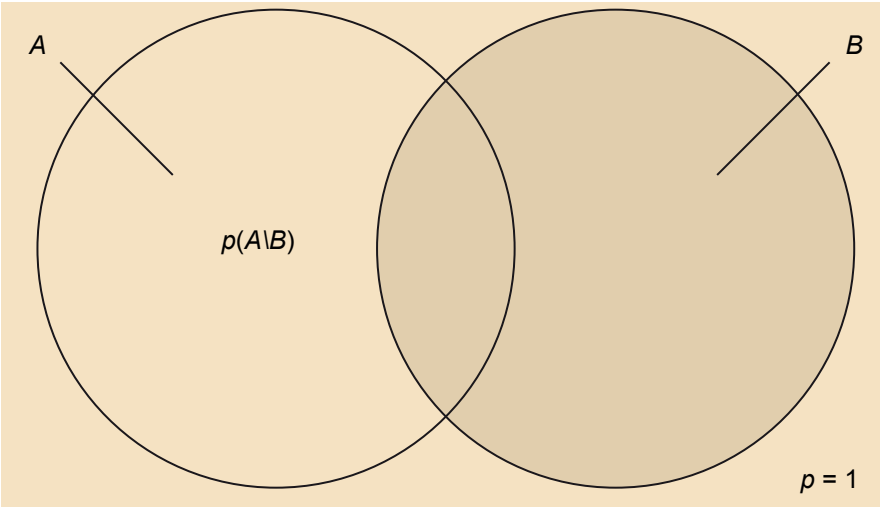
Figura 7



La probabilitat del complementari relatiu de l'esdeveniment B respecte a l'esdeveniment A és igual a la probabilitat de l'esdeveniment A menys la probabilitat de la intersecció dels dos esdeveniments, és a dir:

$$p(A \setminus B) = p(A) - p(A \cap B)$$

Figura 8



Si l'esdeveniment A < esdeveniment B, llavors $p(A) < p(B)$.

Si $A \subset B$, llavors $p(B) \geq p(A)$ i $p(B - A) = p(B) - p(A)$

$p(A \cap B) \leq p(A \cup B)$

3.3. Probabilitat condicionada i independència

3.3.1. Probabilitat condicionada

Donats els esdeveniments A i B, i essent la probabilitat associada a l'esdeveniment B superior a 0 [$p(B) > 0$], la probabilitat que aparegui l'esdeveniment A si s'ha produït l'esdeveniment B es denomina **probabilitat condicionada** d'A donat B, [$p(A/B)$], i es defineix com a:

$$p(A / B) = \frac{p(A \cap B)}{p(B)}$$

Gràficament queda representat en les figures 9 i 10.

Figura 9. Espai mostral Ω

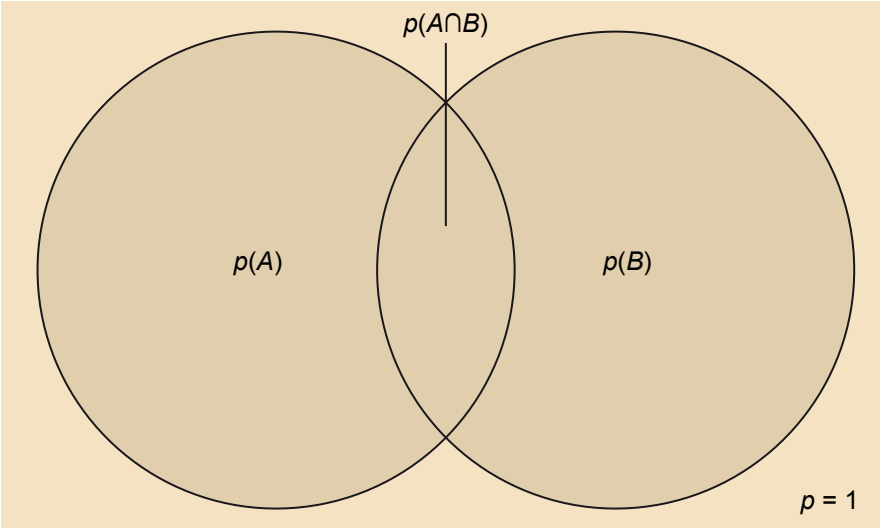
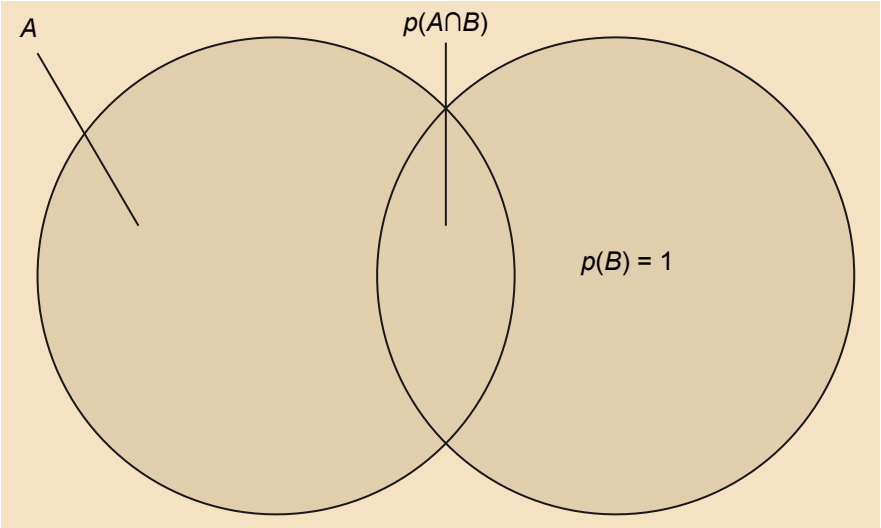


Figura 10. Espai mostral reduït



En aquesta representació gràfica s'observa clarament que l'espai mostral inicial (Ω) queda reduït només en l'espai associat a l'esdeveniment *B*, perquè se sap que aquest ha ocorregut.

3.3.2. Independència

Donats els esdeveniments *A* i *B*, es diu que són independents quan, sabent que se n'ha produït un, no es veu alterada la probabilitat inicial de l'altre, és a dir:

$$p(A/B) = p(A) \text{ o } p(B/A) = p(B)$$

Teorema de la probabilitat total

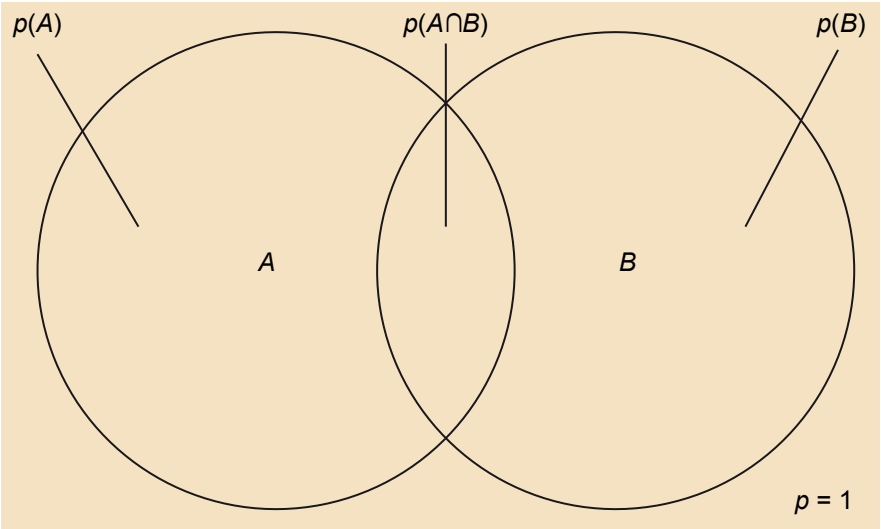
La probabilitat de la unió de dos esdeveniments *A* i *B* és igual a la probabilitat de l'esdeveniment *A* més la probabilitat de l'esdeveniment *B* menys la probabilitat de la intersecció dels dos esdeveniments, és a dir:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Si els esdeveniments A i B són independents o mútuament excloents, la probabilitat de la unió dels dos esdeveniments és igual a la suma de la probabilitat associada a l'esdeveniment A i la probabilitat associada a l'esdeveniment B, és a dir:

$$p(A \cup B) = p(A) + p(B)$$

Figura 11



Teorema del producte

La probabilitat de la intersecció de dos esdeveniments independents és igual al producte de les probabilitats associades als dos esdeveniments, és a dir:

$$p(A \cap B) = p(A) \cdot p(B)$$

La probabilitat de la intersecció de dos esdeveniments no independents és igual al producte de la probabilitat associada a l'esdeveniment A per la probabilitat condicionada de l'esdeveniment B si s'ha donat l'esdeveniment A, o el que és el mateix, la probabilitat associada a l'esdeveniment B per la probabilitat condicionada de l'esdeveniment A, si s'ha donat l'esdeveniment B:

$$p(A \cap B) = p(A) \cdot p(B/A)$$

$$p(A \cap B) = p(B) \cdot p(A/B)$$

3.4. Variables aleatòries

Una variable aleatòria X és una funció definida sobre un espai mostral Ω, de manera que a cada succés elemental d'Ω li fa correspondre un nombre real (R). Aquest espai mostral queda definit per un model o llei de probabilitat que s'estableix a partir de l'associació de cadascun dels valors de la variable aleatòria amb la seva probabilitat corresponent.

S'han de diferenciar dos tipus bàsics de variables aleatòries: les variables **discretes** i les variables **contínues**.

3.4.1. Variables aleatòries discretes

Una variable aleatòria es considera discreta quan el rang de la variable és finit. És a dir, quan s'assignen probabilitats a cada valor concret de la variable X. Una variable aleatòria discreta queda definida per **dues funcions**: la de **probabilitat** i la de **distribució**.

La **funció de probabilitat** assigna a cada valor de la variable discreta la seva probabilitat. La probabilitat associada a cada valor sempre estarà entre 0 i 1. A més, la suma de totes les probabilitats sempre és 1.

Exemple

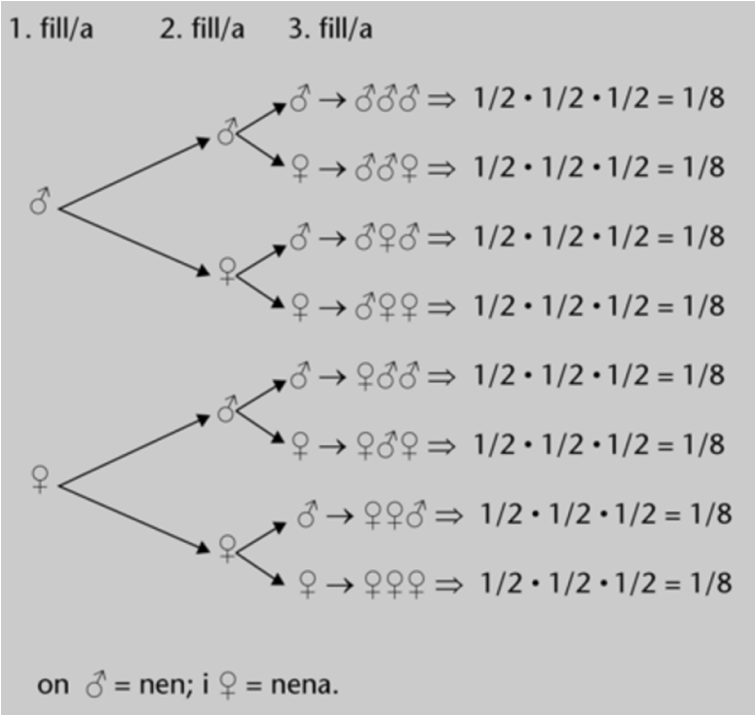
Si una família té tres fills i la probabilitat de ser nen o nena és de 0,5, la funció de probabilitat de la variable «nombre de nens (sexe masculí)» queda reflectida en la taula 6.

Taula 6

Nombre de nens (sexe masculí)			
0	1	2	3
1/8 = 0,125	3/8 = 0,375	3/8 = 0,375	1/8 = 0,125

A partir del diagrama d'arbre es veu clarament com s'obtenen aquestes probabilitats (figura 12).

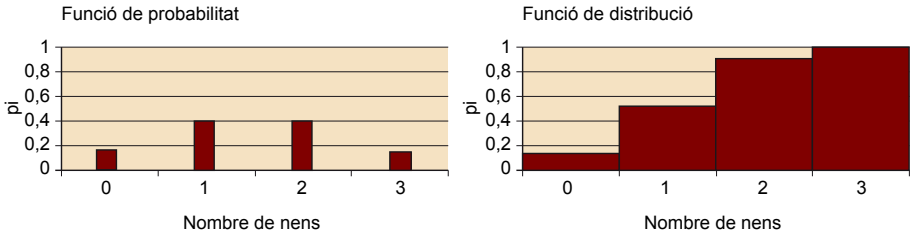
Figura 12



En aquest gràfic s'observa que la probabilitat que dels tres fills que té la família, tots tres siguin nen és d'1/8; que dos siguin nen i un nena és de 3/8 (1/8 + 1/8 + 1/8); que un sigui nen i dues siguin nena és de 3/8 (1/8 + 1/8 + 1/8); i que tots tres siguin nena és d'1/8.

La **funció de distribució** assigna a cada valor de la variable discreta la probabilitat d'obtenir un valor inferior o igual a aquell valor concret. Aquestes dues funcions es representen en la figura 13.

Figura 13

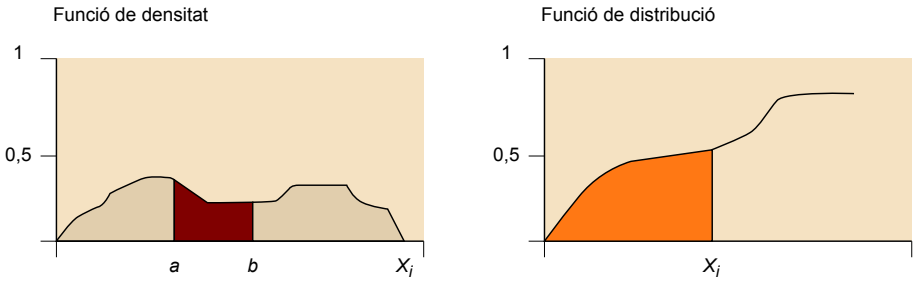


3.4.2. Variables aleatòries contínues

Una variable aleatòria és contínua quan entre dos valors de la variable hi ha un nombre infinit de valors possibles, és a dir, el conjunt imatge és un conjunt continu de nombres, com en un interval. Per exemple, el pes és una variable aleatòria contínua perquè entre dos valors qualsevol (50-51 kg) hi ha un nombre infinit de valors. Les dues funcions que defineixen una variable aleatòria contínua són la **funció de densitat** i la **funció de distribució**.

La **funció de densitat** assigna una probabilitat determinada a un rang o interval de valors de la variable aleatòria contínua [*a*, *b*]. La **funció de distribució** assigna la probabilitat de trobar un valor igual o inferior a x_i; per tant, conceptualment és el mateix que la funció de distribució per a variables aleatòries discretes. Aquestes dues funcions estan representades en la figura 14:

Figura 14



3.4.3. Esperança matemàtica

El concepte d'**esperança matemàtica** (o **esperança** o **mitjana poblacional**) és equivalent al concepte de mitjana en estadística descriptiva, però aplicat a les variables aleatòries. L'esperança matemàtica és el valor mitjà teòric de tots els valors que pot prendre la variable aleatòria. La mitjana de les dades obtingudes amb un experiment aleatori tendirà més al valor de l'esperança matemàtica com més vegades repetim l'experiment.

3.4.4. Variància d'una variable aleatòria

El concepte de **variància d'una variable aleatòria** és equivalent al concepte de variància en estadística descriptiva, però aplicat a les variables aleatòries. La variància d'una variable aleatòria mesura la dispersió mitjana dels valors d'una variable aleatòria respecte a la seva esperança matemàtica. La variància de les dades obtingudes amb un experiment aleatori tendirà més al valor de la variància de la variable aleatòria com més vegades repetim l'experiment. El valor positiu de l'arrel quadrada de la variància és la desviació típica de la distribució de la variable.

3.5. Models de probabilitat

Un model o llei de probabilitat és la correspondència que s'estableix entre cada valor de la variable aleatòria i les probabilitats corresponents. En conseqüència, cada model de probabilitat queda definit per una funció determinada (de probabilitat, o de densitat i de distribució). En alguns casos, les dades s'ajusten a un model de probabilitat perfectament conegut. És per això que en aquest apartat es pretén explicar alguns dels principals models teòrics de probabilitat: la llei binomial, la llei de Poisson i la llei normal o llei de Gauss-Laplace. Les dues primeres s'utilitzen quan es treballa amb variables aleatòries discretes, i l'última quan es treballa amb variables aleatòries contínues.

3.5.1. Models de probabilitat per a variables aleatòries discretes

Distribució binomial

El model de la llei binomial es pot aplicar en el cas de treballar amb variables quantitatives discretes generades a partir d'una variable qualitativa dicotòmica. La variable dicotòmica rep el nom d'**experiment de Bernoulli** (Jacob Bernoulli, 1654-1705). Un dels valors possibles rep el nom d'**èxit** (*E*) i l'altre rep el nom de **fracàs** (*F*).

Cadascun dels dos valors té una determinada probabilitat:

$$p(E) = \pi$$

$$p(F) = (1 - \pi)$$

Si es repeteix l'experiment de Bernoulli un nombre de vegades n i els experiments són independents, és a dir, el valor de π no s'altera en cada repetició, es genera una variable aleatòria discreta amb $n + 1$ valors possibles.

Per exemple, per al cas del llançament d'una moneda, imaginem que l'esdeveniment cara sigui l'«èxit», amb una probabilitat de 0,5. Es pot llançar la moneda tantes vegades com es vulgui (n). De llançament a llançament, el resultat és independent, és a dir, el fet que hagi sortit una cara en el primer llançament no influeix gens quant al resultat cara o creu dels llançaments successius. Al final del procés s'ha generat la variable aleatòria discreta: nombre de cares obtingudes en n llançaments. Aquesta variable es distribueix segons la llei binomial amb $n + 1$ valors (0, 1, 2... n).

Es pot obtenir la probabilitat de la distribució binomial amb la funció «BINOMDIST» de l'aplicació Google Sheets.

Exemple

La probabilitat de patir ansietat davant d'un examen entre la població universitària és de 0,7, i es disposa d'una mostra de quatre persones.

- Quina és la probabilitat que només una de les quatre persones de la mostra pateixi ansietat? Utilitzant la funció «BINOMDIST» amb els paràmetres coneguts (nombre_resultats_correctes: 1; nombre_proves: 4; probabilitat_resultats_correctes: 0,7; i cumulativa: 0) obtenim la probabilitat demanada: 0,0756 (un 7,56%).

- Quina és la probabilitat que hi hagi una o menys persones de la mostra que pateixi ansietat? Utilitzant la funció «DISTR.BINOM» amb els paràmetres coneguts (nombre_resultats_correctes: 1; nombre_proves: 4; probabilitat_resultats_correctes: 0,7; i cumulativa: 1) obtenim la probabilitat demanada: 0,0837 (un 8,37%).

Distribució de Poisson

La distribució de Poisson va formular-se com una particularitat de la distribució binomial pel fet que n tendeix a ser molt gran ($n \to \infty$) i π a ser molt petit ($\pi \to 0$).

Es pot obtenir la probabilitat de la distribució de Poisson amb la funció «POISSON» de l'aplicació Google Sheets.

Exemple

La probabilitat que un nadó neixi amb llavi leporí és de 0,0002 nounats.

- Quina és la probabilitat que en una població de sis mil nounats neixin sis nadons amb aquesta malformació? Utilitzant la funció «POISSON» amb els paràmetres coneguts [x: 6; mitjana (o esperança matemàtica): $E(X)$ = 0,0002 × 6.000 = 1,2; i cumulativa: 0] obtenim la probabilitat demanada: 0,0012 (un 0,12%).

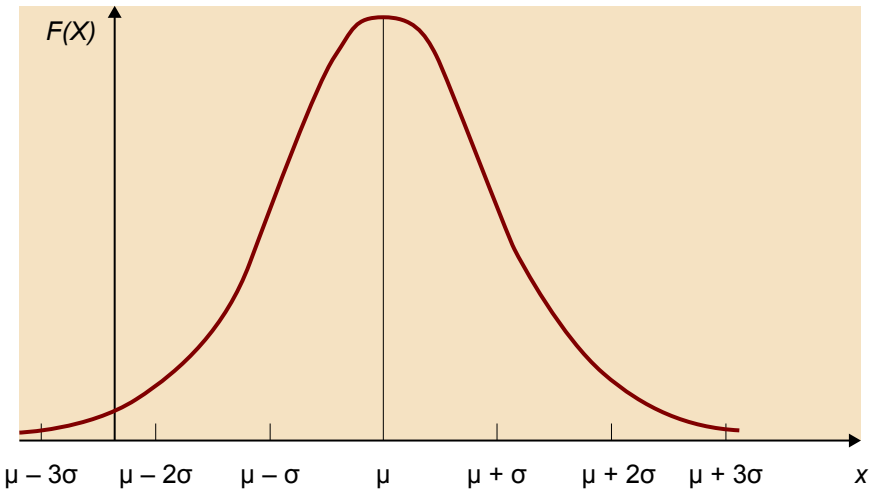
- Quina és la probabilitat que en una població de sis mil nounats neixin sis nadons, o menys, amb aquesta malformació? Utilitzant la funció «POISSON» amb els paràmetres coneguts [x: 6; mitjana (o esperança matemàtica): $E(X)$ = 0,0002 × 6.000 = 1,2; i cumulativa: 1] obtenim la probabilitat demanada: 0,999 (un 99,9%).

3.5.2. Models de probabilitat per a variables aleatòries contínues

Distribució normal o de Gauss-Laplace

La distribució normal és un model teòric de probabilitat a què s'ajusten determinades variables quantitatives contínues. Aquest model també rep el nom de **campana de Gauss**, en honor a Carl Friedrich Gauss, per la forma que presenta i que es mostra en la figura 15.

Figura 15. Campana de Gauss



Les característiques que defineixen la distribució normal són les següents:

- Els seus paràmetres són la mitjana (μ) i la variància (σ^2).

- La variable *X* segueix una distribució normal amb mitjana μ i variància σ^2.
- $X \sim N(\mu, \sigma^2)$.
- Té un punt de màxima altura que coincideix amb la mitjana (μ), la mediana (*Md*) i la moda (*Mo*).
- És simètrica respecte a l'eix d'ordenades. L'eix de simetria està situat a μ.
- És asimptòtica respecte a l'eix d'abscisses, per tant fluctua entre $-\infty$ i $+\infty$.
- Té dos punts d'inflexió situats a $\pm 1\sigma$ de μ. Entre aquests dos punts d'inflexió hi ha el 68,26% del total d'àrea sota la corba normal.

Es pot obtenir el valor de la funció de distribució normal (o de la funció de distribució cumulativa normal) per a un valor, una mitjana i una desviació estàndard especificats amb la funció «NORMDIST» de l'aplicació Google Sheets.

Es pot obtenir el valor de la funció de distribució normal inversa per a un valor, una mitjana i una desviació estàndard especificats amb la funció «NORMINV» de l'aplicació Google Sheets.

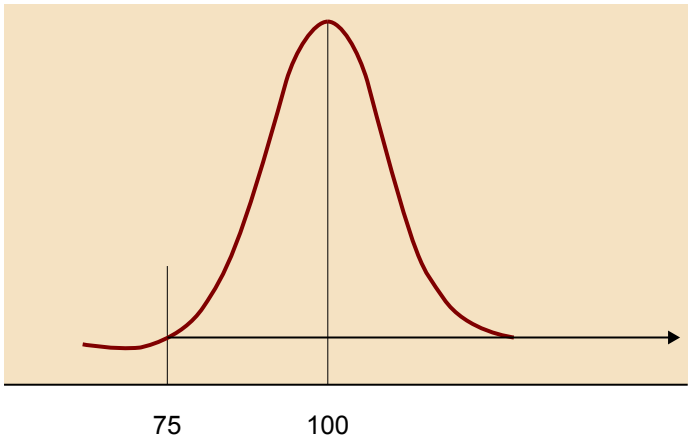
Exemple

El quocient intel·lectual (QI) es distribueix segons la llei normal amb mitjana 100 i desviació tipus 15.

Quina és la probabilitat que una persona presenti un QI superior a 75?

Utilitzant la funció «NORMDIST» amb els paràmetres coneguts (x: 75; mitjana: 100; desviació_estàndard: 15; i cumulativa: 1) obtenim que la probabilitat que una persona presenti un QI menor o igual a 75 és de 0,0477. La probabilitat que una persona presenti un QI superior a 75 és 1 menys la probabilitat que presenti un QI menor o igual a 75, és a dir, 0,9522 (95,22%).

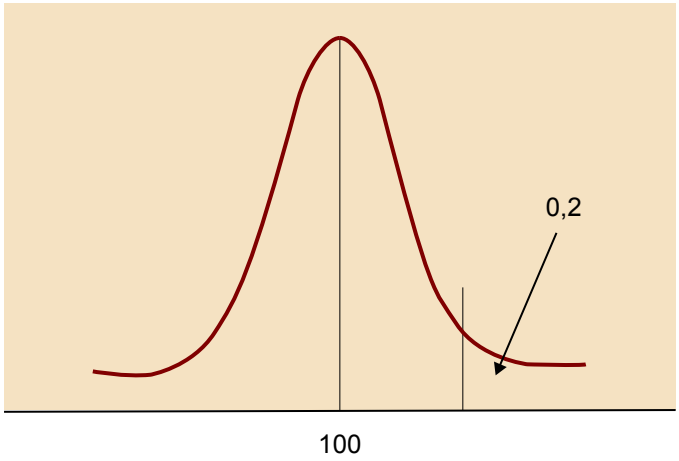
Figura 16



Quin és el QI que delimita el 20% dels subjectes més intel·ligents?

Utilitzant la funció «NORMINV» amb els paràmetres coneguts (probabilitat: 0,8; mitjana: 100; i desviació_estàndard: 15) obtenim que el QI que deixa per sobre seu el 20% dels subjectes més intel·ligents és de 112,6.

Figura 17



4. Estadística inferencial

Mentre que l'estadística descriptiva resumeix i dibuixa la informació amagada en una matriu de dades per tal d'ajudar-nos a entendre-la, l'**estadística inferencial** (o estadística inductiva) pretén aportar conclusions generals aplicables a la població d'estudi a partir de l'anàlisi de mostres diverses d'aquesta població.

Així, doncs, el principal objectiu de l'estadística inferencial és estudiar les característiques numèriques d'una població o verificar afirmacions sobre aquestes característiques a partir de calcular-les en una o diverses mostres escollides a l'atzar. El procés utilitzat en aquest tipus d'estudis ens permet inferir o pronosticar el valor dels paràmetres poblacionals (μ, σ, π, etc.) a partir del valor dels estadístics normals ($\underline{x}$, s_x, p, etc.).

Per això ens caldrà distingir entre els conceptes de **població** i **mostra**. La **població** és el conjunt d'individus, d'elements o de coses amb alguna característica comuna de què es fa un estudi estadístic. Tanmateix, quan no és possible observar tots els elements d'una població cal seleccionar-ne una **mostra** que posi de manifest les característiques estudiades i que permeti inferir dades de la població. El **mostreig** és el procediment seguit per a l'extracció de la mostra. Un **mostreig aleatori**, és a dir, amb l'elecció dels elements de la mostra a l'atzar, afavoreix una millor representativitat de la mostra i evita un possible biaix.

Entre les diferents maneres de treballar la inferència estadística que hi ha, destaquem l'**estimació de paràmetres** i el **contrast d'hipòtesi**.

4.1. Estimació de paràmetres

4.1.1. Distribució mostral d'un estadístic

La **distribució mostral d'un estadístic** (mitjana aritmètica, variància, proporció, etc.) és la distribució de l'estadístic, calculada en mostres infinites de la mateixa mida n, escollides a l'atzar, d'una determinada població. Així, doncs, si de la població d'estudiants de la UOC escollíssim mostres aleatòries de la mateixa mida (per exemple, 30), i de cada mostra calculéssim la mitjana d'edat dels subjectes, obtindríem una distribució de mitjanes d'edat que denominem **distribució mostral de la mitjana**.

La distribució mostral pot obtenir-se per a qualsevol altre estadístic dels estudiats anteriors. Així, també podem parlar de distribució mostral de la variància, distribució mostral de la proporció, distribució mostral de la mediana, etc. Totes elles s'obtindrien calculant el valor de l'estadístic corresponent en cadascuna de les infinites mostres.

Atès que el mostreig és aleatori, el valor de l'estadístic calculat en cada mostra també variarà aleatòriament de l'una a l'altra i, en conseqüència, podem considerar la distribució mostral d'aquest estadístic com la distribució d'una variable aleatòria que pot ajustar-se a un dels models de distribució de probabilitat estudiats en l'apartat anterior.

Distribució mostral de la mitjana aritmètica

L'estadístic més àmpliament utilitzat, com a representatiu d'un conjunt de dades, és la mitjana aritmètica. La distribució mostral de la mitjana té, al seu torn, la seva mitjana aritmètica (anomenada **mitjana de la distribució mostral de la mitjana**, i representada per $\mu_{\underline{X}}$), i la seva desviació estàndard (anomenada **desviació estàndard de la distribució mostral de la mitjana**, **error típic** o **error estàndard de la mitjana**, i representada per $\sigma_{\underline{X}}$).

Lògicament, en poblacions molt àmplies o infinites, el nombre de mostres diferents possibles és també pràcticament infinit (o infinit realment). La mitjana de la distribució mostral de mitjanes tendeix cap a la mitjana de la població (o hi coincideix en poblacions finites), i la desviació estàndard de la distribució (l'error estàndard de la mitjana) disminueix a mesura que augmenta la mida mostral.

El **teorema central del límit** ens diu que, encara que la distribució d'una variable no sigui normal, la distribució mostral de la mitjana basada en mostres de mida n serà aproximadament normal. Aquest teorema és més cert com més grans són les mides mostrals, així que per a n «petits» (per exemple, menys de 10), la distribució mostral de la mitjana només és aproximadament normal, mentre que per a n «grans» (per exemple, de 30), la distribució és pràcticament normal.

Distribució mostral d'una proporció

Quan treballem amb una variable categòrica, no tenim valors numèrics per a cada observació, sinó la presència o l'absència d'un atribut determinat. Així, doncs, per a la variable «sexe» dels subjectes, el que tenim per a cada subjecte és si és home o dona, igual que per a la variable «estat civil» tindrem si està casat o no. Per a aquestes variables dicotòmiques, l'estadístic més representatiu és la proporció (P), que també serà una característica de la població de referència. En aquest context parlarem de la **proporció poblacional** com un paràmetre que es representa per π.

Si escollim a l'atzar diferents observacions d'una variable categòrica i assignem a un dels seus atributs el valor 1 (habitualment, el que és centre del nostre interès), i a l'altre atribut el valor 0, **la seva distribució de probabilitat s'ajustarà a una distribució binomial**. Igual que amb la distribució mostral de la mitjana aritmètica, també és possible calcular la **mitjana** i la **desviació estàndard** de la **distribució mostral de la proporció**.

4.1.2. Intervals de confiança per a l'estimació de paràmetres

Una de les aplicacions més immediates és **l'estimació del valor d'un paràmetre poblacional** a partir de l'obtenció d'una única mostra, escollida aleatòriament, de l'esmentada població.

Quan escollim una mostra aleatòria d'observacions i utilitzem la mitjana o la proporció de la mostra per a estimar el valor poblacional, sabem que si la mostra hagués estat més àmplia, la variabilitat seria més petita i, per tant, l'estimació seria més precisa. Però com podem mesurar la precisió de les nostres estimacions?

La manera de fer-ho és no donant una única estimació del valor poblacional, sinó un interval, i després reforçar aquest interval de valors per mitjà d'una declaració del nostre nivell de confiança que el verdader valor estigui dins d'aquest interval. Això es el que es denomina **interval de confiança**.

L'interval de confiança és un rang entre dos valors al voltant d'un paràmetre mostral entre els quals, amb una probabilitat determinada (o **nivell de confiança**), se situarà aquell paràmetre en la població. El **nivell de confiança** ($1 - \alpha$) es presenta habitualment com a percentatge, en multiplicar el valor d'($1 - \alpha$) per 100, on α és el **nivell d'error** (o **nivell de significació**).

Hi ha un intercanvi entre la precisió que es pot expressar en un interval de confiança i el nivell de confiança (vegeu taula 7). Com més baix sigui el nivell de confiança, més petit serà l'interval de confiança. Per tant, el resultat serà més precís, però la probabilitat que l'interval no inclogui el vertader valor del paràmetre serà més elevada.

Taula 7

Interval de confiança	Nivell de confiança ($1 - \alpha$)	Nivell d'error o de significació (α)	Precisió
↓	↓	↑	↑
↑	↑	↓	↓

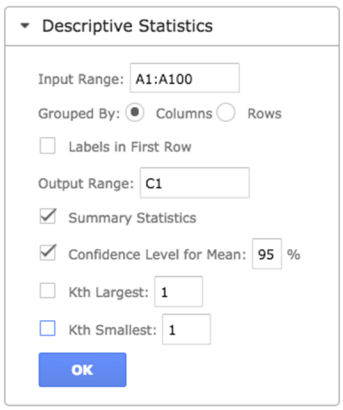
L'única manera de millorar tant la precisió com el nivell de confiança és reduint l'error típic. Si la desviació estàndard poblacional és fixa, únicament podrem reduir l'error típic augmentant la mida de la mostra. Alternativament, si es manté el marge d'error fix, incrementar la mida de la mostra comporta un increment del nivell de confiança.

Interval de confiança per a la mitjana aritmètica

Es pot obtenir un interval de confiança de la mitjana aritmètica amb el complement «XLMiner Analysis ToolPak» de l'aplicació Google Sheets entrant al menú del quadre de diàleg «Descriptive Statistics».

Quan s'inicia el complement, s'obre un quadre de diàleg en el qual cal seleccionar l'opció «Descriptive Statistics» (vegeu figura 18).

Figura 18



- «Input range»: indicar les caselles on s'ubiquen els valors en la matriu de dades.
- «Labels in First Row»: activar si tenim etiquetada la variable en la primera fila.
- «Output Range»: indicar la casella a partir de la qual volem que se'ns mostrin els resultats.
- «Summary Statistics»: activar per a obtenir els resultats descriptius.
- «Confidence Level for Mean»: activar per a obtenir l'interval de confiança. Quan s'activa aquesta opció, s'estableix, per defecte, un nivell de confiança del 95%, però pot substituir-se aquest valor per qualsevol altre.

Una vegada executada aquesta opció, fent clic a «OK» s'obtenen els següents resultats (l'exemple correspon a les dades obtingudes passant el MAS[2] a cent subjectes d'un municipi):

(2)Test d'ansietat manifesta de Taylor (Taylor, 1958).

Taula 8. Anàlisi resultats (MAS)

Mean	22,060
Standard Error	1,288
Median	21,000
Mode	31,000
Standard Deviation	12,878
Sample Variance	165,855
Kurtosis	–0,855
Skewness	0,135
Range	48,000
Minimum	0
Maximum	48,000
Sum	2.206,000
Count	100
Confidence Level (95%)	2,555

La majoria d'informació que ens proporciona aquesta anàlisi correspon a la descripció de la variable. Per al nostre propòsit en aquest apartat, la informació que ens interessa és la que hi ha a les caselles de **mitjana** («Mean») i **nivell de confiança** («Confidence Level»).

Ja coneixem àmpliament la interpretació de la mitjana, però desconeixíem fins ara la interpretació del nivell de confiança. El valor que ens proporciona és el del **marge d'error de l'interval de confiança per al nivell de confiança escollit**. Així, doncs, en el nostre cas, per al nivell de confiança del 95% aquest marge d'error és de 2,555.

Amb aquests resultats, podem concloure que, amb un nivell de confiança del 95%, el grau d'ansietat mitjà de tots els habitants del municipi estarà entre 19,505 i 24,615 punts de l'escala del MAS.

Interval de confiança per a la proporció

Ja hem vist anteriorment que per a variables dicotòmiques es pot considerar la proporció d'una de les seves dues modalitats com la mitjana del conjunt de valors prèviament codificats com 0 i 1, assignant l'1 a la modalitat la proporció de la qual volem estudiar.

Es pot obtenir un interval de confiança d'una proporció amb el complement «XLMiner Analysis ToolPak» de l'aplicació Google Sheets entrant al menú del quadre de diàleg «Descriptive Statistics».

L'exemple que presentem correspon a l'estudi de la proporció d'homes en un municipi. Per tant, el primer que farem en la nostra matriu de dades serà aïllar la variable «sexe» i recodificar-ne els valors, assignant un 1 als homes i un 0 a les dones. Quan iniciem el complement «XLMiner Analysis ToolPak» de l'aplicació Google Sheets, s'obre el mateix quadre de diàleg que en l'exemple anterior. Una vegada executada aquesta opció, fent clic a «OK» s'obtenen els següents resultats:

Taula 9. Anàlisi resultats (sexe masculí)

Mean	0,400
Standard Error	0,049
Median	0,000
Mode	0,000
Standard Deviation	0,492
Sample Variance	0,242
Kurtosis	–1,866
Skewness	0,414
Range	1,000
Minimum	0,000
Maximum	1,000
Sum	40,000
Count	100,000
Confidence Level (95%)	0,098

La major part d'informació que ens proporciona aquesta anàlisi no és pertinent perquè estem analitzant una variable categòrica com és el sexe, però per al nostre objectiu sí que té la informació necessària. Així, **la mitjana de la distribució** (0,40) és la **proporció** (*P*) d'homes de la mostra. També coincideix l'error típic de la proporció amb l'error típic de la mostra.

Haurem de **recalcular el marge d'error** multiplicant el valor de l'error típic pel de la puntuació *z* corresponent al nivell de confiança establert. En el nostre exemple, el marge d'error exacte serà 0,049 × 1,96 = 0,096.

Amb aquest resultat podem concloure que, amb un nivell de confiança del 95%, la proporció d'homes del municipi estudiat està entre 0,304 i 0,496, és a dir, un percentatge d'homes entre el 30,4% i el 49,6%.

4.2. Introducció al contrast d'hipòtesi

El **contrast d'hipòtesi**, també anomenat **prova de significació** o **prova estadística**, és un procediment que ens permet decidir si una afirmació sobre certa característica o característiques de la població pot ser mantinguda o ha de ser rebutjada, d'acord amb les dades obtingudes en una mostra de la població o en diverses mostres.

Amb aquesta breu introducció al contrast d'hipòtesi s'estableixen les bases per a un ampli conjunt de proves estadístiques que estan fora de l'abast d'aquesta assignatura; totes, però, parteixen dels mateixos postulats i segueixen un mateix esquema per a la seva resolució.

Una de les aplicacions més habituals dels contrastos d'hipòtesi és quan es vol comprovar l'efecte d'una determinada intervenció o tractament. Per exemple, podríem plantejar un estudi de com ha influït la nova llei del tabac en el consum de cigarretes al municipi. Podríem comparar la proporció de fumadors abans i després de la promulgació de l'esmentada llei. L'afirmació que posaríem a prova seria la següent: la nova llei del tabac ha disminuït la proporció de fumadors. El contrast d'hipòtesi ens permet prendre una decisió sobre si acceptem o rebutgem l'afirmació anterior (que anomenarem **hipòtesi**), d'acord amb les dades que hem obtingut d'una mostra de 100 subjectes.

4.2.1. Contrast d'hipòtesi: prendre decisions

En l'apartat anterior hem vist les distribucions mostrals de la mitjana i com aquestes distribucions ens permeten definir un interval en el qual confiem que hi haurà la mitjana de la població.

Un exemple que permet il·lustrar els conceptes implicats en la presa de decisions estadístiques podria ser el de les proves d'un laboratori mèdic amb les quals intentem detectar el virus de la sida. Imaginem que s'envia una mostra de sang a un laboratori perquè faci la prova d'anticossos VIH. Tenim dues possibilitats que ens interessen: que els anticossos siguin presents a la sang o que no hi siguin i, en realitat, només una de les possibilitats és certa.

En la taula 10 representem aquestes dues possibilitats per a la situació real en forma de dues files de la taula. Quan s'aplica la prova de laboratori determinada a la mostra de sang, s'arriba a una certa conclusió: la prova és positiva (el virus és present a la sang) o negativa (el virus és absent). Ambdues possibilitats es representen en les dues columnes de la taula 10.

Taula 10

	Prova	
Veritat	Negativa	Positiva
Absència del virus	Correcta	Errònia
Presència del virus	Errònia	Correcta

Les files indiquen el veritable estat de la mostra de sang, mentre que les columnes indiquen la conclusió que el laboratori n'ha tret, que podria ser errònia per moltes raons, per exemple perquè el procediment de laboratori sigui incorrecte o perquè hi hagi manca de detectabilitat del virus. La taula mostra les quatre possibilitats diferents, que depenen de la conclusió a què s'ha arribat i de què és la veritat:

- 1a. fila, 1a. columna: la prova és negativa i la veritat és que no hi ha presència del virus de la sida; és una conclusió correcta.

- 1a. fila, 2a. columna: la prova és positiva, però la veritat és que no hi ha presència del virus de la sida; és una conclusió falsa i els investigadors mèdics sovint parlen d'un positiu fals.

- 2a. fila, 1a. columna: la prova és negativa, però la veritat és que hi ha el virus i la prova ha fracassat a l'hora de detectar-lo; aquesta és una conclusió falsa i s'anomena negatiu fals.

- 2a. fila, 2a. columna: la prova és positiva i veritablement hi ha presència del virus; aquesta és una conclusió correcta.

En el contrast d'hipòtesi tenim la mateixa situació. Nosaltres considerem dos possibles estats de la població, que anomenem **hipòtesis**. A partir de les dades d'una mostra, cal decidir quina de les hipòtesis és la correcta. La nostra decisió també pot ser correcta de dues maneres quan decidim a favor de la hipòtesi que és veritablement correcta, i pot ser equivocada de dues maneres quan decidim a favor de la hipòtesi falsa.

4.2.2. Hipòtesi nul·la i alternativa

La **hipòtesi nul·la**, representada per H_0, és l'expressió formal que es posa a prova en un contrast d'hipòtesi. Indica la «no diferència», o el «sense efecte», i és la que suposem a l'hora de valorar si el resultat es deu a l'atzar. H_0 expressa, per exemple, que un paràmetre de la població, com pot ser la mitjana, pren un valor específic, o que aquesta mitjana és igual en dos grups diferents de subjectes.

La **hipòtesi alternativa**, representada per H_1, és l'expressió de l'efecte, el canvi o la diferència que hi pot haver en les dades estudiades (i que moltes vegades, encara que no en totes, és la que esperem o sospitem). La hipòtesi alternativa diu, per exemple, que un paràmetre de la població, com ara la mitjana, difereix d'un valor especificat, o que el mateix paràmetre obtingut en dos grups diferents difereix en el seu valor, en una direcció determinada (unilateral o d'una cua) o en les dues direccions (bilateral o de dues cues).

Fixeu-vos en una altra característica que diferencia la hipòtesi nul·la de l'alternativa:

- La hipòtesi nul·la, habitualment (però no sempre), consisteix en una igualtat simple entre paràmetres o entre un paràmetre i un valor fix; en aquest cas, la igualtat a la mitjana del grau d'ansietat.

- La hipòtesi alternativa consisteix, normalment, en moltes possibilitats; en aquest cas, que la mitjana d'ansietat dels habitants del municipi sigui diferent de la de la població general.

4.2.3. Ús dels intervals de confiança per a dur a terme un contrast d'hipòtesi

Seguint el nostre exemple anterior, podríem analitzar si el grau d'ansietat (mesurat amb el MAS) dels habitants del municipi és igual al de la població general (hipòtesi nul·la), o bé és diferent (hipòtesi alternativa).

Si definim la mitjana en ansietat de la població general com $\mu = 25$, les expressions formals de la hipòtesi nul·la i l'alternativa serien les següents:

- Hipòtesi nul·la. H_0: $\mu = 25$
- Hipòtesi alternativa. H_1: $\mu \neq 25$

Una vegada més, podem representar les diferents conclusions i les situacions reals en una taula:

Taula 11

	Conclusions a partir del nostre estudi	
Veritat	Igual grau mitjà d'ansietat	Diferent grau mitjà d'ansietat
Igual grau mitjà d'ansietat	Correcta	Errònia
Diferent grau mitjà d'ansietat	Errònia	Correcta

Hem calculat un **interval de confiança** per a la mitjana en ansietat dels subjectes del municipi a partir de la mostra dels 100 subjectes de què disposem. Aquest interval, amb un grau de confiança del 95%, està entre 19,505 i 24,615.

Per tant, confiem en un percentatge del 95% que la veritable mitjana de la població de referència (els habitants del municipi) està entre aquests límits, i veiem que aquest interval no conté el valor 25. Atès que volem comprovar que la mitjana de la població és 25, podem dir amb molta seguretat que no és de 25. A partir dels càlculs fets per a establir els intervals de confiança, diem que rebutgem la hipòtesi nul·la, on $\mu = 25$ i, per tant, acceptem la hipòtesi alternativa, on $\mu \neq 25$. Interpretem aquestes dades, dins del context del nostre exemple general, de manera que el grau mitjà d'ansietat dels habitants del municipi no és igual al grau mitjà de la població general.

4.2.4. Contrast d'hipòtesi i proves de significació

Un **estadístic** de contrast és un instrument estadístic creat per a prendre decisions sobre la hipòtesi nul·la amb certa probabilitat. Es caracteritza per tenir una distribució mostral coneguda (normal, *t* de Student, X^2, etc.). Per a cada tipus de contrast (d'una mitjana, de dues mitjanes, de dues proporcions, etc.), tenim el seu estadístic de contrast corresponent.

Els passos que cal seguir són els següents:

1) Plantejar la hipòtesi nul·la i l'alternativa.

2) Obtenir, a partir de les dades mostrals, l'estadístic de contrast corresponent.

3) Obtenir les regions d'acceptació i de rebuig de la hipòtesi nul·la a partir del valor de l'estadístic de contrast teòric.

A partir de la distribució corresponent de l'estadístic de contrast (normal, *t* de Student, X^2, etc.), i especificant el nivell de confiança assumit (o més habitualment, el seu complementari, que és el nivell de significació α), s'obtenen els dos valors d'aquest estadístic de contrast, que inclouen el percentatge corresponent al nivell de confiança (95%). Aquests dos valors es denominen **valors crítics (superior i inferior) de l'estadístic de contrast**, i la regió compresa entre aquests dos valors es denomina **regió d'acceptació de la hipòtesi nul·la**. La **regió de rebuig de la hipòtesi nul·la** se situa per sobre del valor crític superior i per sota del valor crític inferior.

4) Prendre la decisió d'acceptar o de rebutjar la hipòtesi nul·la.

Prenem aquesta decisió comparant l'estadístic de contrast, calculat amb les dades mostrals, amb els valors crítics de la distribució corresponent. Així, si el nostre estadístic de contrast calculat queda entre aquests valors crítics, acceptem la hipòtesi nul·la (ja que som en la regió d'acceptació de l'esmentada

hipòtesi), mentre que si el nostre estadístic de contrast calculat és més gran que el valor crític superior o menor que l'inferior, rebutgem la hipòtesi nul·la i acceptem l'alternativa.

5) Interpretem el resultat en el context de l'estudi realitzat.

Les dues proves anteriors són equivalents i, lògicament, ens porten al mateix resultat.

El contrast d'hipòtesi, tal com l'hem vist fins ara, pot tenir una variant, que és la utilitzada habitualment en els paquets estadístics informatitzats: la **prova de significació**, que consisteix a obtenir directament la probabilitat de l'estadístic de contrast mostral calculat. Aquesta probabilitat es denomina valor *p*. El **valor *p*** és la probabilitat d'observar el resultat, o un resultat més extrem, quan la hipòtesi nul·la és certa. Com més petit és el valor *p*, més accentuada és la prova contra la H_0 proporcionada per les dades. Els valors *p* per sota del nivell de significació (habitualment de 0,05) s'anomenen convencionalment **significatius**.

El **nivell de significació** és la probabilitat màxima de cometre un error de tipus I (vegeu més endavant). S'estableix en funció del risc que s'està disposat a assumir abans de reunir i analitzar les dades. Si el valor *p* del contrast és menor que α, la hipòtesi nul·la es rebutja i diem que el resultat observat és estadísticament significatiu al nivell de α.

Per a prendre una decisió respecte a la hipòtesi nul·la, simplement comparem aquest valor *p* amb el nivell de significació (α). Si el valor *p* és superior a α acceptem la hipòtesi nul·la, i si és més petit, la rebutgem i acceptem l'alternativa (llavors diem que el resultat és estadísticament significatiu).

Podem esquematitzar els passos per al contrast d'hipòtesi o la prova de significació en la taula 12.

Taula 12

Contrast d'hipòtesi	**Prova de significació**
Determinar la hipòtesi nul·la i l'alternativa	
Calcular el valor de l'estadístic de contrast amb les dades mostrals	
Determinar les regions d'acceptació i de rebuig de la hipòtesi nul·la (amb els valors crítics de l'estadístic de contrast)	Obtenir la probabilitat (valor *p*) de l'estadístic de contrast
Prendre una decisió comparant el valor de l'estadístic de contrast mostral o l'observat amb els valors crítics de la distribució corresponent	Prendre una decisió comparant el valor *p* de l'estadístic de contrast mostral o l'observat amb el nivell de significació assumit
Interpretar la decisió anterior en el context de l'estudi realitzat	

4.2.5. Errors de tipus I i de tipus II

Tal com hem vist en apartats anteriors, sempre que prenem una decisió en un contrast d'hipòtesi o en una prova de significació podem haver encertat o podem haver-nos equivocat, ja que sempre hi ha una probabilitat que la hipòtesi nul·la sigui certa, encara que l'hàgim rebutjat (aquesta probabilitat és el valor *p*), o que no sigui certa encara que l'hàgim acceptat, d'acord amb les nostres dades mostrals.

Podem representar en una taula la situació en el contrast d'hipòtesis estadístiques (taula 13).

Taula 13

	Situació certa	
Decisions basades en les dades	H_0	H_1
H_0	Decisió correcta Probabilitat $1 - \alpha$	Decisió incorrecta Error de tipus II Probabilitat β
H_1	Decisió incorrecta Error de tipus I Probabilitat α	Decisió correcta Probabilitat $1 - \beta$

En aquesta taula hem presentat dos termes que s'utilitzen per a les decisions incorrectes que es poden prendre en aquesta situació:

- **Error de tipus I**: és el que es comet en pronunciar-nos a favor de la hipòtesi alternativa (és a dir, en rebutjar la hipòtesi nul·la) quan de fet la hipòtesi certa és la nul·la. La probabilitat d'aquest error de tipus I és igual a α (nivell de significació), o al valor *p*.

- **Error de tipus II**: és el que es comet en pronunciar-nos a favor de la hipòtesi nul·la quan la hipòtesi alternativa és la certa. La probabilitat de cometre un error de tipus II es representa per β i inicialment és desconeguda.

4.2.6. Potència d'un contrast d'hipòtesi o prova de significació

Com hem vist en l'apartat anterior, la probabilitat de cometre un error de tipus II es denomina β, i en principi l'investigador la desconeix. El complementari d'aquest valor β, és a dir, $1 - \beta$, és l'anomenada **potència de la prova estadística**, i és la probabilitat de no equivocar-nos quan rebutgem una hipòtesi nul·la i acceptem, per tant, l'alternativa. Dit d'una altra manera, és la seguretat

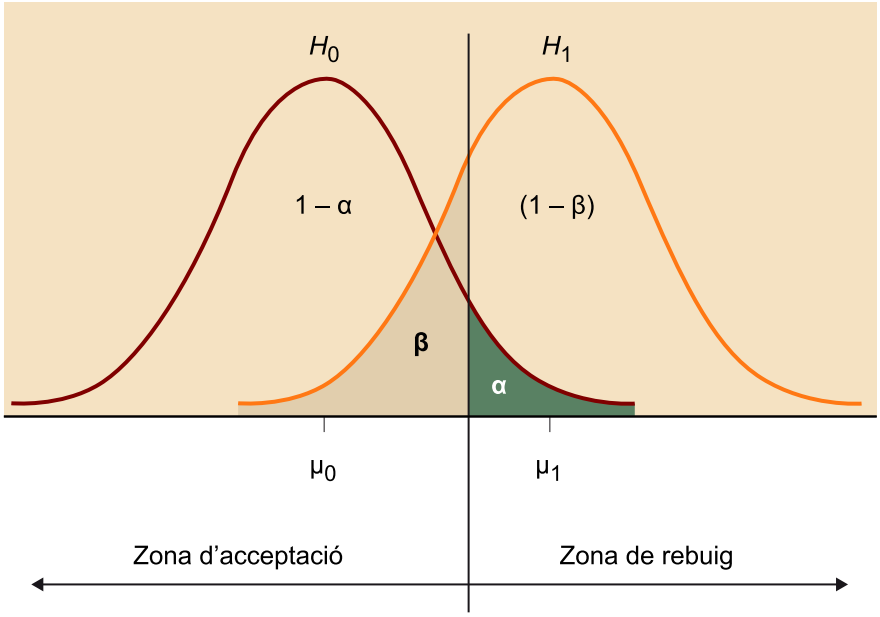
que tenim de no equivocar-nos en acceptar una hipòtesi alternativa (que força vegades representa la hipòtesi de l'efectivitat d'una intervenció determinada perquè expressa la diferència entre dos o més grups o mostres de dades).

Com que el valor de β és desconegut inicialment, també ho és el valor de la potència de la prova (1 – β), encara que sí que sabem la relació que té amb el grau de significació i amb la mida de la mostra, per exemple.

La relació entre α i β, és a dir, entre la probabilitat de cometre un error de tipus I i un error de tipus II, és un altre dels intercanvis característics en estadística, perquè aquesta relació és inversa. Així, si volem disminuir la probabilitat de cometre un error de tipus I (disminuint α), estem augmentant la probabilitat de cometre un error de tipus II (augmentant β), i disminuïm en conseqüència la potència de la prova estadística. Aquest intercanvi entre α i β es pot apreciar millor en la figura 19.

Com es pot observar en la figura 19, la hipòtesi alternativa, quan és certa, també té la seva distribució de densitat (com la hipòtesi nul·la), i aquestes dues distribucions s'encavalquen (en aquest cas pel costat dret de la hipòtesi nul·la, perquè la prova és unilateral per la cua dreta). Així, el valor crític de l'estadístic de contrast és la línia vertical que divideix la gràfica en dos. Per sota (o a l'esquerra) d'aquest valor hi ha la regió (o zona) d'acceptació de la hipòtesi nul·la, i per sobre (o per la dreta), la regió de rebuig de l'esmentada hipòtesi nul·la.

Figura 19



Això determina dues àrees ratllades: una verticalment, que és la proporció de la distribució de la H_0 per sobre del valor crític de l'estadístic de contrast, i que correspon al grau de significació α o al valor *p*; i una horitzontalment, que és la proporció de la H_1 per sota d'aquest valor crític i que denominem β. Si disminuïm la zona ratllada verticalment (és a dir, el grau de significació

α), desplacem la ratlla vertical cap a la dreta, i això comporta que augmenti la zona ratllada horitzontalment, és a dir β, amb la qual cosa disminueix, en conseqüència, la regió 1 – β, que denominem potència de la prova.

L'única manera de disminuir tant la probabilitat de cometre un error de tipus I com de tipus II i augmentar la potència de la prova estadística és, una vegada més, **augmentant les mides mostrals**. Així, doncs, augmentar el nombre de subjectes de les mostres és l'única manera que tenim de disminuir les probabilitats de cometre un error (sigui del tipus I o II) en una prova estadística de significació.

Bibliografia

Adielsson, M.; Barnes, R.; Kupfer, P.; Roberts, I.; Weber, J. H. (2005). «Google Sheets function list». <https://support.google.com/docs/table/25273?hl=en>

Cosculluela, A.; Fornieles, A.; Turbany, J. (2014). *Tècniques d'anàlisi de dades quantitatives*. Material docent de la UOC. Barcelona: Universitat Oberta de Catalunya.