Bioestadística

Graus en:

. Ciència i Tecnologia dels Aliments

. Nutrició Humana i Dietètica

Gener 2019

Dr. Manel Viader Junyent

Secció de Psicologia Quantitativa

Departament de Psicologia Social i Psicologia Quantitativa

Facultat de Psicologia

Universitat de Barcelona

Bioestadística

1.- Conceptes bàsics i panoràmica general

L'estadística és una part de les matemàtiques que es refereix a la recollida, representació, anàlisi i interpretació de dades referents a diversos tipus de variables. Naturalment, el seu objectiu és obtenir informació sobre el comportament d'aquestes variables, les relacions que es puguin establir entre elles i altres aspectes.

La Bioestadística és l'aplicació de les eines pròpies de l'estadística a problemes relacionats amb les ciències de la vida i de la salut.

Un concepte important de l'estadística és el de variable aleatòria, entesa com a qualsevol variable a la qual es pot associar una distribució de probabilitat. L'exemple clàssic és el del llançament repetit d'una moneda i el registre del corresponent resultat, ja sigui cara o creu. Després de repetir el llançament un cert número de vegades podem arribar a concloure que, per exemple, un 53% de les vegades surt cara i un 47% surt creu. Les probabilitats respectives són, doncs, de 0,53 i de 0,47, i això és senzillament la distribució de probabilitat del succés aleatori que anomenem "resultat del llançament d'una moneda".

Cal remarcar que el fet de parlar de variable aleatòria no significa que el fenomen no tingui les seves causes i pugui tenir una explicació determinista. Potser si coneguéssim perfectament tots els factors implicats (posició inicial de la moneda, repartiment de pes entre les seves parts, força i direcció del llançament, resistència de l'aire, etc.) podríem tenir informació segura sobre el resultat. Però en no ser així, estem obligats a un enfocament probabilístic. A la naturalesa poden existir, però, fenòmens realment o irreductiblement aleatoris (com sembla que poden ser alguns dels relacionats amb la física quàntica). En tot cas, tant en un cas com en l'altre l'abordatge probabilístic és necessari.

En el cas del llançament de la moneda, és evident que per tenir una bona aproximació a la distribució de probabilitat necessitem tenir un número important de casos. Llançar una moneda 5 vegades i que surtin 4 cares no significa res. Llançar-la 50.000 vegades i que surtin 40.000 cares ens diria de forma inequívoca que la moneda no està equilibrada. El lògic és pensar que amb una moneda equilibrada les probabilitats d'obtenir cara i creu tendiran a igualar-se a mesura que fem més llançaments, per arribar a un valor límit de 0,5 tant per un resultat com per l'altre.

Aquesta lògica es pot traslladar fàcilment a altres situacions. Imaginem que volem esbrinar quin és el contingut de sal d'un cert producte alimentari. Suposem també que la factoria que prepara aquest producte en produeix 5.000 unitats diàries, per tant més de 100.000 unitats mensuals. Si volem esbrinar si el contingut de sal és el previst i per tant assegurar-nos que no hi ha cap error de producció, el que difícilment podrem fer és comprovar més de 100.000 productes. Necessitarem agafar a l'atzar un cert número d'aquests productes i analitzar-los, i podem suposar que el resultat obtingut en aquests casos serà representatiu del total, sempre que el número de casos estudiat sigui prou gran. Observem aquí alguns conceptes fonamentals: En primer lloc la idea de *població*, que es refereix a tot el conjunt de casos que es produeixen (en

aquest cas, si agafem un període mensual com a unitat d'anàlisi, el total de 100.000 productes que es fabriquen en un mes determinat). Davant de la impossibilitat d'estudiar tots els productes fabricats, extraiem d'aquesta població una *mostra* de casos, és a dir, un cert número de productes, que seran els que realment estudiarem. Per poder extrapolar els resultats de la mostra a tota la població caldrà que aquesta mostra sigui *representativa*, i per aconseguir això necessitem que la *grandària de la mostra* sigui suficient i que la selecció dels productes s'hagi fet amb una *tècnica de mostratge* correcta, normalment de caràcter aleatori. Per exemple, si volem agafar una mostra de 300 productes, seria un error recollir-los tots el mateix dia, caldrà repartir l'obtenció de productes al llarg de tot el mes, i la selecció de quines unitats s'agafen realment es pot fer a l'atzar.

A partir d'aquest exemple es pot distingir fàcilment entre dues branques molt importants de l'estadística: Per una part, l'*estadística descriptiva*, que té com a objectiu classificar, visualitzar o resumir les dades obtingudes sobre un determinat fenomen; per una altra, l'*estadística inferencial*, pensada per intentar extrapolar al conjunt de la població els resultats obtinguts a partir d'una mostra.

L'estadística descriptiva treballa principalment amb eines com el recompte de resultats, la seva representació tabular o gràfica, i el càlcul de certes quantitats que poden informar-nos de les característiques dels resultats obtinguts. Aquestes quantitats, que es refereixen a característiques de la mostra, s'anomenen *estadístics*. Exemples molt coneguts són la *mitjana aritmètica* o la *variància* dels resultats, entre d'altres.

L'estadística inferencial, com ja s'ha indicat, pretén esbrinar, a partir dels resultats d'una mostra, les característiques del conjunt de la població. Si, per exemple, agafem una mostra de 300 productes i el contingut mitjà de sal és de 1,75 grams (per cada 100 grams de producte), podem afirmar que la mitjana de sal en el conjunt de la població (els 100.000 productes fabricats durant aquell mes) és també de 1,75?. Les característiques de la població (mitjana, variància, etc.) s'anomenen *paràmetres*. Per tant, a l'estadística inferencial ens trobarem amb una sèrie d'estadístics (característiques de la mostra) que ens poden servir per *inferir* els corresponents paràmetres, és a dir, les corresponents característiques de la població. És important assenyalar que l'estimació dels paràmetres té un caràcter probabilístic. És a dir, si la mitjana de sal a la mostra és de 1,75 grs., no podrem mai afirmar amb seguretat que la mitjana de sal del conjunt de productes fabricats sigui exactament de 1,75, però sí que podrem establir el que s'anomena un *interval de confiança*. A partir dels càlculs corresponents podríem arribar a dir, per exemple, que amb una probabilitat de 0,95 (o del 95%) la mitjana de sal per als 100.000 productes estarà entre 1,63 i 1,87. L'única manera de saber exactament el valor del paràmetre poblacional seria inspeccionant els 100.000 productes, cosa molt difícil o impossible.

Les tècniques estadístiques ens permeten respondre també altres tipus de preguntes. Seguint amb l'exemple anterior, però amb un objectiu diferent, suposem que agafem una mostra de 300 productes el dia 4 i una altra mostra de 300 productes el dia 14 del mateix mes. Imaginem que el contingut mitjà de sal de la primera mostra és de 1,79 i el de la segona és de 1,68. Aquesta diferència és només producte de l'atzar, o bé ens indica que hi ha algun element diferencial (matèria primera, funcionament del procés productiu, error humà, etc.) que l'explica?. És fàcil pensar en altres exemples: Hi ha diferència en l'efectivitat de dues dietes per perdre pes?; hi ha diferència en la distribució de grups sanguinis entre la població europea i l'americana?; hi ha relació entre el gènere i el contingut de colesterol en sang?, etc, etc. L'estadística inferencial es preocupa també d'aquest tipus de qüestions, com es veurà més endavant.

L'ús de les tècniques estadístiques implica treballar amb conjunts de dades que de vegades poden ser bastant grans, i també la realització de càlculs que en alguns casos poden arribar a ser relativament complexos. Per això és molt important el recurs a aplicacions informàtiques que serveixin de suport a l'usuari. En aquest sentit hi ha moltes possibilitats, algunes relacionades amb software propietari i altres de caràcter lliure. Tant per a la realització d'aquest text com per als exercicis complementaris de l'assignatura s'ha fet ús de tres aplicacions concretes:

. Excel: Com és conegut, forma part de la suite Microsoft Office i, per tant, té caràcter de software propietari, tot i que la seva presència és molt àmplia. Funciona des de la lògica de full de càlcul, però permet fer la majoria d'operacions relacionades amb l'estadística descriptiva i una part de les pròpies de l'estadística inferencial, encara que amb limitacions. Pot ampliar funcionalitats amb alguns complements.

R i Rcommander: R és un software estadístic de caràcter lliure que permet realitzar la gran majoria de les operacions estadístiques. Està disponible per a tots els usuaris de forma completament gratuïta. El seu inconvenient principal és el seu caràcter "poc amigable", ja que funciona a través de la introducció de comandos, cosa que implica conèixer la sintaxi corresponent. Per evitar aquest problema s'han desenvolupat interfícies d'usuari, la més coneguda de les quals és Rcommander, que faciliten enormement l'ús de l'aplicació i eviten la utilització de codi.

SPSS: Probablement és l'eina estadística més utilitzada a l'entorn acadèmic. És molt potent, relativament fàcil d'utilitzar, i permet treballar amb bases de dades de molta grandària. S'actualitza de forma freqüent. És propietat d'IBM i només es pot utilitzar a través de la llicència pertinent. A la Universitat de Barcelona pot ser utilitzada a través de la corresponent llicència de Campus.

Naturalment, hi ha altres aplicacions de caràcter estadístic. Per exemple, a l'àmbit dels fulls de càlcul també es poden realitzar operacions estadístiques en fulls de Google Drive o de LibreOffice (antic OpenOffice), però normalment amb més limitacions que en el cas d'Excel. Altres aplicacions de caràcter propietari són MiniTab, SAS, S-Plus (versió de pagament de R), StatGraphics, etc.

2.- Estadística descriptiva: Tipus de variables i eines principals

El tractament estadístic que es pot fer amb les dades obtingudes sobre una certa variable depèn de forma crucial de les seves característiques i de la forma de mesura que s'hagi utilitzat. És important distingir entre els següents tipus de variables:

. *Variables qualitatives o categòriques*: Les característiques de l'objecte o fenomen es representen mitjançant atributs o categories representatives

Exemples: Grup sanguini, lloc de naixement, tipus d'aliments, etc.

Cal fer un distinció suplementària, en funció de les característiques de mesura de la variable. Moltes variables qualitatives no admeten cap tipus d'ordenació. Si pensem, per exemple, en diferents categories de productes alimentaris (làctics, verdures, pa i derivats, etc.), no té sentit cap mena d'ordenació entre elles, en tot cas qualsevol ordre serà arbitrari. Una altra cosa seria que s'introduís un criteri suplementari (per exemple, el seu contingut de glucosa), ja que en

llavors sí que podríem ordenar les diferents categories. En aquest cas, però, ja no estaríem parlant simplement de "tipus d'aliments", sinó de "contingut de glucosa de diferents tipus d'aliments", una variable diferent. La variable original (tipus d'aliments) no pot, doncs, ser ordenada, o pot ser-ho només per criteris discrecionals sense cap significat concret (per exemple, ordenant les categories per ordre alfabètic).

A diferència del cas anterior, existeixen variables qualitatives que sí poden ser ordenades amb criteris significatius. Suposem que, a partir de diferents criteris diagnòstics, classifiquem les persones afectades per desnutrició en diferents nivells de gravetat del problema: Podem parlar de desnutrició lleu, moderada, greu o molt greu, en funció de la mesura de diferents característiques de les persones. Encara que la informació inicial pugui ser en part quantitativa (per exemple, el pes de la persona o el seu índex de massa corporal), si a partir d'aquesta informació i d'altres criteris el que es fa és categoritzar els casos (lleu, moderat, etc.) la variable resultant és categòrica i, per tant, ha de ser tractada com a variable qualitativa. Ara bé, resulta evident que en aquest cas sí que podem fer una ordenació de les diferents categories pel que a la gravetat del problema de desnutrició: Lleu < Moderada < Greu < Molt greu .

Quan no hi ha ordenació possible de les categories, la variable es sol anomenar *nominal*, mentre que si hi ha possibilitat de fer una ordenació dels seus valors es parla de variable *ordinal*.

. *Variables quantitatives*: En aquestes variables, les característiques que s'estudien s'expressen de forma numèrica

Exemples: Pes, temperatura d'ebullició de diferents líquids, contingut de glucosa d'un aliment, edat, etc.

També es pot fer una distinció rellevant dins de les variables quantitatives. Hi ha variables quantitatives *discretes*, en les quals entre dos valors qualssevol de la variable només hi haurà un número finit d'altres valors (per exemple, l'edat mesurada en anys, o el número de germans). Una altra forma de dir-ho és indicant que entre dos valors consecutius de la variable no n'hi haurà cap altre. Per exemple, una persona té dos germans o en té tres, sense possibilitat de valors intermedis. En oposició a això, hi ha variables quantitatives *contínues*, on entre dos valors qualssevol de la variable n'hi ha infinits més (per exemple, la temperatura, la longitud o la velocitat). En el cas de les variables contínues no té sentit parlar de valors consecutius, perquè entre dos valors qualssevol de la variable, per propers que siguin, sempre n'hi haurà infinits més, encara que òbviament no podrem mesurar la variable amb precisió infinita. Per exemple, entre una temperatura de 35º i una de 36º podem trobar un valor de 35,2. I entre 35,2 i 35,3 podem trobar un valor com 35,26, per exemple. I entre 35,26 i 35,27 podem trobar 35,262, i així successivament.

Cal insistir en que una variable quantitativa pot ser tractada com a qualitativa, però no al contrari. L'índex de massa corporal (IMC) és una variable quantitativa, ja que s'obté dividint en el pes en quilos entre el quadrat de l'alçada, però sovint s'utilitza agrupant els resultats en categories. Un criteri habitual és parlar d'IMC inferior al normal (menor de 18,5), normal (entre 18,5 i 24,99), sobrepès (entre 25 i 29,99) i obesitat (30 o més). El resultat d'això és una variable "qualitativa" de caràcter ordinal. Es pot dir que l'obesitat implica més IMC que el sobrepès, però no que l'obesitat és "el doble" del sobreprès, ni establir cap altra relació quantitativa. La decisió sobre com tractar la variable (com a quantitativa o com a qualitativa) dependrà dels interessos i els objectius de les persones que facin l'estudi.

Com ja s'ha indicat, l'operació inversa no es pot realitzar. Per exemple, l'investigador pot estar interessat en diferents tipus d'aliments, i fins i tot pot utilitzar un codi numèric per identificar-los (per exemple: Làctics=1, verdures=2, pastes=3, etc.), però és evident que això és totalment arbitrari i que no es pot establir cap relació quantitativa entre els diferents valors (les verdures no són "el doble" dels làctics).

2.1 Representació de dades en variables qualitatives

Dades i instruments principals:

A.- Taules de freqüències i percentatges

. Freqüència absoluta

. Freqüència relativa

. Percentatge

. Freqüència absoluta acumulada (només variables ordinals)

. Freqüència i percentatge relatius acumulats (només variable ordinals)

Exemple: Suposem que en una mostra de 40 persones es determina en cada cas quin és el grup sanguini. Imaginem que els resultats obtinguts són els següents:

. Freqüència absoluta

Grup sanguini	**Freqüència**
A-	3
A+	9
AB-	1
AB+	3
B-	2
B+	5
O-	3
O+	14
Total general	**40**

La taula inicial inclou només les freqüències absolutes, és a dir, el número de casos de cada categoria

. Freqüència relativa i percentatge

La freqüència relativa és el número de casos de cada categoria dividit pel número total, i el percentatge és simplement la freqüència relativa multiplicada per 100

Grup sanguini	Freqüència	Freq. relativa	Percentatge
A-	3	0,075	7,5
A+	9	0,225	22,5
AB-	1	0,025	2,5
AB+	3	0,075	7,5
B-	2	0,05	5
B+	5	0,125	12,5
O-	3	0,075	7,5
O+	14	0,35	35
Total general	**40**	**1**	**100**

Cal indicar que la freqüència relativa es pot interpretar directament com a probabilitat. Per exemple, a la taula anterior podem dir que si agafem a l'atzar una persona dins de les 40 estudiades, la probabilitat de que pertany al grup B+ és de 0,05; o la probabilitat de que pertany al grup O+ és de 0,35, etc.

En el cas dels grups sanguinis, al tractar-se d'una variable nominal no té sentit calcular les freqüències i percentatges acumulats. Això es podria fer en el cas d'una variable ordinal, com és el cas de l'IMC quan el categoritzem com s'indicava abans (IN=Inferior al normal; N=Normal; S=Sobrepès; Obesitat). Si obtenim les dades corresponents a aquesta variable en una mostra de 50 persones, els resultats podrien ser els següents:

Categoria IMC	Freq	Freq. rel	%	Freq ac.	F.rel.ac.	% acum
IN	4	0,08	8	4	0,08	8
N	23	0,46	46	27	0,54	54
S	14	0,28	28	41	0,82	82
O	9	0,18	18	50	1	100
Total general	**50**	**1**	**100**	**50**	**1**	**100**

Com es pot observar fàcilment, la freqüència acumulada no és altra cosa que la freqüència corresponent a una categoria donada més les freqüències de totes les categories anteriors. El mateix es pot dir de les freqüències relatives acumulades i dels percentatges acumulats.

En moltes ocasions es plantegen estudis que incorporen més d'una variable qualitativa. Per exemple, es pot intentar esbrinar com es comporta l'IMC en relació amb el gènere. Imaginem doncs que a l'exemple anterior separem els resultats d'homes i dones. Suposem que a la mostra estudiada hi ha 22 homes i 28 dones, i que els resultats obtinguts són els següents:

Categoria IMC/Gènere	Homes	Dones	Total
IN	1	3	**4**
N	10	13	**23**
S	7	7	**14**
O	4	5	**9**
Total	**22**	**28**	**50**

Aquest tipus de taula, que relaciona dues o més variables qualitatives, s'anomena *taula de contingència*.

En aquest cas també és possible, i molt rellevant, fer el càlcul dels percentatges corresponents a cadascuna de les combinacions de categories:

Taula de contingència de percentatges respecte del total de casos:

Categoria IMC/Gènere	Homes	Dones	**% Marginals**
IN	2%	6%	**8%**
N	20%	26%	**46%**
S	14%	14%	**28%**
O	8%	10%	**18%**
% Marginals	**44%**	**56%**	**100%**

Òbviament els percentatges s'obtenen pel procediment habitual, és a dir, dividint el número de casos de la categoria entre el número de casos total i multiplicant per cent. Per exemple, per al grup d'homes amb sobrepès (7 casos): (7/50)*100 = 14%. De la mateixa forma podem llegir qualsevol altra casella: per exemple, un 8% de subjectes de la mostra són homes i presenten obesitat, o un 6% són dones i mostren un IMC inferior al normal. Els percentatges totals per a cada fila i columna s'anomenen *marginals*, i també podem ser útils. Els marginals ens diuen, per exemple, que un 46% del total de la mostra està a la categoria Normal (independentment de que siguin homes o dones), o que un 56% de la mostra són dones.

Aquesta taula no és suficient per entendre correctament la relació entre les dues variables implicades, ja que només ens informa de la distribució de casos per categories respecte del total, però el que ens pot interessar realment és com es distribueixen les categories d'IMC en relació amb el gènere, o viceversa. Per exemple, mirant la taula de casos inicial, podem veure que de les 4 persones que estan a la categoria d'IMC inferior al normal hi ha 1 home i 3 dones. Això significa, traduït a percentatges, que del conjunt de persones de la categoria IN el 25% són homes i el 75% són dones (òbviament: (1/4)*100 = 25%). Si fem el mateix amb les altres tres categories d'IMC el resultat és el següent:

Categoria IMC/Gènere	**Homes**	**Dones**	**Total**
IN	25%	75%	**100%**
N	43,5%	56,5%	**100%**
S	50%	50%	**100%**
O	44,4%	55,6%	**100%**

De la mateixa forma podem establir la distribució de categories d'IMC dins de cadascun dels dos gèneres. En aquest cas, per exemple, podem veure que dins de la categoria de dones, amb un total de 28 persones, hi ha 13 casos que són la categoria d'IMC normal. Per tant, podem dir que el 46,4% de les dones de la mostra estan a aquesta categoria de l'IMC (naturalment: (13/28)*100=46,4%). Per al conjunt dels resultats la taula seria la següent:

Categoria IMC/Gènere	Homes	Dones
IN	4,5%	10,7%
N	45,5%	46,4%
S	31,8%	25,0%
O	18,2%	17,9%
Total	**100%**	**100%**

Aquí podem veure, per exemple, que la distribució d'IMC per als dos gèneres no és idèntica, hi ha diferències sobretot pel que fa dues categories, la IN i la S. Només el 4,5% dels homes mostren un IMC inferior al normal, mentre que això els passa al 10,7% de les dones. En canvi, el 31,8% dels homes mostren sobrepès, mentre que això es produeix només a un 25% de les dones. Això ens dona una primera indicació de que pot existir relació entre les dues variables, però per afirmar-ho amb major seguretat hauríem de fer una prova estadística de significació, tal i com veurem més endavant al parlar de l'estadística inferencial.

Tots els percentatges calculats a qualsevol taula de contingència es poden transformar fàcilment en probabilitats, simplement dividint per 100. Per exemple, la probabilitat de que un home presenti un IMC normal és de 0,455. La mateixa probabilitat per a una dona és de 0,464.

B.- Representació gràfica

Les dades de variables qualitatives es poden representar gràficament sobretot a partir de dues eines, com són els diagrames de barres i els diagrames de sectors.

Si agafem l'exemple de la distribució de grups sanguinis, les dades obtingudes es poden representar en un diagrama de barres, ja sigui vertical o horitzontal.

Diagrama vertical:

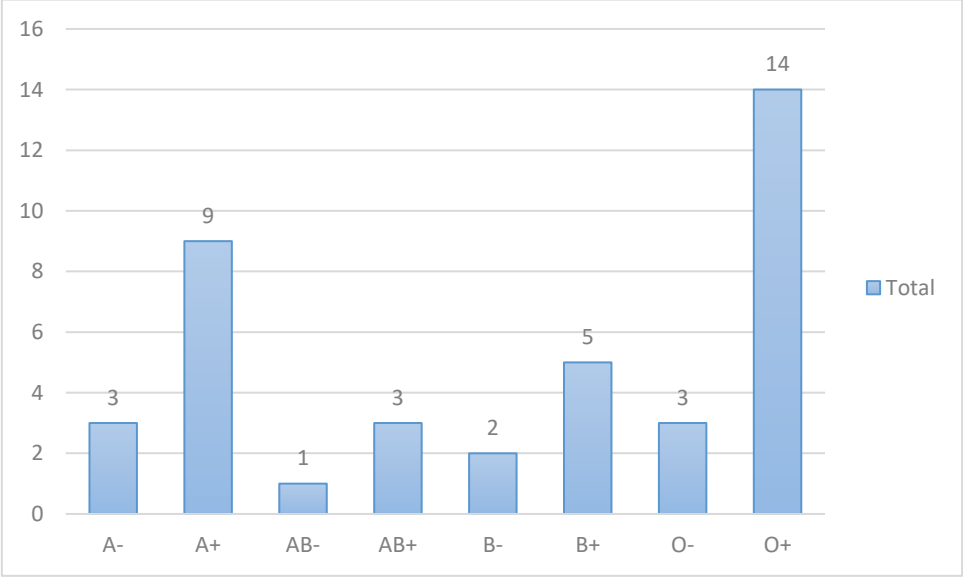
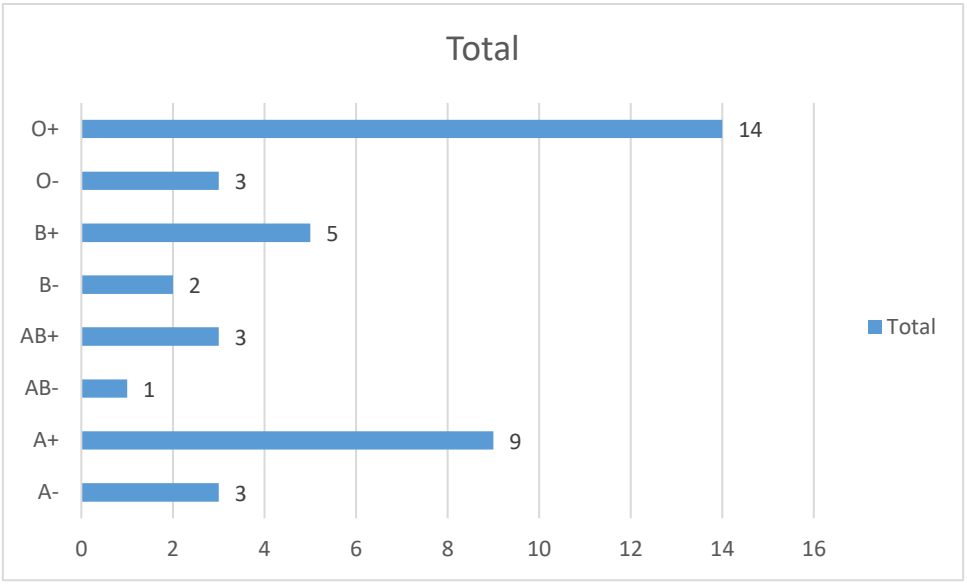
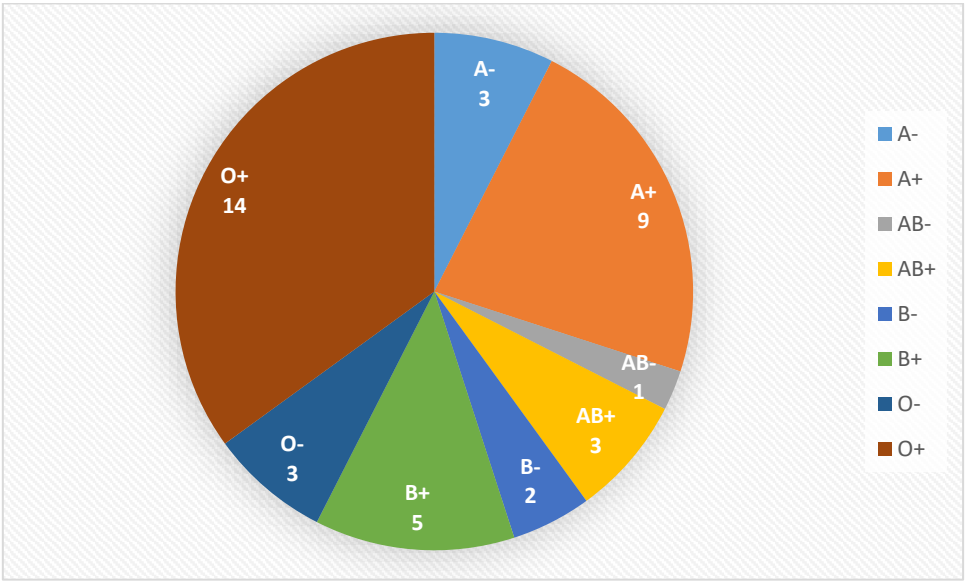


Diagrama horitzontal:

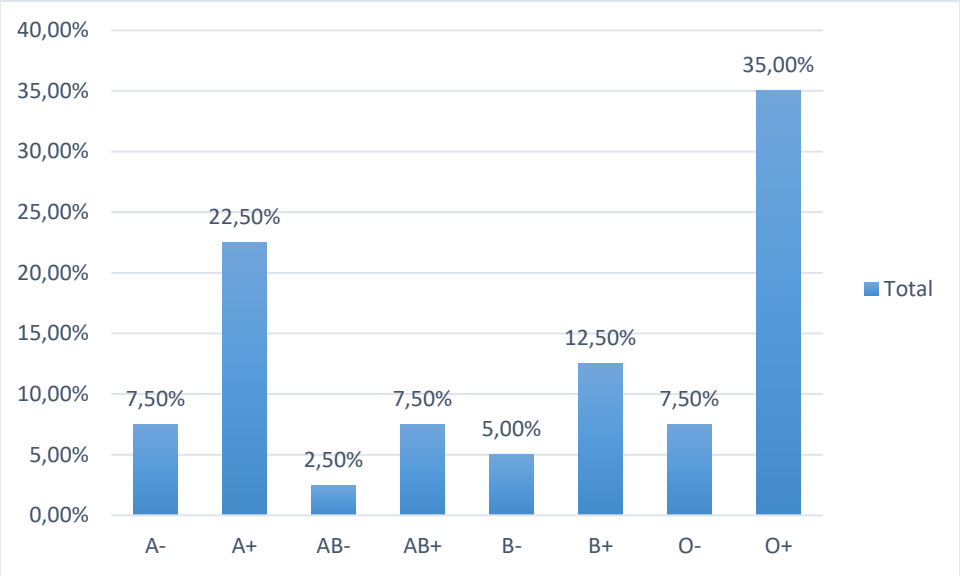


Un diagrama de sectors tindria la forma següent:

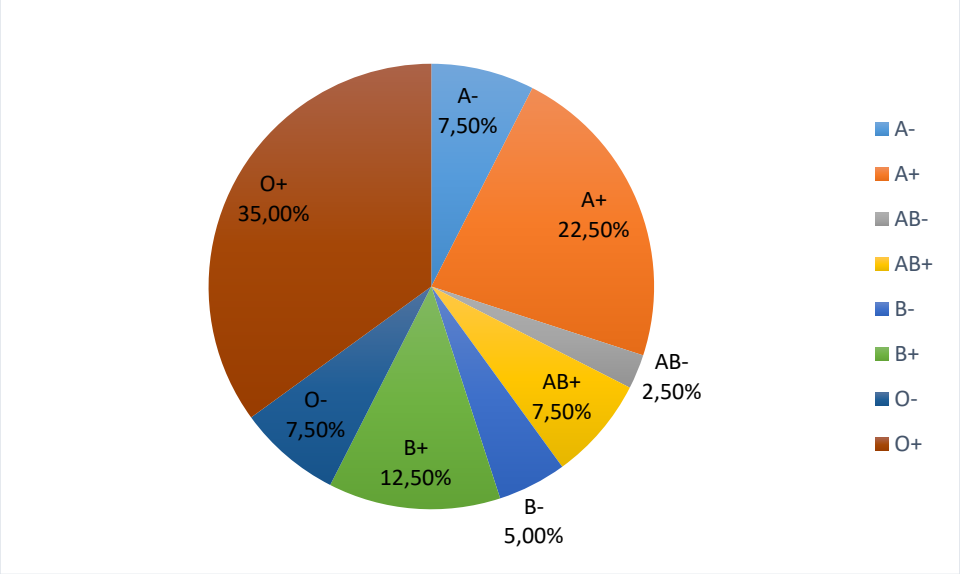


Com es pot veure, es tracta de representacions analògiques basades en la longitud de les barres o en l'àrea dels sectors, en els dos casos proporcionals a la magnitud del resultat.

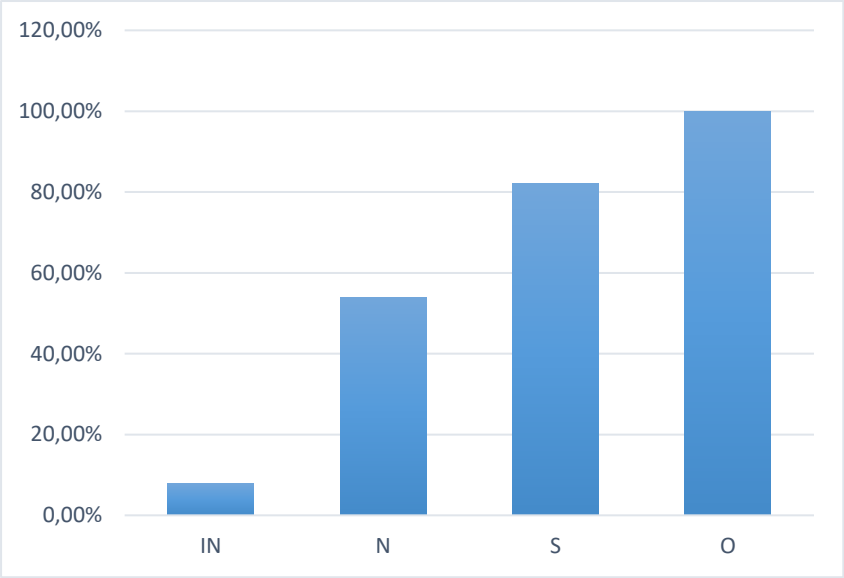
Naturalment, també es poden fer gràfics basats no en les freqüències sinó en els percentatges. L'aspecte del gràfic seria similar, però introduint percentatges en lloc de freqüències absolutes. Per exemple, el diagrama de barres per als percentatges seria:



I el diagrama de sectors corresponent és:

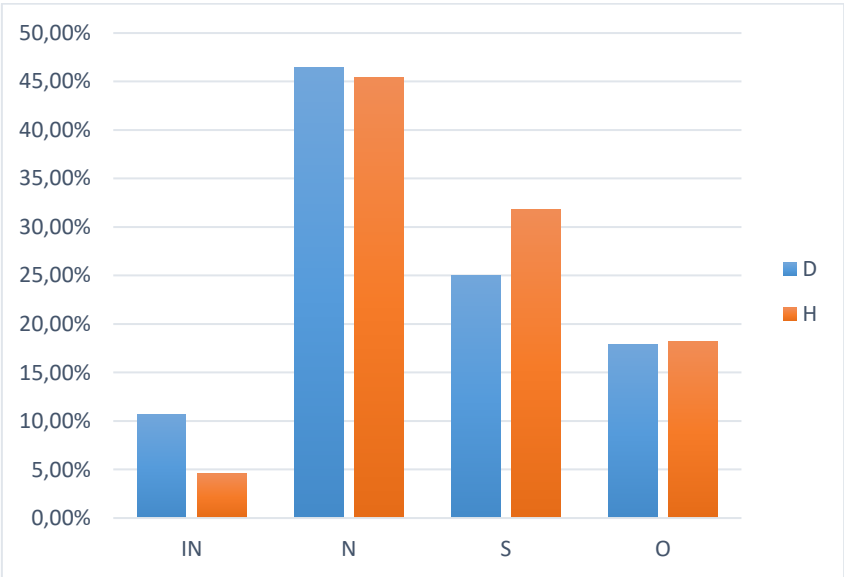


En el cas de variables de caràcter ordinal, com és l'IMC, també és possible fer una representació de les freqüències o dels percentatges acumulats. Per exemple, recollint les dades de percentatges acumulats de la taula que hem vist abans, la representació seria:



Categoria IMC	%	% acum
IN	8	8
N	46	54
S	28	82
O	18	100
Total general	**100**	**100**

Si treballem amb dues variables es poden fer representacions separades per a cadascuna d'elles, però són especialment interessants els gràfics que reuneixen informació sobre les dues variables i permeten veure el seu comportament conjunt. En aquest cas els gràfics més útils són els de percentatges, ja que mostren millor la possible relació entre les variables. Una representació útil pot ser la següent, agafant les dades d'una de les taules que hem utilitzat i comentat anteriorment:



Categoria IMC/Gènere	Homes	Dones
IN	4,5%	10,7%
N	45,5%	46,4%
S	31,8%	25,0%
O	18,2%	17,9%
Total	**100%**	**100%**

La inspecció del gràfic ens mostra novament, com ja havíem detectat, una distribució desigual de les categories d'IMC per a homes i dones i, per tant, la possibilitat d'existència de relació entre les dues variables.

C.- Estadístics descriptius

A més de les taules de freqüències i percentatges i de la representació gràfica, de vegades pot ser útil utilitzar algun tipus d'estadístic per oferir una visió general dels resultats obtinguts. Cal dir, però, que en el cas de les variables qualitatives aquesta possibilitat està molt limitada, a diferència del que passa amb les variables quantitatives.

En el cas de les variables nominals, l'únic índex que es pot utilitzar sempre és la *moda*. La moda no és altra cosa que la categoria que té la major freqüència absoluta. Per exemple, en el cas de la distribució de grups sanguinis, la moda és la categoria O+, ja que recull el número més gran de casos de la mostra analitzada. Si hi haguessin dues categories amb la mateixa freqüència absoluta, hauríem de dir que hi ha dues modes i, per tant, la distribució seria bimodal. Com és evident, la moda no ens diu res que no fos conegut abans, ja que amb les taules de freqüències es pot veure perfectament la distribució de casos per a les diferents categories, incloent quina o quines reuneixen un major número de casos.

Naturalment, a les variables de caràcter ordinal també es pot calcular la moda. Per exemple, en el cas de l'IMC, la moda correspon al grup catalogat com a Normal, que reuneix el número més gran de casos.

En el cas de variables qualitatives ordinals alguns autors plantegen també la possibilitat de calcular algun altre índex com seria el cas de la mediana, tot i que aquest estadístic és molt més propi de les variables quantitatives. La mediana és aquell valor que deixa per sobre i per sota de sí el 50% de casos registrats. és a dir, és un valor central. Quan el número de categories és petit no té utilitat plantejar-se un índex com la mediana, però si treballéssim, per exemple, amb 10 o 20 categories, potser sí que podria tenir interès identificar quina d'elles és la que es troba en una posició central, ni que sigui de forma aproximada. En tot cas, el tema de la mediana, com el d'altres estadístics descriptius, es tractarà amb molt més detall al parlar de variables quantitatives, on es molt més habitual i natural l'ús d'estadístics descriptius.

2.2 Representació de dades en variables quantitatives

A) Definicions generals, tabulació i representació gràfica

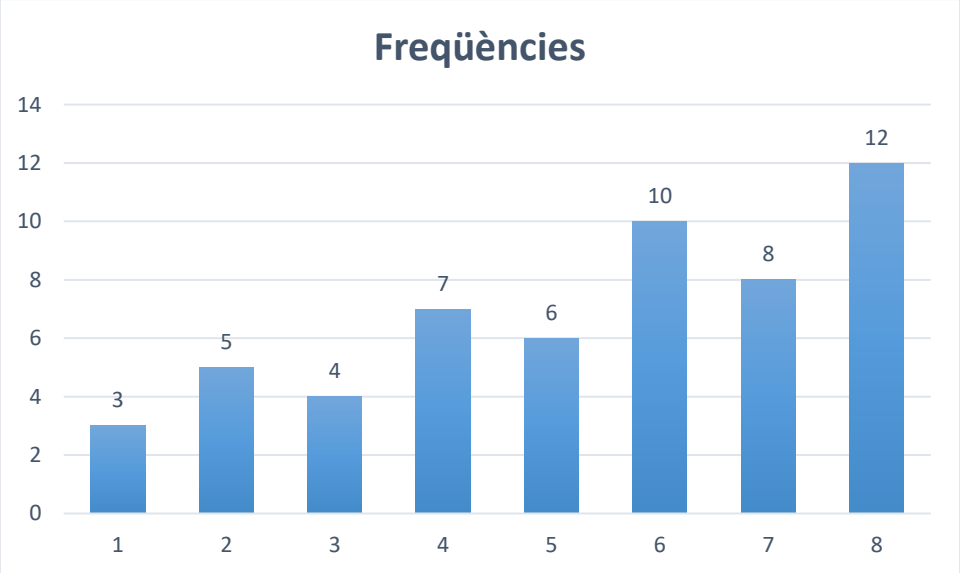
En el cas de les variables quantitatives cal tenir present en primer lloc que, encara que les variables quantitatives discretes i les contínues tenen molts punts en comú, hi ha algunes distincions importants pel que fa a la seva naturalesa i tractament.

Quan treballem amb una variable discreta, el primer que cal tenir present és que el número de resultats diferents possibles que podem obtenir és finit. Si volem saber quin és el número de germans que té cadascun dels subjectes en una mostra de 30 persones, és evident que la immensa majoria dels resultats es mourà entorn de valors petits (segurament entre 0 i 5), i els valors superiors estaran poc representats. En tot cas, el número de resultats diferents serà finit. Imaginem un segon exemple, en el qual proposem a un grup de 55 persones afectades d'algun tipus de trastorn alimentari que realitzin un conjunt de sessions de suport. El número de sessions dependrà de l'evolució de cada cas, però el mínim de sessions a realitzar és d'una i el màxim de vuit. Suposem que els resultats obtinguts són els següents (no s'ofereixen les dades brutes, cas per cas, sinó directament la taula de freqüències absolutes):

Número de visites	Freqüències
1	3
2	5
3	4
4	7
5	6
6	10
7	8
8	12
Total general	**55**

Això ens diu, per exemple, que hi ha 3 persones que han vingut únicament a una sessió, mentre que n'hi ha 12 que han vingut a les 8 sessions previstes, i el mateix per a tots els altres valors.

Aquests resultats són fàcilment representables, per exemple amb un diagrama de barres similar als que realitzàvem amb les variables qualitatives:



Naturalment, com passava amb les variables qualitatives, podem fer també les taules de freqüències relatives, de percentatges absoluts i relatius, i també podem calcular les freqüències i percentatges acumulats., com fèiem amb les variables qualitatives ordinals:

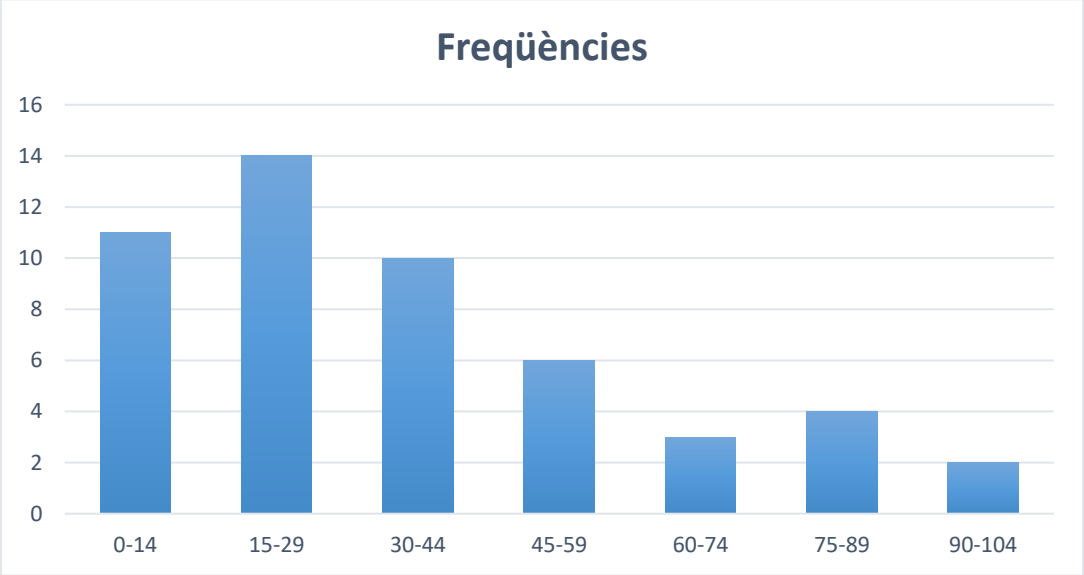
Nº visites	Freq. Abs.	Freq. Rel.	%	Freq. Acum.	F. Rel. Acum.	% Acum
1	3	0,054	5,4	3	0,054	5,4
2	5	0,091	9,1	8	0,145	14,5
3	4	0,073	7,3	12	0,218	21,8
4	7	0,127	12,7	19	0,345	34,5
5	6	0,109	10,9	25	0,454	45,4
6	10	0,182	18,2	35	0,636	63,6
7	8	0,146	14,6	43	0,782	78,2
8	12	0,218	21,8	55	1	100
Total general	**55**	**1**	**100**	**55**	**1**	**100**

Quan el número de valors diferents obtingut en un estudi és molt gran, tot i ser finit en tractar-se d'una variable discreta, pot ser útil treballar amb intervals, per simplificar el tractament de les dades. Per exemple, suposem que es fa un estudi sobre el consum d'alcohol i es donen a tots els participants les instruccions necessàries per tal que facin el registre durant un mes de totes les ingestes d'alcohol que realitzin (es registra una ingesta cada vegada que es consumeix una beguda alcohòlica, ja sigui amb els menjars, festes, a la feina, a casa o a qualsevol altra situació). Suposem també que els resultats obtinguts estan entre 0 i 95 ingestes. En aquest cas, pot ser útil agrupar les dades en intervals: per exemple, de 0 a 5, de 6 a 10, d'11 a 15, etc. D'aquesta manera reduïm la complexitat de les dades a efectes de la seva representació.

Suposem que els resultats obtinguts són els següents, agrupant-los en intervals d'un rang de 15 punts (de 0 a 14, de 15 a 29, etc.):

Número d'ingestes	**Freqüència**
0-14	11
15-29	14
30-44	10
45-59	6
60-74	3
75-89	4
90-104	2
Total general	**50**

Naturalment, en aquest cas es podrien calcular també les freqüències relatives, percentatges, etc. La representació gràfica de les freqüències absolutes corresponents als diferents intervals seria, utilitzant un diagrama de barres (també es podria fer un diagrama de sectors, si es vol):



Una altra forma que pot ser útil per representar les dades és el diagrama *Stem and Leaf* (que se sol traduir en català com a *diagrama de tronc i fulles o de tiges i fulles*, i en castellà com a *diagrama de tallo y hojas*). En aquest tipus de diagrama es treballa també per intervals, definits

per la primera o primeres xifres de les puntuacions obtingudes, segons el criteri que adoptem. Per exemple, en el cas de les dades d'ingesta d'alcohol, les 50 puntuacions obtingudes es poden representar així:

n: 50

0 | 00347788

1 | 134567888

2 | 23456889

3 | 02349

4 | 022345

5 | 04567

6 | 16

7 | 368

8 | 13

9 | 25

La part esquerra de cada fila és el tronc, i la part dreta les fulles. Si agafem la fila 4, aquí tindrem representats tots els casos els quals el número d' ingestes d'alcohol ha estat d'entre 40 i 49. Comptant les fulles, veiem que hi ha 6 casos, i ajuntant el tronc i les fulles, veiem que els resultats són: 40, 42, 42, 43, 44, 45. Aquest diagrama ens permet representar de forma pràctica les dades cas per cas i ens indica també la freqüència absoluta que podem trobar a diferents intervals.

Les aplicacions estadístiques ofereixen altres possibilitats de representació gràfica de les dades que poden ser explorades fàcilment i no seran tractades aquí.

Si treballem amb variables quantitatives contínues, la forma lògica de procedir és mitjançant intervals. Des d'un punt de vista teòric, una variable quantitativa contínua ens pot oferir un número infinit de resultats diferents, ja que entre dos valors qualssevol de la variable, per propers que siguin, se'n podem trobar infinits més. Naturalment, des d'un punt de vista pràctic això està limitat per la precisió de les mesures, que sempre és finita. Si agafem l'exemple de la temperatura corporal, entre una temperatura de 36 graus i una de 37 hi ha infinits valors possibles. Però entre 36,4 i 36,5 també hi ha infinits valors; i entre 36,44 i 36,45 també, i així successivament. Per tant, l'única forma de procedir és definint intervals i treballant a partir d'ells. Suposem que podem obtenir la temperatura amb una precisió de dos decimals. En aquest cas, podríem definir, per exemple, els intervals següents:

(35,50-35,99) , (36,00-36,49) , (36,50-36,99), etc.

i a partir d'ells podrem classificar els diferents resultats obtinguts.

Una altra forma de representar els mateixos intervals seria aquesta:

(35,5-36,0(, (36,0-36,5(, (36,5-37,0(, etc.

El parèntesi invertit al final de l'interval indica que aquest últim valor està exclòs de l'interval i, per tant, forma part del següent. En conseqüència, un valor de 36,00 graus formaria part del

segon interval, i no del primer. Aquesta forma de mostrar els intervals pot ser útil quan es treballa amb variables mesurades amb molts decimals, ja que permet simplificar la representació dels intervals i deixar clar quins són els seus límits superior i inferior.

Naturalment, en el cas d'una variable quantitativa contínua es poden realitzar les mateixes tabulacions i representacions esmentades anteriorment, i això es faria de forma molt semblant al cas de variables quantitatives discretes agrupades per intervals. Per exemple, suposem que mesurem el pes en kg. a una mostra de 40 persones. Imaginem que els resultats s'obtenen amb una precisió d'un decimal, i que són els següents:

78,3	68,3	58,3	81,3
86,5	74	67,2	83,5
55,6	80,1	74,3	71,4
93,4	63,5	71	73,5
67	68,3	70,4	92,4
72,3	56,3	77,3	74,1
101,7	72,2	62,8	67,8
88,7	49,5	88,3	83,6
75,4	81,2	74,6	92,4
59,9	76,3	79,1	61,4

Per a major facilitat per a algunes de les operacions que es faran més endavant, és molt útil ordenar les puntuacions de menor a major:

49,5	67,8	74,1	81,3
55,6	68,3	74,3	83,5
56,3	68,3	74,6	83,6
58,3	70,4	75,4	86,5
59,9	71	76,3	88,3
61,4	71,4	77,3	88,7
62,8	72,2	78,3	92,4
63,5	72,3	79,1	92,4
67	73,5	80,1	93,4
67,2	74	81,2	101,7

Si agrupem els resultats en intervals de 5 kgs. a partir del valor més petit, la taula de freqüències absolutes seria la següent:

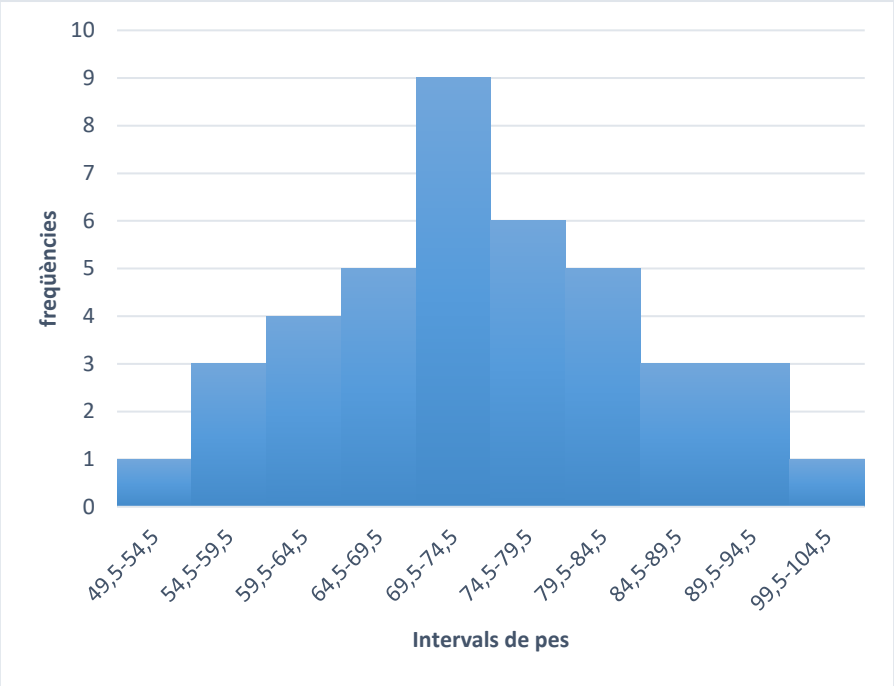
Pes en kg.	Freqüències
(49,5-54,5(	1
(54,5-59,5(	3
(59,5-64,5(	4
(64,5-69,5(	5
(69,5-74,5(	9
(74,5-79,5(	6
(79,5-84,5(	5
(84,5-89,5(	3
(89,5-94,5(	3
(94,5-99,5(	0
(99,5-104,5(	1
Total general	**40**

Naturalment, podem utilitzar un número d'intervals diferent, en funció del grau de precisió que necessitem. Per exemple, les mateixes dades es podrien agrupar en 5 intervals:

Pes en kg.	Freqüències
(49,5-61,5(	6
(61,5-73,5(	12
(73,5-85,5(	15
(85,5-97,5(	6
(97,5-109,5(	1
Total general	**40**

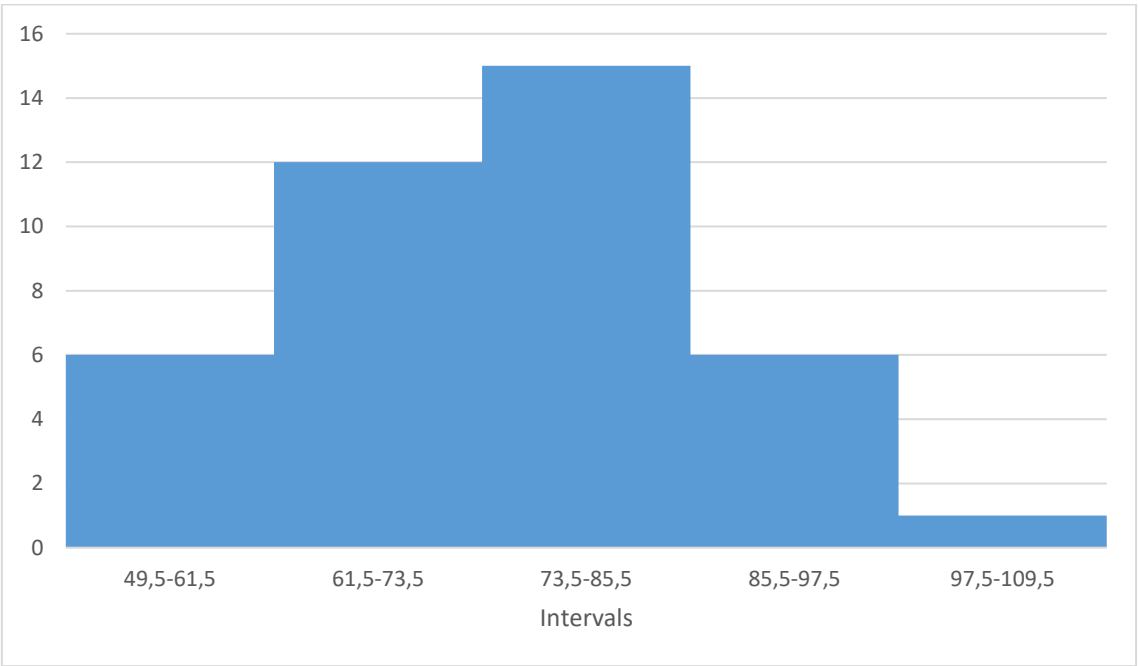
En el cas de dades agrupades per intervals en variables quantitatives contínues, la representació gràfica se sol fer mitjançant un *histograma*, que no és sinó una variant de la representació ordinària per diagrama de barres. Si l'amplitud dels intervals és idèntica, com és el cas més habitual, l'alçada de les barres o rectangles ve donada simplement per les freqüències, com als diagrames convencionals. Si els intervals tinguessin amplades desiguals caldria fer una correcció, i l'alçada de cada rectangle es determinaria amb l'expressió Alçada = freqüència / amplitud de l'interval. Normalment amb variables contínues no s'inclouen separacions entre els rectangles corresponents a cada interval, emfatitzant d'aquesta manera el concepte d'àrea i de distribució contínua.

Si utilitzem l'agrupació en 10 intervals de la distribució de pesos, i considerant que l'amplada dels intervals és idèntica, l'histograma seria:



Cal tenir en compte que en aquest cas hi ha un interval (el que aniria de 94,5 fins a 99,5) on no hi ha cap cas (freqüència 0). Moltes aplicacions estadístiques per defecte no mostren els intervals amb freqüència 0 a les representacions gràfiques, cosa que cal tenir present al valorar-les.

En el cas de l'agrupació en 5 intervals, l'histograma és:



B) Estadístics descriptius en variables quantitatives

A més de les tabulacions habituals i de la representació gràfica de resultats, les variables quantitatives, siguin discretes o contínues, ens ofereixen la possibilitat d'utilitzar un número d'estadístics descriptius molt superior al propi de les variables qualitatives. Els més importants són els següents:

B1. Mesures de tendència central: Moda, mediana, mitjana

B2. Mesures de dispersió i de posició: Amplitud o rang, variància, desviació típica, coeficient de variació, percentils, rang interquartílic

B3. Mesures de forma: Asimetria i curtosi

B1. Mesures de tendència central

Com indica la seva denominació, intenten identificar o calcular valors representatius de la centralitat de la distribució de resultats.

Moda

Com ja s'ha indicat anteriorment, la moda és el valor més freqüent en una determinada distribució. En variables quantitatives contínues estrictament no té sentit buscar la moda a partir dels casos individuals. Fins i tot encara que la mesura tingui una precisió limitada, la informació que obtindrem de la moda a partir de casos individuals serà molt pobre. Si agafem, per exemple, la distribució de pesos que hem vingut comentant, podem veure que la majoria de valors individuals tenen una freqüència de 1, i només en dos casos (les puntuacions 68,3 i 92,4) hi ha una freqüència de 2. Clarament això no ens ofereix cap informació d'interès. En aquest cas és molt més lògic treballar per intervals i veure quin d'ells té la major freqüència. Això ens permet veure que, en la distribució amb 10 intervals, l'interval amb major freqüència és el de 69,5-74,5, que té 9 casos. Per tant, la moda correspondria a aquest interval. En cas d'utilitzar només 5 intervals, el de major freqüència és l'interval 73,5-85,5. Es podria calcular fins i tot un valor numèric concret per a la moda, però això és de poca utilitat.

Si es treballa amb una variable discreta i no s'agrupen els resultats per intervals, sí que es pot establir quin es valor concret al qual correspon la moda. Per exemple, en el cas del número de sessions de suport realitzades, que s'ha tractat abans, està clar que la moda és de 8, ja que és el resultat que ha obtingut una major freqüència.

Mediana

La mediana és el valor de la variable que divideix la distribució de resultats en dues parts amb el mateix número de casos o, com s'ha indicat abans, és la puntuació que deixa un 50% de casos per sota i un 50% per sobre.

Per calcular la mediana cal ordenar els valors obtinguts de menor a major i buscar quin es el valor central. Suposem els següents resultats de qualsevol variable, ordenats de menor a major:

3, 4, 6, 7, 7, 8, 10, 11, 13, 15, 15

Aquí la mediana és 8, ja que per sota d'aquest valor hi ha cinc casos i per sobre uns altres cinc.

Si el número de casos és parell, llavors cal fer la mitjana de les dues puntuacions centrals. Per exemple:

3, 4, 6, 7, 7, 8, 9, 10, 11, 13, 15, 15

En aquest cas no hi ha un únic valor central, sinó dos (8 i 9). Si els agafem en conjunt, la resta de valors queden distribuïts equitativament, cinc per sota i cinc per sobre. En aquest cas la mediana és 8,5, resultat de fer la mitjana dels dos valors centrals.

En el cas de la distribució de resultats de pes, disposem de 40 dades. En tractar-se d'un número parell, haurem de localitzar els dos valors centrals i fer-ne la mitjana. Si ordenem els valors de menor a major (des de 49,5 fins a 101,5), veurem que els valors centrals són 74,0 i 74,1 (posicions 20 i 21). En conseqüència, la mediana serà de 74,05.

Com a norma general, si tenim un total de N casos, en cas que N sigui imparell la mediana és la puntuació que està a la posició (N+1)/2. Si, en canvi, N és un número parell, la mediana serà la mitjana de les puntuacions situades a les posicions (N/2) i (N/2)+1. En el cas de la distribució de pesos, la mediana és la mitjana de les puntuacions situades a la posició 20 (40/2) i 21 (40/2 + 1), una vegada ordenats els resultats de menor a major.

Mitjana aritmètica

Cal dir en primer lloc que hi a més d'un tipus de mitjana (mitjana geomètrica, mitjana quadràtica, mitjana aritmètico-geomètrica, etc.), però la que s'utilitza gairebé sempre al treball estadístic és la *mitjana aritmètica*.

La mitjana aritmètica és simplement el resultat de sumar tots els resultats obtinguts i dividir aquesta suma entre el número total de casos. És a dir, en el cas d'una mostra amb *n* casos:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Si agafem l'exemple de la distribució de pesos en una mostra de 40 casos, caldria en primer lloc obtenir el sumatori de totes les puntuacions: 78,3 + 86,5 + 55,6 + ... + 92,4 + 61,4 = 2973,2. Dividint aquesta suma entre el número de casos (40) obtenim la mitjana, que és de 74,33.

Si tenim la taula de freqüències absolutes el càlcul de la mitjana es pot simplificar. Per exemple, en el cas del número de sessions terapèutiques, a partir de la taula de freqüències absolutes es podria procedir de la forma següent:

Nº visites	**Freq. Abs.**
1	3
2	5
3	4
4	7
5	6
6	10
7	8
8	12
Total general	**55**

En aquest cas, la fórmula a aplicar es basa en cadascun dels valors i la seva freqüència :

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i \cdot n_i}{n}$$

És a dir:

$$\frac{(1 \cdot 3) + (2 \cdot 5) + (3 \cdot 4) + \cdots + (8 \cdot 12)}{55} = 5{,}36$$

És fàcil veure que els dos procediments són equivalents i es poden utilitzar indistintament.

A l'hora d'interpretar una mitjana aritmètica cal tenir molt presents dues coses:

. En primer lloc, la seva representativitat depèn molt de la variabilitat dels resultats. Si els resultats són poc variables i s'agrupen bastant entorn de la mitjana, aquest valor es pot considerar força representatiu. Ara bé, si els resultats són molt dispersos aquesta representativitat es pot posar en dubte.

Per posar un cas senzill de poca representativitat de la mitjana, imaginem que tenim els resultats següents: 2, 4, 18, 20. La mitjana d'aquests resultats és de 11. Com es pot veure, aquest valor no ens diu gairebé res de la distribució de resultats originals, la mitjana obtinguda és molt diferent dels 4 valors originals. En canvi, si els valors fossin 8, 10, 12 i 14, la mitjana seria igualment de 11, però en aquest cas el valor de la mitjana sí que ens dona informació rellevant, ja que les diferents dades es mouen en valors força propers als de la mitjana aritmètica.

. En segon lloc, la mitjana aritmètica és molt sensible als valors extrems, i això pot complicar la seva interpretació. Imaginem, per exemple, els resultats següents: 2, 5, 3, 6, 4, 5, 18. La mitjana d'aquests 7 valors és de 6,14. Ara bé, és evident que hi un valor extrem (18), molt diferent dels altres, i que influeix molt en aquest resultat. Si obtenim la mitjana de les 6 primeres dades, sense tenir en compte el 18, el resultat és de 4,17, substancialment diferent del primer. Per això es diu que la mitjana és molt sensible a les puntuacions extremes, tot i que lògicament aquest efecte s'atenua quan el número de registres és gran. En tot cas cal revisar sempre la distribució de dades per detectar la possible presència de valors extrems, tant per dalt com per baix.

A diferència de la mitjana aritmètica, la mediana es veu poc afectada per la presència de valors extrems, ja que en realitat és un índex ordinal, que es refereix a la posició de les diferents puntuacions i no a la seva magnitud. En el cas anterior, si ordenem els 7 valors originals de menor a major tindrem: 2, 3, 4, 5, 5, 6, 18. La mediana és de 5. Si l'últim valor fos encara més extrem (per exemple, 45 en lloc de 18), la mediana continuaria essent igual a 5, en canvi la mitjana aritmètica canviaria de forma important.

B2. Mesures de dispersió

Estan destinades a valorar el grau de variabilitat que presenten les dades. Una distribució uniforme, en la qual tots els valors fossin iguals, tindria una dispersió o una variabilitat nul·les. Per exemple, si analitzem 80 unitats d'un cert producte i trobem que el contingut de sucre de cadascuna d'elles és exactament de 0,18 mg., llavors la variabilitat dels resultats seria igual a 0. Òbviament això no passarà normalment, i del que es tracta és de saber fins a quin punt els resultats obtinguts són més o menys variables entre sí. Els principals índexs de dispersió són els següents:

. Amplitud o rang

És la mesura de dispersió més senzilla. Consisteix en la diferència entre el registre més gran (màxim) i el més petit (mínim). Per exemple, en el cas de la distribució de pesos el rang seria: 101,7 – 49,5 = 52,2.

Cal tenir present que, lògicament, el rang és molt sensible a l'existència de valors extrems, ja que són precisament aquests valors els que el determinen. Per tant, si trobem valors anormalment alts o baixos, caldrà tenir-ho en compte a l'hora de determinar i interpretar el rang.

. *Variància i desviació típica*

Són les dues mesures de dispersió més importants i utilitzades, i serveixen de base per al càlcul d'altres. Tant l'una com l'altra ens indiquen quin és el grau de dispersió dels resultats entorn de la mitjana aritmètica i, per tant, són molt importants a l'hora de valorar la representativitat d'aquesta mesura central.

La variància d'una mostra amb n casos i una mitjana ȳ s'obté a partir de la fórmula següent:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

És a dir, al numerador de la fórmula cal fer la diferència entre cadascuna de les puntuacions i la mitjana de la distribució, elevar-la al quadrat i fer el sumatori per a les n puntuacions. Per aquest motiu aquest numerador s'anomena també *suma de quadrats*. La simple divisió de la suma de quadrats entre el número de casos ens dona el resultat de la variància.

Cal assenyalar que la fórmula anterior és la que s'ha utilitzat clàssicament, però actualment és més habitual emprar una lleugera variant d'aquesta expressió:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Com es veurà a la part d'estadística inferencial, el motiu de l'ús d'aquesta segona formulació és que ofereix una estimació no esbiaixada de la variància poblacional. Des d'un punt de vista pràctic, la majoria de les aplicacions estadístiques utilitzen per defecte aquesta segona fórmula, de manera que serà també la que s'utilitzarà normalment en aquest text. Cal dir també que quan es treballa amb un número de casos elevat, la diferència de resultat entre una i l'altra és petita.

En el cas de la distribució de pesos, i recordant que la seva mitjana és de 74,33, el procediment seria:

Pes	Diferència	Dif. al quadrat
78,3	78,3-74,33= 3,97	15,76
86,5	86,5-74,33= 12,17	148,11
55,6	55,6-74,33= -18,73	350,81
.....	.....	.....
.....	.....	.....
92,4	92,4-74,33= 18,07	326,52
61,4	61,4-74,33= -12,93	167,18
	∑ = 0	∑ = 5106,70

La variància serà el resultat de dividir la suma de quadrats entre el número de casos menys un: 5106,70 / 39 = 130.94

Per la seva part, la desviació típica o desviació estàndard (que representarem amb la lletra s) s'obté simplement fent l'arrel quadrada de la variància. Per tant, en el nostre cas :

$$s = \sqrt{130{,}94} = 11{,}44$$

Vista la forma de calcular tant la variància com la desviació típica, cal fer algunes indicacions:

. Si fem la suma de les diferències entre cada puntuació i la mitjana aritmètica, sense elevar-les al quadrat, aquesta suma serà necessàriament igual a zero, per la pròpia definició i naturalesa de la mitjana.

. La variància, tot i que com es veurà més endavant és un índex molt útil, és difícil d'interpretar ja que es basa en diferències elevades al quadrat (diferències quadràtiques), i per tant no manté les unitats de mesura de la variable original

. En canvi, la desviació típica sí que està expressada en les mateixes unitats originals, i per tant es pot interpretar de forma més natural. Per exemple, veurem més endavant que en moltes distribucions la majoria de les puntuacions tendiran a agrupar-se entorn a la mitjana dins del rang d'una desviació típica per sobre o per sota, tot i que això depèn de la forma concreta de la distribució. En el nostre cas, el lògic seria pensar que la majoria de registres estarien entre els valors 74,33 ± 11,44, és a dir entre 62,89 i 85,77, com passa realment.

. *Coeficient de variació o de dispersió*

A partir de la desviació típica es pot derivar molt fàcilment el coeficient de variació, que es calcularia segons l'expressió següent:

$$C.V. = \left[\frac{s}{\bar{y}}\right] . 100$$

És a dir, el coeficient de variació s'obté dividint la desviació típica entre la mitjana, i multiplicant el resultat per cent. En el nostre cas: C.V. = (11,44/74,33) · 100 = 15,39.

El coeficient de variació ens permet valorar la magnitud de la desviació típica en relació amb la mitjana. Això té interès en diferents situacions. Per exemple, en el cas dels resultats de pesos, imaginem que obtenim una segona mostra de casos, que la seva mitjana és de 65,5 i la desviació típica de 11,23. Es pot veure que les desviacions típiques de les dues mostres (11,44 i 11,23) són bastant semblants, tot i que la segona és menor que la primera. Els coeficients de variació són, respectivament, de 15,39 (ja calculat abans) i de 17,14. Per tant, en termes relatius a les mitjanes, la segona distribució és més dispersa que la primera, ja que el seu coeficient de variació és superior.

El coeficient de variació és especialment útil quan es volen comparar resultats de dispersió de variables registrades amb unitats de mesura diferents. Per exemple: en una mateixa mostra, quina variable mostraria més dispersió, el pes o l'IMC?. El C.V. seria útil per respondre aquesta pregunta.

. *Percentils, quartils, decils i rang interquartílic*

El percentil k d'una variable és el valor que deixa sota d'ell el k per cent dels casos. Per exemple, el percentil 30 és la puntuació que deixa per sota el 30% dels casos i, per tant, té el 70% per sobre.

Per relacionar aquesta qüestió amb temes anteriors, cal assenyalar que la mediana no és altra cosa que el percentil 50 ja que, com s'ha indicat, deixa per sobre i per sota el 50% dels casos.

Els percentils que s'utilitzen més sovint són els anomenats *quartils* els quals, com indica el seu nom, divideixen la distribució en quarts: Els quartils són els percentils 25 (primer quartil), 50 (segon quartil o mediana) i 75 (tercer quartil). També s'utilitzen de vegades els *decils*, que divideixen la distribució en 10 parts: decil 1 (percentil 10), decil 2 (percentil 20), etc.

El *rang interquartílic* és la diferència entre el tercer i el primer quartil, és a dir, entre els percentils 75 i 25.

Per a molts dels càlculs i explicacions relatius a percentils ens serà imprescindible l'ordenació de les puntuacions obtingudes de menor a major, com ja hem fet anteriorment. Recuperem aquí els resultats de l'exemple dels pesos en una mostra de 40 casos, ordenats de menor i major i indicant a més el número d'ordre de cada puntuació:

Pes	Posició	Pes	Posició	Pes	Posició	Pes	Posició
49,5	1	67,8	11	74,1	21	81,3	31
55,6	2	68,3	12	74,3	22	83,5	32
56,3	3	68,3	13	74,6	23	83,6	33
58,3	4	70,4	14	75,4	24	86,5	34
59,9	5	71	15	76,3	25	88,3	35
61,4	6	71,4	16	77,3	26	88,7	36
62,8	7	72,2	17	78,3	27	92,4	37
63,5	8	72,3	18	79,1	28	92,4	38
67	9	73,5	19	80,1	29	93,4	39
67,2	10	74	20	81,2	30	101,7	40

A partir de les dades ordenades es poden calcular els percentils, però cal fer algunes precisions. Si pensem, per exemple, en el percentil 10 (o decil 1), aquesta és la puntuació que deixa el 10% de les puntuacions restants a sota i el 90% a sobre. El 10% de 40 són 4. Això ens podria fer pensar que el percentil 10 és la cinquena puntuació (59,9), ja que en deixa 4 per sota. Ara bé, si ho analitzem detingudament, veurem que la cinquena puntuació no pot ser el percentil 10, ja que deixa 4 puntuacions per sota (el 10,25% de les 39 puntuacions restants) i 35 per sobre (un 89,75% de les restants). Si fem els mateixos càlculs per a la quarta puntuació (58,3) veurem que tampoc compleix amb els requisits necessaris per ser el percentil 10 (deixa sota d'ella 3 puntuacions de les 39 restants, és a dir, un 7,69%). Per tant, queda clara que el percentil 10 "real" (és a dir, la puntuació que realment deixaria un 10% de les restants a sota) es troba en algun punt entre la quarta posició (que deixa un 7,69% per sota) i la cinquena (que deixa un 10,25% per sota), encara que aquest valor intermedi no es correspongui amb cap resultat obtingut en aquesta mostra (però sí és un valor que podria existir a la població d'origen de la mostra).

Per calcular el valor exacte del percentil podem utilitzar l'expressió següent:

$$P_x = \frac{x\,Y_i + (100 - x)Y_s}{100}$$

Per aplicar aquesta expressió el primer que cal és esbrinar entre quines dues puntuacions es troba el percentil. Una vegada les tinguem localitzades, anomenarem Y_i a la puntuació inferior i Y_s a la superior. En el nostre cas, la puntuació inferior és 58,3 i la superior és 59,9, ja que sabem que el percentil es troba entre les dues. En cas de dubte, es pot esbrinar fàcilment quina és la puntuació inferior fent el producte n · p/100, és a dir, es multiplica el número de casos pel percentil que es busca i es divideix entre 100. En el nostre cas busquem el percentil 10, per tant la puntuació inferior serà: 40 · 10 / 100 = 4, és a dir, la que ocupa la quarta posició (58,3), i el percentil estarà entre aquesta i la següent (59,9).

Aplicant la fórmula indicada anteriorment, el percentil 10 serà:

$$P_{10} = \frac{10 \cdot\ 58{,}3 + (100 - 10) \cdot 59{,}9}{100} = 59{,}74$$

És a dir, per sota de la puntuació "teòrica" de 59,74 hi haurà el 10% de les puntuacions (a la nostra mostra, 4 puntuacions) i per sobre hi haurà el 90% (a la nostra mostra, 36 puntuacions). És una puntuació "teòrica" perquè en realitat a aquesta mostra no hi ha ningú que tingui un pes de 59,74, però a la població d'origen de la mostra sí que es podria donar realment aquesta puntuació.

Tècnicament, el que s'ha fet és una *interpolació* entre els valors inferior i superior.

Si, per exemple, es vol calcular el primer quartil (Q1), és a dir, el percentil 25, determinarem en primer lloc la puntuació inferior: 40 · 25 / 100 = 10. Per tant, el quartil 1 està entre les puntuacions 10 i 11, una vegada ordenades (o sigui, entre 67,2 i 67,8). Per obtenir el valor exacte utilitzarem l'expressió ja coneguda:

$$P_{25} = \frac{25 \cdot\ 67{,}2 + (100 - 25) \cdot 67{,}8}{100} = 67{,}65$$

De la mateixa manera procedirem amb tots els percentils.

Un cas particular de percentil és la mediana, com s'ha indicat abans. La mediana és el percentil 50 (o el quartil 2), per tant deixa per sota un 50% dels casos i per sobre l'altra 50%. Com s'ha indicat també, quan el número de casos és imparell no cal fer interpolació, ja que hi ha un valor central únic que compleix amb aquesta condició. En canvi, quan el número de casos és parell la mediana es troba entre els dos valors centrals, i quan fem la mitjana entre aquests dos valors en el fons el que fem és interpolar entre ells, donant el mateix pes al valor inferior i al superior.

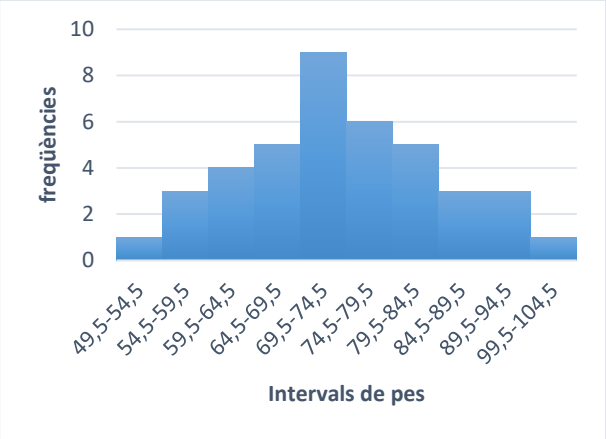
També és possible calcular els percentils a partir de les dades agrupades per intervals, però aquest procediment no es desenvoluparà aquí.

Finalment, podem calcular el rang inter-quartílic, és a dir, la diferència entre els quartils 3 i 1, o entre els percentils 75 i 25. Com ja hem vist, el quartil 1 és igual a 67,65. Si fem de la mateixa forma el càlcul del quartil 3, el resultat és de 81,225. Per tant el rang inter-quartílic és: 81,225 – 67,65 = 13,575.

B3. Mesures de forma

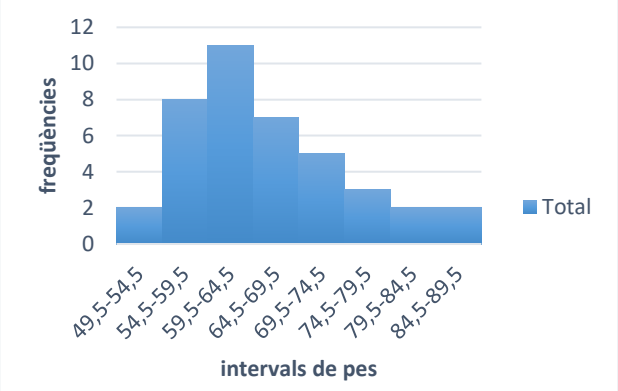
Un primer element a considerar en relació amb la forma d'una distribució és la seva major o menor simetria. El *coeficient d'asimetria* mesura fins a quin punt la distribució dels valors de la variable és aproximadament igual als dos costats dels seus valors centrals, i en quin grau es desvia d'una simetria perfecta.

Si recuperem el gràfic de la distribució de pesos amb 10 intervals, veurem que la distribució és relativament simètrica, però no de forma perfecta. Hi ha uns intervals centrals que reuneixen la major part dels casos, i les freqüències disminueixen a mesura que anem a valors més extrems, tot i que no exactament de forma simètrica.

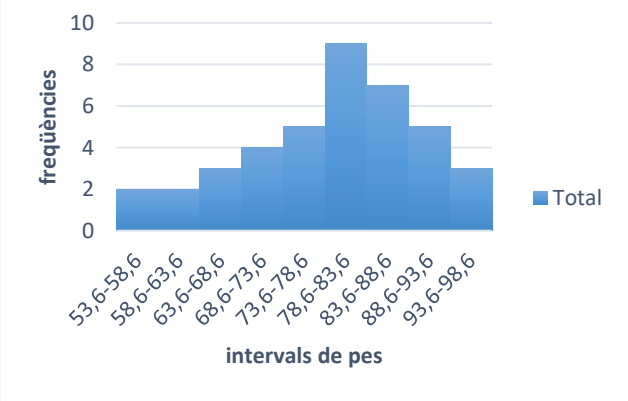


Distribució de les dades obtingudes. Es pot considerar "relativament" simètrica

Amb unes altres dades podríem trobar, però, distribucions molt més asimètriques. Per exemple:



Distribució amb asimetria positiva, o esbiaixada per la dreta (la "cua" de la distribució és a la dreta)



Distribució amb asimetria negativa, o esbiaixada per l'esquerra (la cua de la distribució és a l'esquerra)

Tot i que la visualització dels gràfics pot ser força il·lustrativa, és molt més precís calcular un índex o coeficient numèric d'asimetria. Hi ha diferents propostes al respecte, en funció de les característiques de la variable, si es treballa o no per intervals, etc. També cal advertir que les diferents aplicacions informàtiques per a estadística poden oferir possibilitats diferents, i de vegades utilitzen per defecte índexs també diferents.

Un dels índexs més utilitzats és el coeficient d'asimetria de Fisher, que es calcularia de la forma següent:

$$A_s = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^3}{n\, s^3}$$

on, com s'ha vist anteriorment, n és el número de casos a la mostra, y_i és cadascuna de les puntuacions, $\bar{y}$ és la mitjana i s és la desviació típica.

Cal insistir en que hi ha variants de la fórmula per obtenir el coeficient d'asimetria, de forma que el resultat pot ser variable. En tot cas, per al coeficient de Fisher i les seves variants, la interpretació és la següent:

$A_s > 0$ Asimetria positiva $A_s = 0$ Simetria perfecta $A_s < 0$ = Asimetria negativa

Cal indicar que l'asimetria és detectable també, tot i que de forma imperfecta, a partir de la comparació de la mitjana aritmètica i la mediana de la distribució. Si la simetria és perfecta, la mitjana i la mediana coincideixen; si hi ha asimetria positiva, la mitjana és més gran que la mediana; i si hi ha asimetria negativa, la mitjana és més petita que la mediana.

Per a les tres distribucions de pesos representades més amunt, el coeficient d'asimetria és:

Primer cas: $A_s = 0{,}14$ Hi ha una lleugera asimetria positiva

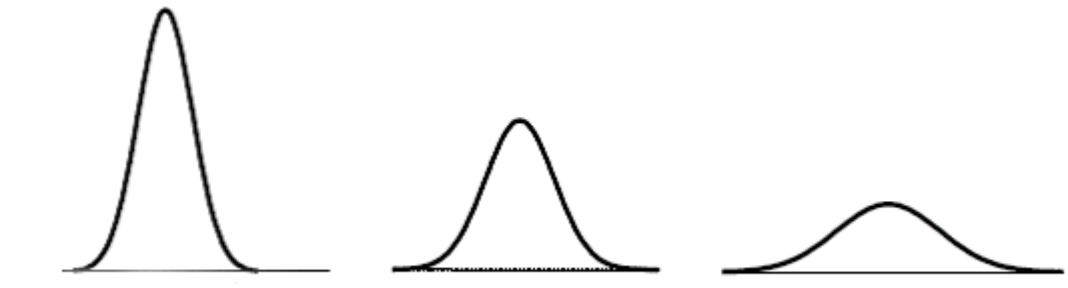
Segon cas: $A_s = 0{,}70$ En aquest cas l'asimetria positiva és molt més marcada

Tercer cas: $A_s = -0{,}49$ Es tracta d'un cas d'asimetria negativa

És interessant remarcar que en el primer cas, que correspon a les dades de pes que s'han vingut utilitzant al llarg d'aquest apartat, la mitjana aritmètica és de 74,33 i la mediana de 74,05. Ja que la mitjana és superior a la mediana, això ens confirma l'asimetria positiva d'aquesta distribució, tal i com ens indica el coeficient obtingut.

A més de l'asimetria, també pot ser d'interès valorar una segona característica de les distribucions, anomenada *curtosi*. La curtosi es refereix al caràcter més o menys "pla" de la distribució, però només té interès interpretar-la en distribucions més o menys simètriques. La curtosi ens indica fins a quin punt les puntuacions es concentren més o menys entorn dels valors centrals, i per tant té relació amb la variabilitat de les dades. En funció del grau de concentració de les puntuacions podrem parlar de distribucions leptocúrtiques (molt concentrades), mesocúrtiques (amb una concentració intermèdia) i platicúrtiques (poc concentrades).

Gràficament:



A l'igual que passava amb l'asimetria, es pot calcular també un coeficient de curtosi i, com també passa amb l'asimetria, hi ha més d'una forma de fer aquest càlcul. Una fórmula força utilitzada i fàcil d'interpretar és la següent:

$$A_c = \left[\frac{1}{ns^4}\sum_{i=1}^{n}(y_i - \bar{y})^4\right] - 3$$

On *n* és el número de casos de la mostra, *s* és la desviació típica, y_i és cadascuna de les puntuacions i $\bar{y}$ és la mitjana.

El coeficient s'interpreta fàcilment segons els criteris següents:

$A_c > 0$ Distribució leptocúrtica $A_c \approx 0$ Distribució mesocúrtica $A_c < 0$ = Distribució platicúrtica

En el cas de la nostra distribució original de pesos, el coeficient de curtosi és igual a -0,07, és a dir, força proper a zero, per tant la distribució és aproximadament mesocúrtica. El fet que el coeficient sigui negatiu ens indica que, en tot cas, tendiria més cap a una forma platicúrtica o poc concentrada, i no a una forma leptocúrtica o molt concentrada.

3.- Conceptes generals de probabilitat i distribucions de probabilitat. Distribució normal

3.1. Conceptes generals de teoria de la probabilitat

La revisió d'alguns conceptes bàsics sobre la probabilitat i sobre les distribucions de probabilitat ens permetrà fer la transició des de l'estadística descriptiva, pensada sobretot per a la representació de les dades, fins a l'estadística inferencial, que ens permetrà prendre *decisions estadístiques* a partir de la informació disponible.

Per entendre correctament la importància de la probabilitat cal tenir present, entre altres coses, que en estadística treballem gairebé sempre amb mostres de casos provinents d'una certa població. Com s'indicava a l'apartat 1, habitualment la grandària de les poblacions fa inviable treballar amb tots els casos, i ens obliga a agafar-ne una o diverses mostres, normalment amb algun tipus de criteri basat en l'atzar (mostratge aleatori simple, per quotes, etc.). Si el mostratge es fa bé, el resultat ha de ser una mostra aleatòria i representativa de la població, i és a partir d'aquesta premissa que ens podem plantejar extreure conclusions generals a partir de la mostra, que és precisament el que fa l'estadística inferencial.

Un concepte bàsic en la teoria de la probabilitat és el de *succés o experiment aleatori*. Un experiment aleatori és una prova o fenomen que presenta variació en els seus resultats, i en el qual aquests resultats no són coneguts a priori. Un exemple molt senzill, com s'indicava a l'apartat 1, és el de llençar una moneda.

La probabilitat és la funció que mesura l'expectativa de que succeeixi cadascun dels esdeveniments possibles, als quals s'assigna un valor entre 0 i 1. En cas d'una moneda correctament equilibrada, les probabilitats esperades són p = q = 0,5, expressió en la qual p és la probabilitat d'obtenir una cara, i q la probabilitat d'obtenir una creu. Ja que aquests dos successos són els únics possibles en el nostre experiment, la suma de les seves probabilitats ha de ser igual a 1 i, per definició, q=(1-p).

Naturalment, en experiments més complexos poden donar-se més resultats. Si tirem un dau de 6 cares, tenim 6 resultats possibles (des de 1 fins a 6), i si el dau està correctament equilibrat, cadascun d'aquests successos té la mateixa probabilitat, igual a 1/6 (0,16666.....). Això dóna lloc a la *funció de probabilitat*, que en aquest darrer cas seria:

Resultat	Probabilitat
1	$0,1\hat{6}$
2	$0,1\hat{6}$
3	$0,1\hat{6}$
4	$0,1\hat{6}$
5	$0,1\hat{6}$
6	$0,1\hat{6}$
Total	**1**

La funció de probabilitat es pot calcular de diverses formes. En casos senzills, com els que estem estudiant fins ara, hi ha bàsicament dues maneres d'obtenir la funció:

. Si els successos són equiprobables, com suposem que passa al tirar una moneda o un dau, podem aplicar la regla de Laplace: p = Successos favorables / Successos possibles. En el cas de la moneda, en principi hem de suposar que els dos successos possibles tenen la mateixa probabilitat, per tant, p = ½ (probabilitat de cara) i q = ½ (probabilitat de creu). En el cas del dau de 6 cares, la probabilitat de cada succés individual és p = 1/6, com ja s'ha vist.

. Si els successos no són equiprobables o senzillament no n'estem segurs, una forma d'aproximar-se a la funció de probabilitat és realitzar l'experiment aleatori múltiples vegades, és a dir, podem adoptar una aproximació empírica. Si llancem una determinada moneda 2000 vegades i obtenim 996 cares i 1004 creus, les probabilitats corresponents són p = 996/2000 = 0,498, i q = 1004/2000 = 0,502. Obtenim així una bona aproximació a les probabilitats que podíem suposar teòricament. Naturalment, si en realitzar l'estudi observem una desviació important respecte de l'esperat, això voldrà dir que la moneda no està correctament equilibrada.

Quan obtenim una mostra d'una certa població i realitzem un estudi, els resultats obtinguts poden ser expressats també en forma d'una distribució de probabilitat. Recuperem, per exemple, l'estudi sobre els grups sanguinis en una mostra de 40 persones, els resultats del qual eren els següents:

Grup sanguini	**Freqüència**	**Freq. relativa**	**Percentatge**
A-	3	0,075	7,5
A+	9	0,225	22,5
AB-	1	0,025	2,5
AB+	3	0,075	7,5
B-	2	0,05	5
B+	5	0,125	12,5
O-	3	0,075	7,5
O+	14	0,35	35
Total general	**40**	**1**	**100**

És fàcil veure que els resultats es poden interpretar en forma de probabilitats. Per exemple, si agafem a l'atzar una persona d'entre les 40 de la mostra, quina és la probabilitat de que el seu grup sanguini sigui B+?. Només cal veure que les freqüències relatives es poden interpretar immediatament com a probabilitats, amb suma 1. La probabilitat d'escollir una persona amb grup B+ és de 0,125. O, dit d'una altra forma, la probabilitat de que això passi és igual a 5/40 = 0,125, que és precisament la manera com es calcula la freqüència relativa. Igualment, la probabilitat d'escollir una persona amb grup O+ és de 0,35, i així successivament.

Per a una variable qualitativa o bé quantitativa discreta, podem definir doncs la funció de probabilitat f(x), que ens indica la probabilitat de que es produeixi un determinat valor x. En el cas anterior, i utilitzant les dades sobre freqüències relatives:

x (grups sang.)	f(x) (probabilitats)
A-	0,075
A+	0,225
AB-	0,025
AB+	0,075
B-	0,05
B+	0,125
O-	0,075
O+	0,35

Naturalment, a partir d'aquí es poden respondre diferents preguntes. Per exemple, quina probabilitat hi ha de que una persona de la mostra tingui Rh negatiu?. Només cal sumar les probabilitats de tots els grups amb Rh negatiu: 0,075 + 0,025 + 0,05 + 0,075 = 0,225.

Si treballem amb una variable quantitativa discreta podem procedir de la mateixa forma. Per exemple, si recuperem les dades de l'exemple del número de visites terapèutiques i considerem les freqüències relatives com a probabilitats, la distribució de probabilitat corresponent és:

x (nº visites)	f(x) (probabilitats)
1	0,054
2	0,091
3	0,073
4	0,127
5	0,109
6	0,182
7	0,146
8	0,218

En el cas d'una variable quantitativa discreta podem introduir dos conceptes que en principi són nous, però que en realitat es corresponen amb idees que ja s'han tractat. L'esperança matemàtica d'una variable aleatòria és el valor mitjà teòric dels valors de la variable. Si treballem amb una mostra determinada, l'esperança matemàtica no és altra cosa que la mitjana dels resultats de la mostra. En aquest cas, la mitjana del número de visites era de 5,36, i per tant podem dir que l'esperança matemàtica de la distribució és precisament aquesta. Un altre tema serà preguntar-se quina seria la mitjana de la població d'origen d'aquesta mostra. Això només ho podrem esbrinar fent una estimació a partir de la mitjana mostral, com es veurà més endavant.

De la mateixa forma, es pot parlar de la variància d'una variable aleatòria, que mesura la dispersió dels valors de la variable respecte de l'esperança matemàtica. És evident que aquest concepte coincideix amb el de variància de la mostra, que ja hem comentat anteriorment.

Si treballem amb variables aleatòries contínues, com és el cas del pes, no podem parlar de funció de probabilitat, ja que per la pròpia definició de variable contínua, per a qualsevol valor x que considerem, p(x)=0. En aquest cas hem de parlar de *funció de densitat de probabilitat*, que ens permet calcular la probabilitat de que x prengui un valor entre dos punts concrets.

Podem fer una aproximació a valors de la funció de densitat a partir de les dades empíriques, però amb limitacions. Recuperem, per exemple, els resultats de la variable pes en una mostra de 40 persones els quals, com fem sempre amb les variables contínues, estan agrupats per intervals. Podem calcular la freqüència relativa (i, per tant, la probabilitat) per a cada interval:

Pes en kg.	**Freqüències**	**Freq. Relatives (probabilitats)**
(49,5-54,5(	1	0,025
(54,5-59,5(	3	0,075
(59,5-64,5(	4	0,1
(64,5-69,5(	5	0,125
(69,5-74,5(	9	0,225
(74,5-79,5(	6	0,15
(79,5-84,5(	5	0,125
(84,5-89,5(	3	0,075
(89,5-94,5(	3	0,075
(99,5-104,5(	1	0,025
Total general	**40**	**1**

A partir d'aquí podem afirmar, per exemple, que la probabilitat de que una persona d'aquesta mostra tingui un pes entre 59,5 i 64,5 kgs. és de 0,1.

En aquest cas podem calcular també l'esperança matemàtica i la variància de la variable aleatòria, que es correspondran amb la mitjana i la variància que ja coneixem.

Cal recordar aquí novament que una cosa és treballar amb els resultats d'una mostra, i una altra intentar inferir a partir d'ells les característiques de la població. Si mirem els resultats de pesos de la mostra veurem que, per exemple, no hi ha ningú que tingui un resultat comprès a l'interval (94,5- -99,5(. Per tant, la funció de densitat de probabilitat entre aquests dos valors seria igual a zero. Això vol dir que a la població de referència d'aquesta mostra no hi ha ningú que pesi entre 94,5 i 99,5 quilos?. Evidentment no és així, per tant cal tenir clar que si volen obtenir resultats generalitzables necessitaríem una funció de densitat de probabilitat que ens descrigui el conjunt de la població, i no solament la mostra.

Si treballem amb dues variables, a partir de la taula de contingència podem calcular també algunes probabilitats d'interès. Recuperem, per exemple, les dades de relació entre IMC i gènere que s'utilitzaven a l'apartat 2:

Categoria IMC/Gènere	Homes	Dones	**Total**
IN	1	3	**4**
N	10	13	**23**
S	7	7	**14**
O	4	5	**9**
Total	**22**	**28**	**50**

Utilitzant la definició més elemental de probabilitat (successos favorables / successos possibles) podem respondre diverses preguntes:

Si agafem una persona a l'atzar dins d'aquests mostra de 50:

. Quina és la probabilitat de que sigui home?: p(h) = 22 /50 = 0,44. Naturalment, la probabilitat de que sigui dona és: p(d) = 28/50 = 0,56, o bé 1 – 0,44 = 0,56

. Quina és la probabilitat de que l'IMC correspongui a la categoria sobrepès?: p(sob) = 14/50 = 0,28

Podem calcular també la probabilitat d'*intersecció* de dos successos a i b, que representarem com a p(a∩b). Aquesta és la probabilitat de que els dos successos es produeixin simultàniament. Per exemple: si agafo un cas a l'atzar, quina és la probabilitat de que la persona sigui dona i presenti un IMC normal?:

p(dona∩normal) = 13/50 = 0,26

Igualment podem calcular la probabilitat d'*unió* de successos o resultats, que representarem com a p(aUb). Aquesta és la probabilitat de que es produeixi algun dels dos successos (a o b) o de que es produeixin els dos simultàniament (a i b). Aquí cas distingir dues situacions, en funció de si els successos a i b són o no excloents.

Per exemple, si agafo un cas a l'atzar, quina probabilitat hi ha de que sigui tingui sobrepès o obesitat?. Aquests dos successos són excloents, és a dir, una persona estarà a la categoria sobrepès o a la categoria obesitat, però no a les dues. Per això en aquest cas la probabilitat p(sobUobs) és igual a la suma de les probabilitats de cadascun dels successos:

p(sobUobs)= p(sob) + p(obs) = 14/50 + 9/50 = 0,28 + 0,18 = 0,46

o, si es prefereix: p(sobUobs) = 14+9/50 = 0,46

Què passa en el cas que parlem de dos successos no excloents?. Per exemple, quina és la probabilitat de que una persona escollida a l'atzar dins de la mostra sigui dona i/o estigui a la categoria normal?. En aquest cas cal pensar que les dues coses poden passar per separat, però també simultàniament, i per això ja no són successos excloents. Per tant, quins serien els casos que complirien amb les condicions?: home-normal (10 casos), dona-normal (13), dona-IN (3), dona-sobrepès (7),dona-obesitat (5). Els casos favorables són doncs: 10 + 13 + 3 + 7 + 5 = 38, i la probabilitat buscada és 38/50 = 0,76.

Tècnicament, la probabilitat d'unió amb successos no excloents es pot obtenir aplicant la fórmula:

p(aUb) = p(a) + p(b) - p(a∩b)

En el nostre cas:

p(donaUnormal) = p(dona) + p(normal) - p(dona∩normal) = 28/50 + 23/50 – 13/50 = 0,76

Cal dir que en el cas de successos excloents aquesta expressió també és correcta, ja que llavors p(a∩b)=0. Per exemple, si recuperem el cas vist unes línies més amunt: Quina és la probabilitat de que una persona sigui a la categoria sobrepès o a la categoria obesitat?. Els dos successos

són excloents, per tant p(sob∩obs)=0, ningú pot estar simultàniament a les categories sobrepès i obesitat. Per tant:

p(sobUobs)= p(sob) + p(obs) - p(sob∩obs) = 14/50 + 9/50 – 0 = 0,28 + 0,18 – 0 = 0,46, com ja havíem vist.

Cal afegir que la unió de tots els successos incompatibles possibles té una probabilitat de 1. Per exemple, és evident que qualsevol persona estarà en alguna de les categories que hem definit per a l'IMC, i que només pot estar a una d'elles. Per tant: p(inUnorUsobUobs) = 4+23+14+9/50 = 1.

Un tipus molt important de probabilitats són les probabilitats condicionades, que ens indiquen la probabilitat d'un succés quan se n'ha produït un altre, i representarem com a p(b/a). Quina és la probabilitat de b una vegada s'ha produït a?. Per exemple, quina probabilitat hi ha de que una persona tingui un IMC inferior al normal (IN) si és un home?:

p(IN/home) = 1/22 = 0,045

És important veure que aquí la referència ja no és el total general de casos, sinó el total del primer succés (en aquest cas, el total d'homes, que és 22, dels quals 1 mostra un IMC per sota del normal).

Quina és la probabilitat de que una persona amb obesitat sigui dona?. El primer succés és que tingui obesitat; una vegada produït aquest succés, quina probabilitat hi ha de que sigui una dona?. Ens hem de referir a la distribució d'homes i dones dins de la categoria d'obesitat. Veiem que de les 9 persones obeses, 4 són homes i 5 són dones. Per tant:

p(dona/obs) = 5/9 = 0,555

Cal afegir encara algunes definicions importants:

Dos successos són independents si el resultat d'un no té cap relació ni influència sobre el resultat de l'altra. Per exemple, si tiro una moneda dues vegades, el resultat de la segona tirada no es veurà influït en absolut pel resultat de la primera:

Si la primera tirada ha estat cara: p(cara2/cara1)=0,5 , p(creu2/cara1)=0,5

Si la primera tirada ha estat creu: p(cara2/creu1)=0,5 , p(creu2/creu1)=0,5

Dos successos són dependents si el que passi amb un ofereix informació sobre l'altre.

Per exemple, com hem vist abans, si agafem una persona a l'atzar, quina és la probabilitat de que tingui sobrepès?:

p(sob) = 14/50 = 0,28

Ara bé, si sabem que aquesta persona és un home, quina és la probabilitat de que tingui sobrepès?:

P(sob/home) = 7/22 = 0,32

Per tant, el fet de saber que es tracta d'un home modifica la probabilitat de que tingui sobrepès, ens afegeix informació suplementària.

Si les dues variables fossin totalment independents entre sí, la situació seria diferent. Per exemple, suposem que els resultats obtinguts al mesurar els IMCs en una mostra de 63 persones fossin els següents:

Categoria IMC/Gènere	Homes	Dones	**Total**
IN	2	4	**6**
N	8	16	**24**
S	7	14	**21**
O	4	8	**12**
Total	**21**	**42**	**63**

En aquest cas:

. Si agafem un subjecte a l'atzar, quina és la probabilitat de que sigui obès?: p = 12/63 = 0,19

. Quina és la probabilitat de que sigui obès si és un home?: p = 4 / 21 = 0,19

. I si és una dona?: p = 8 / 42 = 0,19

Per tant, en aquest cas el gènere no ens diu res sobre la probabilitat de mostrar obesitat. Tant si la persona és dona com si és home, la probabilitat de mostrar obesitat és la mateixa, la variable gènere no ens afegeix cap informació respecte d'això. Podem dir, doncs, que amb aquestes dades el fet de mostrar o no obesitat és independent del gènere.

Com veurem més endavant, hi ha proves estadístiques que ens permetran establir l'existència o no de relació entre diferents variables.

3.2 Algunes distribucions teòriques de referència

Quan s'estudien determinats fenòmens es pot constatar que molts d'ells tendeixen a oferir resultats que segueixen unes certes lleis i donen lloc a distribucions de probabilitat identificables.

Per exemple, suposem que sabem que el 40% d'una determinada població està afectada d'estrès, i que això pot influir en els seus hàbits alimentaris. Si agafem una mostra de 10 persones d'aquesta població, quina és la probabilitat de que 6 d'elles presentin estrès?. Aquest fenomen es pot descriure matemàticament a partir d'una *distribució binomial*, aplicable quan estem davant de dos successos excloents (en el nostre cas, presentar o no presentar estrès, a partir dels criteris que s'estableixin). Per establir la funció de probabilitat que descriu aquest tipus de fenòmens necessitem només dos paràmetres: *n* seria el número de casos i π és la probabilitat del succés que interessa. En el cas plantejat, *n* seria igual a 10, ja que agafem 10 persones, i π seria igual a 0,4, ja que la probabilitat de que una persona concreta estigui afectada d'estrès és aquesta (cal recordar que el 40% de la població té aquesta afectació). La funció de probabilitat binomial s'obté de la forma següent:

$$\boldsymbol{p}(\boldsymbol{x_i}) = \binom{n}{x_i} \pi^{x_i} (1-\pi)^{n-x_i}$$

on $\binom{n}{x_i} = \frac{n!}{x_i!\ (n-x_i)!}$

En el nostre cas volem saber quina probabilitat hi ha de que en una mostra de 10 persones (n=10) n'hi hagi 6 (x_i=6) que presentin estrès, sabent que la probabilitat de que una persona qualsevol tingui estrès és de π=0,4. Per tant:

$$\boldsymbol{p}(\boldsymbol{x_i = 6}) = \binom{10}{6} 0{,}4^6 (1-0{,}4)^4 = 0{,}1115$$

Com que es coneix la forma de calcular la funció de probabilitat de la distribució binomial, es pot calcular també la seva esperança matemàtica (equivalent a la mitjana), la seva variància i desviació típica, i també construir unes taules on es pugui consultar directament qualsevol probabilitat, coneixent simplement *n* i π.

Es pot veure que el model binomial és aplicable a successos discrets. En aquest cas s'ha aplicat al número de persones amb estrès dins d'una mostra. Un altre exemple seria el del resultat d'un control de qualitat: quina probabilitat hi ha d'obtenir 4 productes defectuosos o menys en una mostra de 30 productes si la probabilitat de generar un producte defectuós és de 0,1?.

Hi ha altres funcions de probabilitat útils per a variables discretes, com és el cas de la distribució de Poisson, que en realitat no és sinó un cas límit de la distribució binomial, la distribució multinomial, la distribució hipergeomètrica, etc., tot i que no seran estudiades aquí.

Igualment, hi ha distribucions pròpies de variables quantitatives contínues, entre les quals cal esmentar de forma especial la *distribució normal*.

. La distribució normal de probabilitat

Aquesta distribució és sense dubte una de les més importants en el camp de l'estadística, ja que existeixen molts fenòmens que s'hi ajusten de manera més o menys aproximada. Suposem, per exemple, que volem estudiar el contingut de glucosa en sang en persones no diagnosticades de diabetis en una determinada població (per exemple, la població d'habitants de Catalunya) i que obtenim una mostra aleatòria de 5000 casos per tal de fer aquest estudi. Suposem que la mitjana de glucosa en sang obtinguda és de 77,8. El que s'espera és que els valors de glucosa a la mostra (i suposem també que a la població) es distribueixin entorn d'aquesta mitjana de forma que a mesura que ens allunyem d'ella el número de casos sigui cada vegada menor. És a dir, la pressuposició és que la majoria de valors estaran agrupats entorn de la mitjana (ja sigui per sobre o per sota) i que el número de casos anirà baixant a mesura que ens allunyem de la mitjana. En altres paraules, el que s'espera és que els valors es distribueixin entorn de la mitjana en forma del que anomenem *corba normal* o *campana de Gauss*, anomenada així en honor del matemàtic Karl Friedrich Gauss. Aquest autor va desenvolupar la teoria relativa a la distribució normal durant el segle XIX, i també ho va fer de manera independent el matemàtic francès Pierre Simon de Laplace, de forma que també es parla de vegades de la *campana de Laplace-Gauss*.

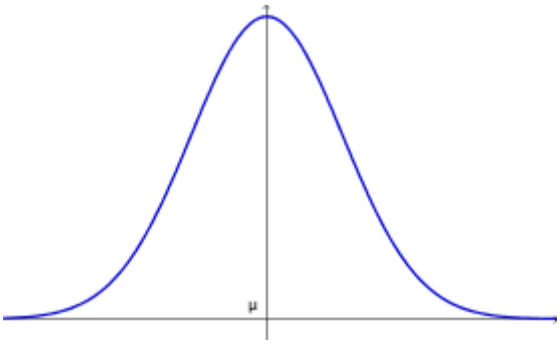
Cal tenir present que en tractar-se d'una distribució contínua no parlarem estrictament de probabilitat, sinó de *densitat de probabilitat*. Això vol dir que el que farem és intentar calcular

quina és la probabilitat d'obtenir un resultat que estigui entre dos valors donats. Per exemple, en el cas anterior, podríem intentar calcular quina és la probabilitat, si agafem una persona a l'atzar, de que el seu nivell de glucosa estigui entre 60 i 70, o entre 64 i 66, o entre 64,9 i 65,0, etc. El que conceptualment no seria correcte és pensar que podem calcular la probabilitat associada a un valor exacte, ja que al ser una variable continua la probabilitat d'obtenir un valor concret teòricament és igual a 0 (considerant que qualsevol valor de la variable ha de tenir infinits decimals, encara que la limitada precisió de la mesura ens impedeixi conèixer-los).

La funció de densitat de probabilitat de la distribució normal és:

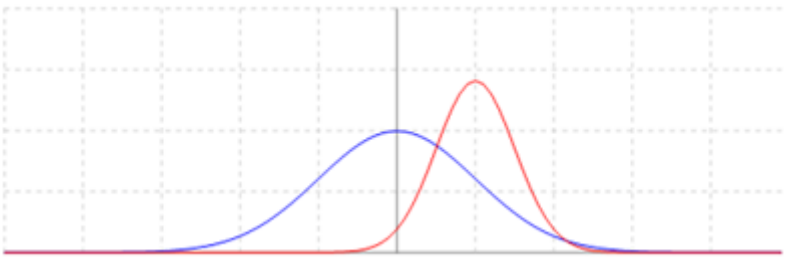
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

La campana de Gauss té una forma molt característica i ben coneguda:



Aquesta corba normal és perfectament simètrica, tot i que naturalment a la pràctica no sol ser així. El punt central i més alt correspon a la mitjana de la distribució. La major o menor amplitud de la corba depèn de la dispersió de les dades. Les mesures d'asimetria i curtosi que s'han mostrat en un apartat anterior ens permeten descriure bé les característiques d'una distribució normal. En el cas d'una distribució normal perfectament simètrica, la mitjana, la mediana i la moda serien coincidents, tot i que a la pràctica això no passa gairebé mai.

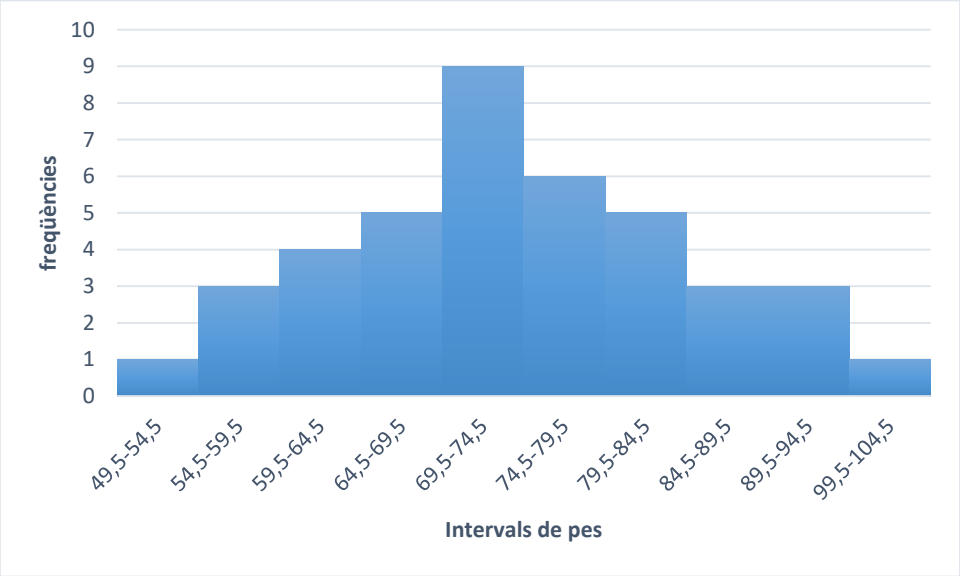
Naturalment, la forma exacta de la corba normal pot variar entre un cas i un altre, en funció de la mitjana i desviació típica de les dades i de la seva distribució concreta. Per exemple:



Suposem que aquestes dues corbes corresponen a dues mostres diferents, a les quals s'ha fet una mesura d'una certa variable (per exemple, la concentració de glucosa en sang). Podem veure que les dues distribucions semblen normals. Ara bé, la blava té una

mitjana inferior a la vermella (està situada més a l'esquerra), i en canvi té una desviació típica superior (les dades estan més disperses entorn de la mitjana). Per tant, aparentment els resultats obtinguts a les dues mostres, tot i ajustar-se gràficament a una llei normal, mostren diferències que caldrà estudiar.

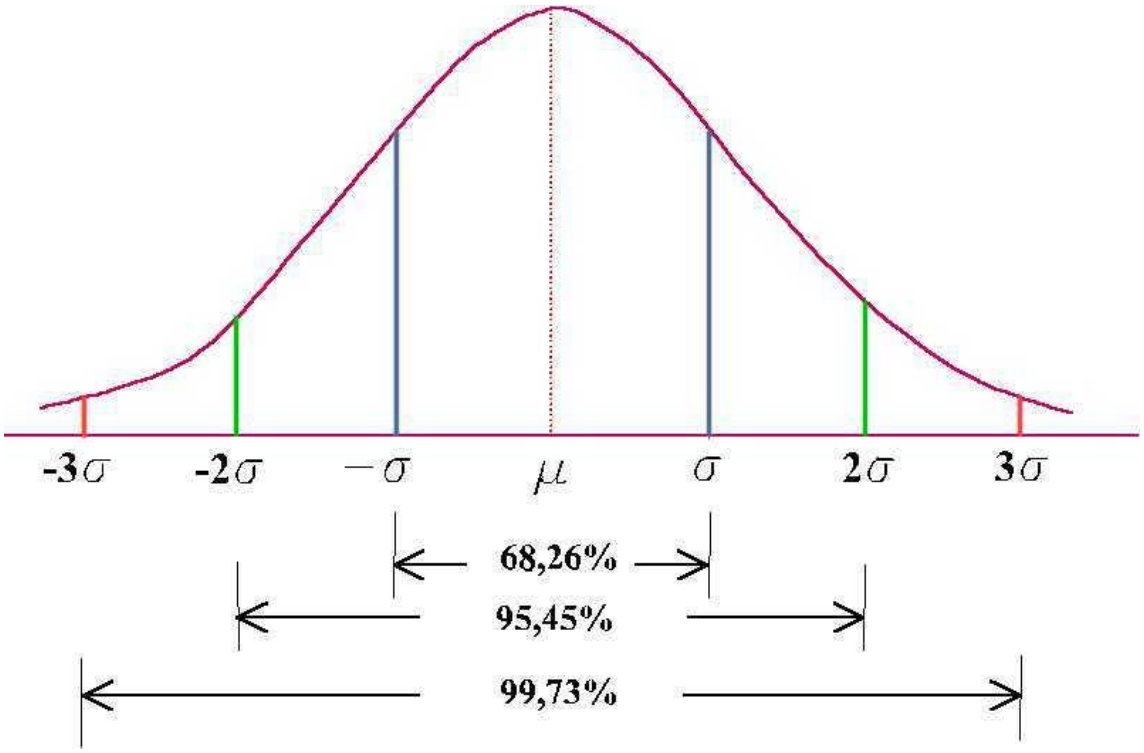
La representació gràfica dels resultats d'una variable ens pot ajudar a fer una primera aproximació a la forma de la distribució. Recordem que la distribució de pesos amb la qual s'ha treballat anteriorment es representava segons l'histograma següent:



Recordant que la mitjana d'aquesta distribució era de 74,33, observem com els valors es distribueixen entorn d'aquest valor mig, amb una disminució progressiva de freqüències a mesura que ens apropem als extrems. Per tant, tot i que l'histograma no ho permet apreciar bé, la distribució en principi té una certa aparença de normalitat. A l'apartat següent es veurà com és possible establir de forma molt més precisa si qualsevol distribució s'ajusta o no a una llei normal. En tot cas, convé recordar que el coeficient d'asimetria per a la nostra distribució, calculat anteriorment, era de 0,14, cosa que ens indica una lleugera asimetria positiva o esbiaixada per la dreta. Per la seva part, el coeficient de curtosi era igual a -0,07, és a dir, proper a zero, cosa que ens indicava una distribució bastant mesocúrtica, tot i que amb una petita tendència cap a l'aplanament (distribució platicúrtica), ja que el coeficient és negatiu.

La distribució normal té diferents propietats d'interès. Per exemple, a una distribució normal "perfecta" amb mitjana μ i desviació típica σ, entre la puntuació mitjana i les puntuacions que estiguin "una desviació típica" per sobre i per sota hi haurà el 68,26% dels casos. Entre les puntuacions situades dues desviacions típiques per sobre i per sota i haurà el 95,45% dels casos, etc.

Gràficament:



Recuperem novament la distribució de pesos que hem vingut utilitzant. Cal recordar que la seva mitjana era de 74,33, i la desviació típica de 11,44. La puntuació que està una desviació típica (o sigma) per sota de la mitjana és (74,33-11,44) = 62,89; i la que està una sigma per sobre és (74,33+11,44)=85.77. Per tant, d'acord amb el que s'ha indicat, podem dir que a la població d'origen d'aquesta mostra, suposant que s'ajusti perfectament a una distribució normal, el 68,26% dels casos estarà entre les puntuacions 62,89 i 85,77.

Com veurem més endavant, si una distribució s'ajusta a una llei normal podrem calcular la probabilitat associada a qualsevol interval que ens interessi (per exemple, la probabilitat d'obtenir una puntuació inferior o superior a qualsevol valor determinat, o la probabilitat d'obtenir una puntuació situada entre dos valors donats).

. Proves de normalitat

Una pregunta important a fer-se davant de qualsevol distribució de dades és: fins a quin punt aquesta distribució s'ajusta a una corba normal o campana de Gauss?. Per respondre a aquesta pregunta s'han ideat diferents *proves de normalitat*. El que fan les proves de normalitat és comparar qualsevol distribució empírica amb una distribució normal "ideal", per tant el que ens diuen és fins a quin punt la distribució empírica es desvia d'una campana de Gauss perfecta.

Algunes proves de normalitat més o menys habituals són les de Shapiro-Wilk (una de les més potents per a mostres petites), Shapiro-Francia, Anderson-Darling, Cramer-Von Mises, Kolmogorov-Smirnov o χ^2 de Pearson, entre d'altres. La majoria d'aplicacions

estadístiques poden calcular alguns o tots aquests índexs de normalitat, que no seran explicats en detall per la seva complexitat. El més important és realitzar una interpretació correcta dels seus resultats, i per tal de fer-ho cal avançar aquí alguns conceptes més propis de l'estadística inferencial, i concretament de les proves estadístiques de significació.

Cadascuna de les proves de normalitat ens ofereix dues dades: En primer lloc, el valor de l'estadístic que calcula cadascuna de les proves (per exemple, l'estadístic w a la prova de Shapiro-Wilk i altres, o l'estadístic d a la prova de Kolmogorov-Smirnov). A més d'això, cadascuna de les proves ens ofereix una probabilitat p, anomenada *significació estadística*, que és la base de la interpretació correcta dels resultats.

Per exemple, si realitzem dues de les proves citades a les nostres dades de la distribució de pesos, els resultats són els següents:

Prova de Shapiro-Wilk: w = 0,992 , p = 0,99

Prova de Kolmogorov-Smirnov: d = 0,066 , p = 0,93

La forma d'obtenir els estadístics w i d és força diferent, però l'important és la interpretació que se'n pot fer a partir de les probabilitats p obtingudes. Cal recordar que el que fan aquestes proves, per un o altre procediment, és comparar la distribució empírica que tenim (en el nostre cas, la distribució de pesos), amb una distribució normal teòrica. Per tant, el que ens permet decidir la prova és fins a quin punt la nostra distribució difereix (o no) d'una distribució normal perfecta. La lògica de la decisió estadística s'explicarà més endavant, però de moment n'hi ha prou amb una interpretació senzilla: convencionalment, si la probabilitat p obtinguda és inferior a 0,05 hem d'interpretar que, amb molt poca probabilitat d'error, les dues distribucions són diferents. En canvi, si el valor de p és superior a 0,05, podem acceptar la idea que les dues distribucions no difereixen significativament. En el nostre cas, podem veure que les probabilitats obtingudes són molt altes (superiors a 0,9), i això ens porta a concloure que la distribució de pesos s'ajusta a una distribució normal de probabilitat. Més endavant s'aprofundirà en la lògica de la decisió estadística i es podrà entendre millor el que implica.

. Què podem fer a partir d'una distribució normal?

Si sabem que una distribució qualsevol s'ajusta a una llei normal, això ens permet respondre a una sèrie de preguntes que poden ser d'interès. Si agafem el cas de la distribució de pesos, i sempre suposant que es tracti d'una mostra representativa, podrem respondre qüestions com, per exemple, les següents:

Si agafem una persona a l'atzar dins de la població d'origen de la mostra:

. Quina probabilitat hi ha que el seu pes sigui inferior a 55 kg.?

. Quina probabilitat hi ha que el seu pes estigui entre 80 i 85 kg.?, etc.

O, expressant les mateixes preguntes en altres termes:

. Quin percentatge de la població té un pes inferior a 55 kg.?

. Quin percentatge de la població té un pes entre 80 i 85 kg.?

La resposta a aquestes qüestions es pot obtenir perquè la distribució normal té una funció de densitat coneguda, tal i com s'ha indicat abans. Això ens permet establir les densitats de probabilitat per al conjunt de la distribució, i respondre qualsevol pregunta del nostre interès. A la pràctica, qualsevol aplicació estadística ens permet calcular les probabilitats associades a una distribució normal qualsevol.

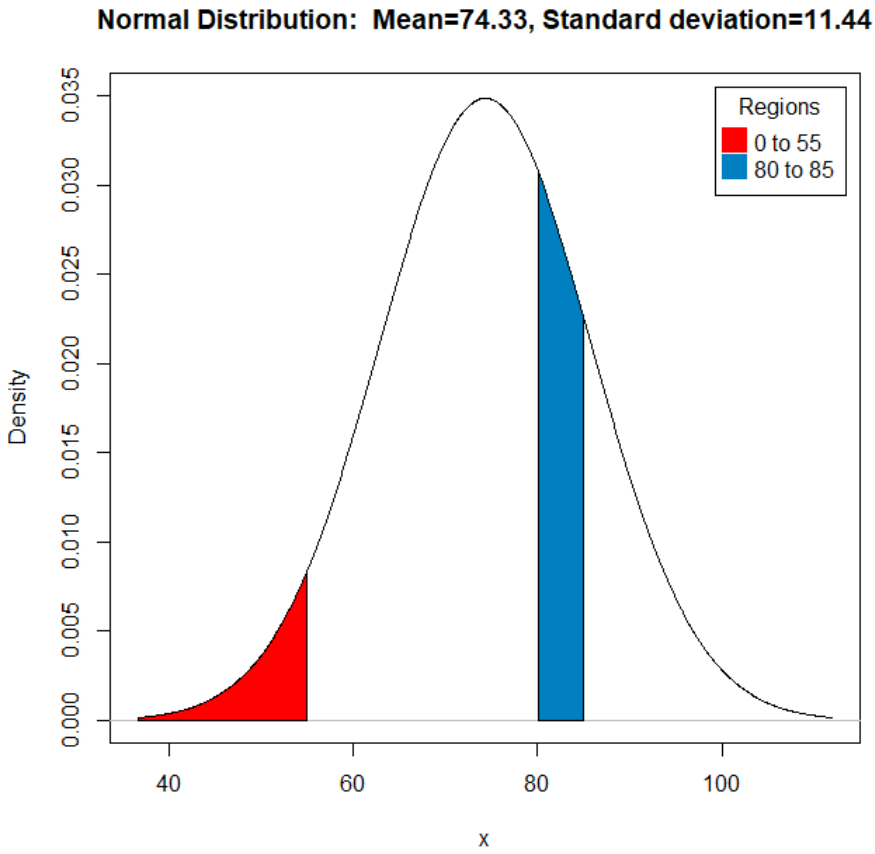
Si consultem qualsevol d'aquestes aplicacions, la resposta a les dues preguntes formulades abans és la següent:

p (pes < 55 kgs.) = 0,046

p (80 < pes < 85) = 0,135

Si utilitzem l'aplicació R, en el primer cas el resultat es pot obtenir directament, en el segon cal per una operació intermèdia, tal i com es veurà immediatament.

És important tenir present que qualsevol probabilitat a una distribució continua es representa mitjançat una àrea a la distribució corresponent. Per exemple, per a les dues probabilitats que s'han calculat, les àrees corresponents són:



Com s'indicava, en el cas de la primera pregunta plantejada, p (pes < 55), les aplicacions estadístiques normalment ens donen una resposta directa, ja que es pot demanar quina

és l'àrea situada a l'esquerra (per tant, per sota) de qualsevol valor donat. La densitat de probabilitat a l'esquerra del valor pes = 55 és de 0,046 (àrea vermella del gràfic).

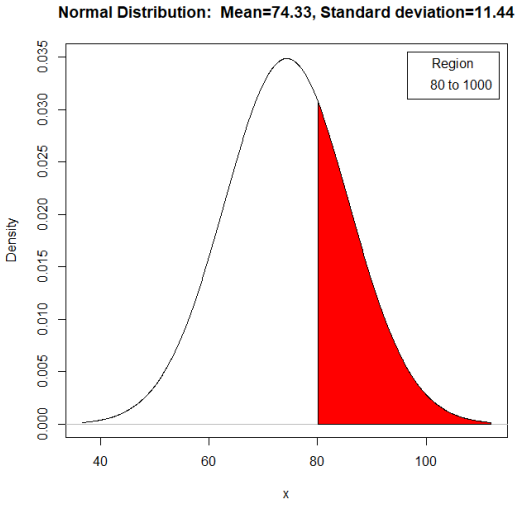
Pel que fa a la segona pregunta, s'ha de respondre en dues passes: En primer lloc, cal determinar l'àrea que queda a la dreta del valor 80 i també la que queda a la dreta del valor 85; posteriorment, la diferència entre les dues àrees ens oferirà el valor buscat. Els valors obtinguts són:

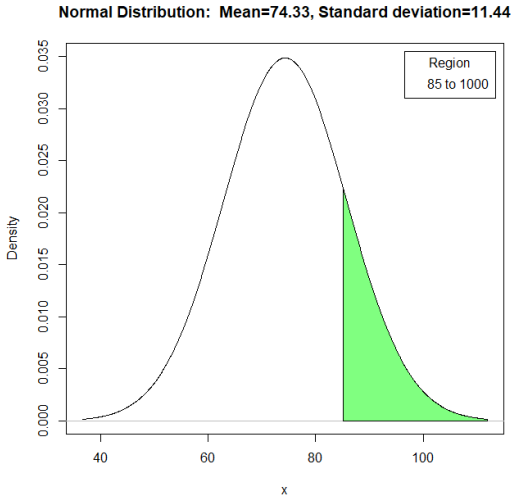
p (pes > 80) = 0,310

p (pes > 85) = 0,175

Per tant: p (80 < pes < 85) = 0,310 – 0,175 = 0,135

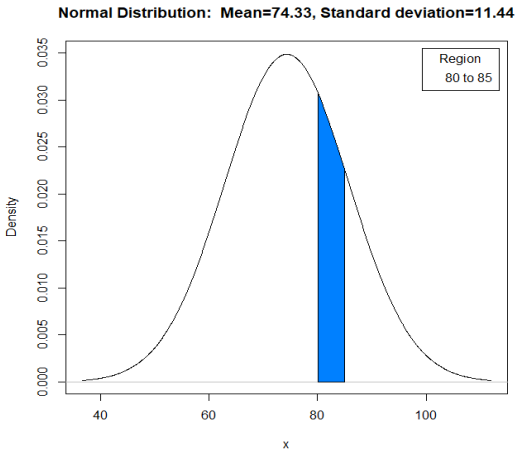
Gràficament:





Àrea des de 80 cap a la dreta: 0,310 Àrea de 85 cap a la dreta: 0,175

Si de l'àrea del primer gràfic es sostreu l'àrea del segon, el resultat és l'àrea compresa entre els valors 80 i 85, que resulta correspondre a una densitat de 0,135 (0,310-0,175):



Naturalment, aquesta operació es pot fer també calculant les àrees a l'esquerra dels dos valors i fent-ne la diferència.

Qualsevol distribució normal pot ser transformada en una *distribució normal tipificada o estàndard*. Una distribució normal estàndard és aquella que té una mitjana de 0 i una desviació típica de 1. Com es veurà en altres apartats, la distribució normal tipificada ens serà útil en diferents operacions estadístiques.

Per transformar una distribució normal qualsevol en una distribució estàndard, cal transformar cada puntuació en una puntuació tipificada o estàndard *z*, d'acord amb la fórmula següent:

$$z_i = \frac{x_i - \mu}{\sigma}$$

És a dir, cal fer la diferència entre cada puntuació i la mitjana i fer el quocient entre aquesta diferència i la desviació típica. Per exemple, en el cas de la distribució de pesos que venim treballant, per a la puntuació x_i = 67,2, el càlcul seria: (67,2-74,33)/11,44 = -0,623.

Si fem el mateix per totes les puntuacions de la mostra el resultat és (ordenat de menor a major):

Pes	z	Pes	z	Pes	z	Pes	z
49,5	-2,170	67,8	-0,571	74,1	-0,020	81,3	0,609
55,6	-1,637	68,3	-0,527	74,3	-0,002	83,5	0,801
56,3	-1,576	68,3	-0,527	74,6	0,024	83,6	0,810
58,3	-1,401	70,4	-0,343	75,4	0,094	86,5	1,064
59,9	-1,261	71	-0,291	76,3	0,172	88,3	1,221
61,4	-1,130	71,4	-0,256	77,3	0,260	88,7	1,256
62,8	-1,008	72,2	-0,186	78,3	0,347	92,4	1,579
63,5	-0,946	72,3	-0,177	79,1	0,417	92,4	1,579
67	-0,641	73,5	-0,073	80,1	0,504	93,4	1,667
67,2	-0,623	74	-0,028	81,2	0,600	101,7	2,392

Tenint com a referència que la mitjana d'aquesta mostra és de 74,33, podem veure com les puntuacions que es troben sota de la mitjana tenen puntuacions estandaritzades negatives i les que estan per sobre són positives, cosa lògica ja que la mitjana de la distribució normal estandaritzada és de 0. Naturalment, quan més lluny de la mitjana es trobi una puntuació, més gran serà el seu valor absolut, tant si és positiva com negativa.

Una de les utilitats que pot tenir la distribució normal tipificada és la de fer càlculs de les probabilitats associades a diferents valors de la distribució, ja que disposem d'unes taules calculades a aquest efecte. Recuperant algun dels exemples anteriors, si volem saber quina és la probabilitat d'obtenir una puntuació de pes inferior a 55, s'hauria de calcular el valor z corresponent i, posteriorment, consultar les taules de la distribució normal tipificada per determinar l'àrea que queda a l'esquerra d'aquesta puntuació. Cal

treballar amb precaució amb les taules de la distribució normal tipificada, ja que es poden presentar de diferents formes. Per aquest motiu, i també per comoditat i precisió, actualment el més habitual i aconsellable és fer els càlculs corresponents amb qualsevol aplicació estadística. No obstant això, es veurà més endavant que la distribució normal tipificada té altres utilitats, i per això s'hi ha fet referència aquí.

. Altres distribucions contínues de probabilitat

Al marge de la distribució normal, hi ha altres distribucions contínues de probabilitat que són molt importants per al treball estadístic. Entre elles cal destacar la distribució t de Student i la distribució chi-quadrat les quals, com es veurà als següents apartats, són molt importants per a determinades operacions estadístiques. Concretament, la distribució t de Student serà útil per a la comparació de dues mitjanes, mentre que la distribució chi-quadrat ens ajudarà per portar a terme la comparació entre proporcions. També la distribució F de Fisher-Snedecor ens resultarà útil a l'hora de comparar dues variàncies. Com es veurà, alguna d'aquestes distribucions (concretament la distribució *t* de Student) són força semblants a la normal, mentre que altres són bastant diferents. En tot cas, es farà una breu referència a aquestes distribucions en el moment en què calgui utilitzar-les.

4. Mostratge estadístic. Distribucions mostrals i estimació de paràmetres. Introducció a la prova d'hipòtesis

Com s'ha indicat anteriorment, un objectiu fonamental de l'estadística inferencial és el d'extreure conclusions vàlides per a una població a partir d'una o diverses mostres que se n'extreguin.

A grans trets, hi ha tres elements fonamentals a tenir presents quan parlem d'estadística inferencial:

. El procés de mostratge: Quina és la població de referència?. Com s'ha extret una determinada mostra d'aquesta població?. Es pot considerar la mostra com a representativa?.

. Les estimacions puntuals i per interval. Suposem que la ingestió mitjana de kilocalories en una determinada mostra de *n* persones és de 2750 kcal/dia. Podem pensar que la mitjana a la població és també de 2750 (estimació puntual) però, de forma més realista, podem dir també que la ingestió a la població estarà, per exemple, entre 2600 i 2900 kcal amb una certa probabilitat (estimació per intervals). Un concepte fonamental en relació amb això és el de *distribució mostral de l'estadístic*.

. La prova d'hipòtesis. Per exemple: hi ha diferència en el nivell mitjà de colesterol entre homes i dones?. Hi ha relació entre nivell d'estudis i hàbits d'alimentació?. És més àcid el producte A que el producte B?. Hi ha diferència de productivitat entre dos terrenys diferents?. Aquestes i moltes altres qüestions es poden respondre (sempre en termes probabilístics) mitjançant les *proves estadístiques de significació*.

4.1 El mostratge

Com s'ha indicat al primer apartat d'aquest text, el concepte bàsic relatiu al mostratge és el de *representativitat*. Evidentment, per extrapolar els resultats d'una mostra a la seva població d'origen és necessari que aquesta mostra sigui, tant com sigui possible, representativa de la població de referència, és a dir, ha de compartir-ne les característiques principals. Per aconseguir això hi ha dos elements fonamentals: l'ús de procediments basats en l'*aleatorització* i l'assoliment d'una *grandària mostral* suficient:

El procediment aleatori més directe és escollir un conjunt de casos entre la població d'origen mitjançant, per exemple, una sèrie de números aleatoris o qualsevol procediment de sorteig o similar. Des d'un punt de vista pràctic, però, això no sempre és fàcil o possible. Alguns exemples són els següents:

. Si es vol realitzar una enquesta telefònica sobre hàbits alimentaris, és evident que només es podrà comptar amb les respostes d'aquelles persones que vulguin contestar. Es pot considerar això com una mostra representativa, o bé podem pensar que les persones que responen tenen algun tipus de motivació especial, o qualsevol característica que les allunyi del conjunt de la població?.

. De vegades un investigador només pot treballar amb aquells casos que pot tenir al seu abast. Per exemple, en un estudi sobre anorèxia i bulímia és possible que només es pugui comptar amb els o les pacients que acudeixin al centre de salut on treballi l'investigador. Es pot afirmar, per exemple, que les persones tractades d'anorèxia o bulímia a l'hospital Clínic de Barcelona són

representatives del conjunt de persones afectades a tot Catalunya?. És possible que responguin, per exemple, a perfils sociodemogràfics concrets que poden ser diferents dels trobats en altres hospitals de Catalunya?. Aquest tipus de "mostratge", on es treballa amb els casos "més propers" o accessibles, s'anomena de vegades *mostratge incidental*, i és evident que crea molts dubtes sobre la representativitat de la mostra

. Si demanem persones voluntàries per fer una prova de tast d'aliments d'entre els estudiants d'una Universitat, i oferim a canvi una recompensa econòmica o de qualsevol altre tipus, és lògic pensar que les persones que responguin a aquesta convocatòria puguin estar especialment motivades per la recompensa oferida. Això pot fer que difereixin en algun aspecte del conjunt de la població?

Hi ha diverses variants del mostratge aleatori, amb la idea d'assegurar millor la representativitat de la mostra. Una pràctica habitual consisteix en assegurar la presència a la mostra de subgrups rellevants, mitjançant diferents variants de mostratges estratificats. Per exemple, en molts estudis és important assegurar la presència tant d'homes com de dones, o de les diferents franges d'edat, etc. Si, a més, coneixem la distribució aproximada d'aquestes característiques a la població, podem intentar mantenir-la a la mostra. Per fer això podem definir els subgrups que ens interessen i assegurar la seva presència proporcional a la mostra. Per exemple, podem assegurar-nos que hi ha hagi un 50% d'homes i un 50% de dones o, pel que fa a la variable edat, podem procurar que la presència de les diferents franges d'edat sigui similar a la seva distribució a la població. Per fer això hem de definir primer els sub-grups o estrats, decidir quina proporció correspon a cadascun d'ells a partir del coneixement que tinguem de la població d'origen, i posteriorment seleccionar a l'atzar els casos a la població, dins de cada estrat. En aquest cas estem realitzant un mostratge estratificat proporcional.

Altres variants de mostratge són el mostratge per conglomerats, o el de rutes aleatòries, tots dos basats en la selecció de zones geogràfiques. Per exemple, si volem fer un estudi a nivell de Catalunya, podem optar per escollir certes zones geogràfiques (lògicament, diverses entre sí) i fer un mostratge aleatori dins de cadascuna d'elles. Les zones geogràfiques poden ser comarques, ciutats o qualsevol altra agrupació, mantenint sempre un criteri de diversitat i representativitat de les zones escollides. Aquest procediment simplifica molt les coses, ja que ens evitar haver de treballar amb tot el territori i ens permet concentrar-nos en zones concretes.

L'altre element important a l'hora de valorar la representativitat d'una mostra és la seva grandària, és a dir, el número de casos que inclou. Normalment representarem com a *n* el número de casos de la mostra, i com a *N* el número de casos total de la població, en cas que el coneguem. És evident que quan més s'apropi *n* a *N*, més representativa serà la mostra, sempre que el mostratge s'hagi fet correctament. Ara bé, també és molt clar que a la pràctica és difícil treballar amb mostres molt grans, per raons simplement de temps i recursos. No hi ha una norma rígida sobre la grandària de les mostres, i sovint es treballa amb els casos que és possible assumir, sense fer més consideracions. Cal tenir en compte, però, que la grandària de la mostra és important pel que fa a la precisió de les estimacions i la probabilitat d'error que tenim al fer-les. Com veurem una mica més endavant, és possible fer un càlcul de la grandària desitjable de la mostra en funció del nivell de precisió que vulguem per a les nostres estimacions.

A la pràctica, és habitual trobar estudis amb grandàries mostrals molt diverses, des de mostres petites (per exemple, n<30) fins a altres molt grans, de milers de casos.

En qualsevol cas, l'important és, quan es dissenya un estudi o quan s'interpreten els resultats d'un estudi anterior, tenir clar quina tècnica de mostratge es vol utilitzar o s'ha utilitzat i quina

és la representativitat de la mostra obtinguda, per tal de valorar fins a quin punt i amb quines limitacions els resultats poden ser generalitzables. No es tracta únicament d'aplicar les tècniques estadístiques adients, sinó d'interpretar correctament els seus resultats i valorar fins a quin punt ens poden portar a fer afirmacions de tipus general, més enllà de la mostra estudiada.

4.2 Estimacions puntuals i per interval

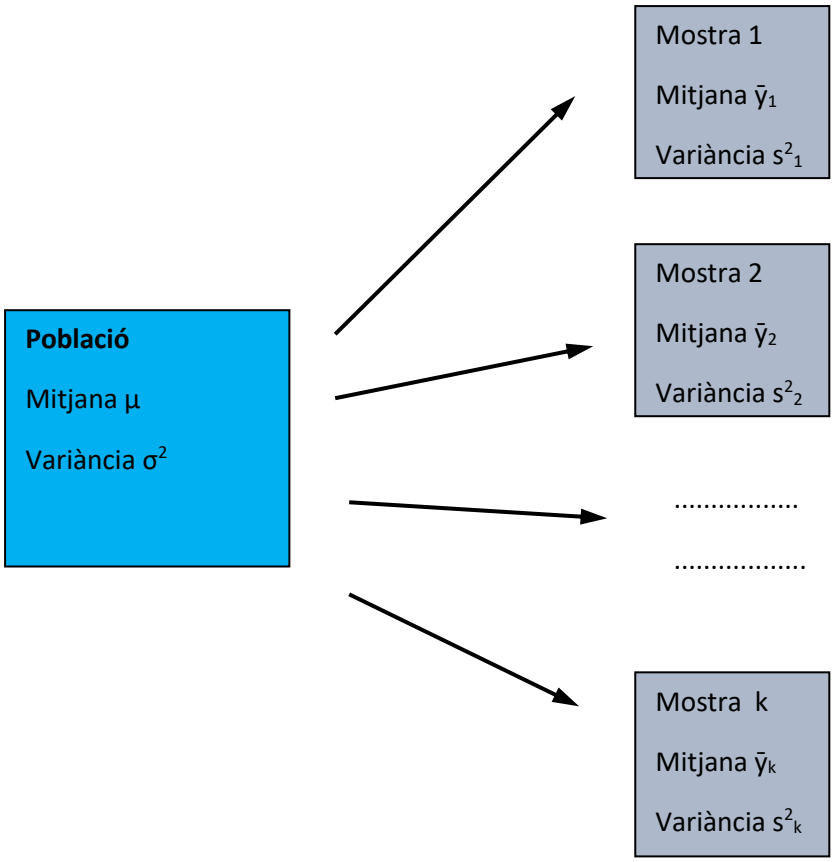
Si disposem d'una mostra que considerem prou representativa, ens podem plantejar obtenir informació sobre les característiques de la població (*paràmetres*) a partir de la informació obtinguda sobre la mostra (*estadístics*). Hem distingit també entre estimació puntual i estimació per intervals. En aquest apartat es desenvoluparan els conceptes més importants respecte dels dos tipus d'estimació.

L'estimació puntual és aquella que es fa utilitzant únicament el valor de l'estadístic mostral, sense establir cap mena d'interval de valors. Si, com s'indicava abans, la mitjana d'ingesta de kcal diàries en una mostra de subjectes és de 2750, podem concloure que aquesta és una aproximació a la mitjana poblacional, tot i admetre que es comet un error més o menys gran. És a dir, assumim que $\mu = \bar{y}$, essent μ la mitjana poblacional i $\bar{y}$ la mitjana mostral. Com es pot veure, habitualment els paràmetres són designats amb lletres gregues i els estadístics amb lletres llatines.

Si en lloc d'una mostra en tinguéssim cinc, o deu, veuríem immediatament que les mitjanes de les diferents mostres habitualment no coincideixen. En aquest cas podríem fer una mitjana global per al conjunt de casos, i tindríem una aproximació més afinada a la mitjana poblacional.

El fet que les mitjanes obtingudes amb diferents mostres provinents de la mateixa població no siguin coincidents ens porta a un concepte fonamental, que és el de *distribució mostral de l'estadístic*. La distribució mostral no és altra cosa que la distribució de valors que s'obtenen d'un determinat estadístic a les diferents mostres disponibles o possibles. Si, per exemple, per a una determinada població nosaltres obtinguéssim (teòricament, és clar) totes les mostres possibles de grandària n=40, per a cadascuna d'aquestes mostres tindríem una mitjana, i els valors d'aquestes mitjanes seguirien una certa distribució. La mitjana de totes les mitjanes mostrals ens permetria estimar la mitjana poblacional.

Gràficament, podem representar aquesta idea de la forma següent, considerant els paràmetres mitjana i variància, i suposant que hem agafat *k* mostres:



Lògicament, quan més gran sigui la grandària de les diferents mostres, més s'aproparan tant la mitjana com la variància de cadascuna d'elles als corresponents paràmetres poblacionals, i més semblants seran les diferents mitjanes i variàncies mostrals entre sí. Tècnicament, aquesta idea s'expressa en el *teorema del límit central*. Seguint amb l'exemple de les mitjanes, el teorema del límit central ens indica que en mostres aleatòries de grandària *n*, la mitjana de les mostres fluctua entorn de la mitjana poblacional amb un *error estàndard* de $\sigma/\sqrt{n}$, on σ és la desviació típica de la variable a la població. A mesura que *n* augmenta, la distribució mostral de mitjanes es concentra cada vegada més entorn de la mitjana poblacional i els seus valors s'ajusten més a una distribució normal. L'error estàndard no és altra cosa que la desviació típica de la distribució mostral.

El teorema del límit central és aplicable a altres estadístics, com és el cas de la variància, la proporció de casos a les diferents categories en una variable qualitativa, etc., tot i que lògicament el càlcul de l'error estàndard és diferent per a cadascun d'ells.

Tècnicament, qualsevol estimació que es faci d'un paràmetre poblacional ha de reunir tres característiques fonamentals:

. No ha de ser esbiaixat. Això vol dir que l'esperança matemàtica de la distribució mostral si utilitzem totes les mostres possibles ha de ser igual al paràmetre poblacional, com ja s'ha explicat en el cas de les mitjanes. El mateix és aplicable a l'estimació de qualsevol altre paràmetre.

. Ha de ser consistent. Un estimador és consistent quan la diferència entre l'estimador i el paràmetre poblacional és cada vegada menor a mesura que incrementem la grandària mostral.

Naturalment, el cas límit és aquell en el qual treballem amb tota la població i tenim la possibilitat de calcular directament els paràmetres que ens interessin.

. Ha de ser eficient. Un estimador eficient és aquell que presenta un error estàndard mínim. Això vol dir que la distribució mostral de l'estadístic estigui molt concentrada entorn del corresponent paràmetre poblacional que es vol calcular. Lògicament, una forma d'aconseguir això és, novament, incrementar en el que es pugui la grandària de la mostra, ja que el l'error estàndard és igual, en el cas de les mitjanes, a $\sigma/\sqrt{n}$, però a igualtat de grandàries mostrals hi ha estimadors més eficients que altres.

. Ha de ser suficient. Això vol dir que no cal obtenir cap altre estimador que permeti millorar l'estimació feta, és a dir, l'estimació obtinguda és la millor possible en les condicions donades.

El compliment d'aquestes condicions i l'ús de diferents procediments d'estimació (bàsicament els de mínims quadrats i màxima versemblança) han permès als teòrics de l'estadística establir quins són els millors estimadors dels paràmetres poblacionals. Per exemple, en el cas de les variàncies, com s'havia indicat al parlar de mesures de dispersió a l'apartat 2, el millor estimador de la variància poblacional és la variància mostral obtinguda amb l'expressió:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Aquesta expressió, amb el denominador *n-1* en lloc del més natural *n*, és l'estimador del paràmetre poblacional que millor compleix les condicions indicades anteriorment. Per tant, podem dir que:

$$\hat{\sigma}^2 \;= s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Aquí l'expressió $\hat{\sigma}^2$ indica l'estimació del paràmetre poblacional σ^2.

. Estimació per interval d'una mitjana: Interval de confiança

El concepte de distribució mostral d'un estadístic i el teorema del límit central ens permeten passar de l'estimació puntual dels paràmetres poblacionals a fer-ne una estimació per intervals. En una estimació per intervals el que fem és calcular entre quins valors es troba el paràmetre poblacional amb una determinada probabilitat, és a dir, establim un *interval de confiança*. Podem parlar, per exemple, d'un interval de confiança del 95%, que ens indicarà entre quins valors es troba, amb una probabilitat de 0,95, el paràmetre poblacional, assumint per tant una probabilitat d'error de 0.05, o del 5%. Lògicament, quan major sigui el nivell de confiança amb què volem treballar, més ampli serà l'interval obtingut. Sempre un interval amb p=0,95 serà més ampli que l'interval amb p=0,90 per a una mateixa variable i amb una grandària mostral fixa. Per la mateixa lògica, quan més gran sigui la grandària de la mostra, més estrets seran els intervals obtinguts, ja que podrem fer l'estimació amb major precisió.

L'interval de confiança per a una mitjana s'obté a partir de l'expressió següent:

$$I.C. = \bar{y} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

On ȳ és la mitjana de la mostra, α és el nivell o risc d'error que s'admet en l'interval de confiança, Z és la puntuació normal tipificada que correspon al valor de α, n és el número de casos de la mostra, i σ és la desviació típica poblacional.

En el cas habitual en què la desviació típica de la població sigui desconeguda, podem treballar amb la seva estimació, és a dir, amb la desviació típica de la mostra:

$$I.C. = \bar{y} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Ara bé, cal dir que aquesta aproximació només és vàlida (i encara amb reserves) quan la grandària de la mostra és gran, i en tot cas sempre superior a 30 casos. Com veurem més endavant, hi ha una alternativa que ens serà útil quan la mostra sigui inferior a 30 casos, i que de fet podrem aplicar de forma general a qualsevol situació.

Suposem que s'utilitza novament la distribució de pesos que ja s'ha vist en apartats anteriors. Cal recordar que es disposava d'una mostra de 40 persones, que la seva mitjana era de 74,33 kgs. i la seva desviació típica era de 11,44. Imaginem que es vol establir l'interval de confiança corresponent a p=0.95. Per tant, la probabilitat α de cometre un error al fer l'estimació per interval serà α = 1 – 0.95 = 0.05. Com que un error es pot cometre tant per la banda alta com per la banda baixa (la mitjana poblacional podria estar per sobre o per sota de l'interval de confiança), aquest valor es "reparteix" entre els dos extrems de la distribució, de manera que α/2 = 0,025. L'interval de confiança, doncs, tindrà una probabilitat 0,95 d'incloure la mitjana poblacional, una probabilitat 0,025 de que la mitjana poblacional estigui per sota de l'interval, i una probabilitat també de 0,025 de que estigui per sobre.

L'única dificultat radica en establir el valor de $Z_{\alpha/2}$. Cal recordar que partim de la base de que la distribució mostral de les mitjanes s'ajusta a una distribució normal, amb mitjana μ i desviació típica (error estàndard) igual a $\sigma/\sqrt{n}$. Com ja s'ha indicat anteriorment, qualsevol distribució normal pot ser tipificada, és a dir, convertida en una distribució normal amb mitjana 0 i desviació típica 1. Les puntuacions tipificades obtingudes són representades amb la lletra Z. Només cal establir, doncs, quina és la puntuació tipificada que deixa per sota de sí el 2,5% dels casos (p = 0,025) i quina deixa per sobre el mateix percentatge. Aquesta puntuació resulta ser (d'acord amb les taules de la distribució normal tipificada), z = 1,96. Per tant, l'interval de confiança que es busca és:

$$I.C. = 74{,}33 \pm 1{,}96 \cdot \frac{11{,}44}{\sqrt{40}} = (70{,}78 - -77{,}88)$$

Podem concloure doncs que, amb una probabilitat p=0,95, la mitjana de pes de la població estarà entre els valors de 70,78 i 77,88 kgs.

Si es vol treballar amb un nivell de confiança diferent, lògicament el valor Z corresponent també serà diferent. Els més habituals són els següents:

Nivell de confiança	Valor Z
90%	1,64
95%	1,96
99%	2,57
99,9%	3,29

És fàcil veure com a mesura que incrementem el nivell de confiança, l'amplitud de l'interval és cada vegada més gran. Per exemple, si en lloc de treballar amb un nivell de confiança del 95% volem fer-ho amb un nivell del 99%, l'interval corresponent serà (utilitzant el valor Z de 2,57) igual a (69,68 -- 78,98).

En tot cas, pràcticament qualsevol aplicació estadística pot calcular directament l'interval de confiança que es busqui, de forma que no cal recórrer a les taules de la distribució normal tipificada.

Si coneixem la desviació típica de la població és possible calcular també quin número de casos necessitarem per estimar la mitjana amb la precisió que considerem necessària. Com hem vist anteriorment, l'interval de confiança del 95% era de (70,78 – 77,88). Considerant que la mitjana de la mostra és de 74,33, l'interval es pot representar també com a: 74,33 ± 3,55. Naturalment, el valor de 3,55 és el resultat del càlcul $Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}$. Aquesta "desviació" respecte de la mitjana (3,55) s'anomena també *hemi-interval*, i se sol representar amb la lletra *e*.

Suposem, per exemple, que en lloc de tenir un hemi-interval de ± 3,55 volem fer una estimació més precisa i obtenir un interval de confiança més "estret". Imaginem que volem arribar fins a un hemi-interval *e* = 2.

Considerant que *e* = $Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}$, està clar que podem dir que $n = \left(Z\alpha_{/2} \cdot \frac{\sigma}{e}\right)^2$

Com que no coneixem la desviació típica de la població, ens veiem en l'obligació de substituir-la per la seva estimació, és a dir, la desviació típica de la mostra. Per tant, per aconseguir un valor *e* de 2 necessitarem:

$$n = \left(1{,}96 \cdot \frac{11{,}44}{2}\right)^2 = 125{,}69$$

És a dir, necessitaríem aproximadament 125 subjectes per establir un interval de confiança del 95% amb un hemi-interval *e* = 2. Cal dir, en tot cas, que aquesta és una aproximació discutible, ja que realment no coneixem la desviació típica de la població i n'hem fet una estimació, però en tot cas serveix per mostrar clarament que si volem més precisió al fer una estimació per interval de la mitjana poblacional hem d'incrementar necessàriament el número de casos de la mostra.

Com s'ha assenyalat abans, l'aproximació basada en la distribució normal només és utilitzable quan es conegui la desviació típica poblacional o, en cas contrari, quan el número de casos de la mostra sigui gran. Quan el número de casos de la mostra és inferior a 30, la distribució mostral de mitjanes s'ajusta millor a una altra distribució contínua, concretament la distribució *t* de Student-Fisher. La distribució *t* es relaciona molt amb la normal, i descriu la distribució mostral de les mitjanes amb mostres petites (n<30) i quan la desviació típica de la població és desconeguda (que és gairebé sempre). Per tant, en el cas de n<30 l'expressió utilitzada anteriorment es modifica de la forma següent:

$$I.C. = \bar{y} \pm t_{(\nu,\alpha/2)} \cdot \frac{s}{\sqrt{n}}$$

Com es pot veure, l'única diferència és que la distribució de referència no és la distribució normal tipificada (puntuacions *z*), sinó la distribució *t* de Student. El valor de *t* que cal utilitzar ve donat, com abans, per α/2, i per un segon valor, representat per *v* i que s'anomena *graus de llibertat* de la distribució. En el cas de la distribució *t* de Student, *v* = *n-1*, és a dir, el número de casos de la mostra menys un. Amb aquests dos valors (*v*,α/2) es pot anar a la taula de la distribució d'Student i obtenir el valor corresponent que caldrà incloure a la fórmula anterior. Si, per exemple, treballéssim amb una mostra de n=25 casos i volem establir l'interval de confiança amb p=0,95 per a la mitjana, caldria anar a la taula de la distribució d'Student amb *v* = *n - 1* = 24, i α/2 = 0,025. Novament, les aplicacions estadístiques ens permeten evitar aquest pas i ens ofereixen directament el càlcul corresponent.

Tot i que, com s'ha indicat, el recurs a la distribució *t* de Student-Fisher és molt adient en els casos en què n<30, això no impedeix que es pugui utilitzar també amb mostres més grans. En realitat, el seu ús és indicat, sigui quina sigui la grandària de la mostra, quan no coneguem la desviació típica poblacional, és a dir, gairebé sempre. Moltes aplicacions estadístiques utilitzen de forma habitual i per defecte la distribució de Student. Els resultats obtinguts no són habitualment gaire diferents dels que es poden trobar a partir de la distribució normal.

D'acord amb això, si apliquem aquest procediment a la distribució de pesos (tot i que la grandària de mostra sigui de 40), per obtenir l'interval de confiança amb p=.95 caldria utilitzar els valors (*v=39*,α/2=0.025). D'acord amb la taula de la distribució *t* de Student, el valor *t* corresponent és de 2,0227. Per tant:

$$I.C. = 74{,}33 \pm 2{,}0227 \cdot \frac{11{,}44}{\sqrt{40}} = (70{,}67 - -77{,}99)$$

Com es pot veure, el resultat obtingut és molt similar al que havíem trobat utilitzat la distribució normal tipificada (z).

. Estimació per interval d'una proporció

Suposem que es vol implantar un mecanisme de seguiment de l'acceptació d'un determinat producte alimentari entre els consumidors. Imaginem que el número de vendes mensuals del producte en qüestió és aproximadament de 300.000 unitats. Suposem també que s'estableix un sistema d'enquestes a determinats supermercats, i es demana als clients que classifiquin el producte com a Excel·lent, Bo, Regular o Deficient, a més de fer també algunes preguntes complementàries. Es fan les enquestes durant una setmana fins a obtenir 300 respostes. Suposem que els resultats obtinguts són els següents, amb expressió també de les proporcions corresponents:

Categories	**Excel·lent**	**Bo**	**Regular**	**Deficient**	**Total**
Nº de casos	144	72	39	45	300
Proporcions	0,48	0,24	0,13	0,15	1

Aquests resultats s'han obtingut amb una mostra relativament gran, però en tot cas ens pot interessar establir quines són les proporcions que podem esperar trobar a la població total de consumidors del producte. Com fèiem en el cas de la mitjana, podem establir també un interval de confiança per a les proporcions poblacionals, de manera que a partir de les proporcions mostrals *p* puguem fer una aproximació a les proporcions poblacionals π.

Suposem que ens interessa especialment la proporció de persones que consideren el producte com a deficient. Per fer l'estimació per interval, el primer que cal es dicotomitzar els resultats. És a dir, a partir dels resultats de la mostra definim la proporció p(Deficient) = 0,15, i la proporció p(No deficient) = 0,85. Naturalment, aquest últim resultat prové de reunir totes les valoracions que no són de Deficient, és a dir, de reunir els resultats Excel·lent, Bo i Regular.

Sota determinades condicions, que es veuran més endavant, la distribució mostral de l'estadístic proporció segueix una distribució normal amb un error estàndard igual a $\sqrt{\frac{p\cdot(1-p)}{n}}$ D'acord amb això, el càlcul de l'interval de confiança és semblant al que es feia en el cas de les mitjanes. L'expressió corresponent és la següent:

$$I.C.= \ p \ \pm \ Z\alpha_{/2} \cdot \ \sqrt{\frac{p \cdot (1-p)}{n}}$$

Suposem que volem calcular l'interval amb un nivell de confiança del 95%, que és el més habitual. Cal recordar que el valor Z corresponent és de 1,96. Per tant, en el nostre cas:

$$I.C.= \ 0{,}15 \ \pm \ 1{,}96 \ \cdot \ \sqrt{\frac{0{,}15 \cdot 0{,}85}{300}} \ = (0{,}109 - -0{,}191)$$

Per tant, amb un nivell de confiança del 95% (i, en conseqüència, amb una possibilitat d'error del 5%), la proporció poblacional de persones que consideren deficient el producte (π) està entre 0,109 i 0,191 o, si es prefereix, entre el 10,9 i el 19,1%.

Per tal de poder utilitzar correctament aquest procediment cal que es compleixin les condicions següents:

(π_{inf} . n) ≥ 5 , (1- π_{inf} . n) ≥ 5

(π_{sup} . n) ≥ 5 , (1- π_{sup} . n) ≥ 5

π_{inf} i π_{sup} són es límits inferior i superior de l'interval de confiança obtingut. Per tant, π_{inf} = 0,109 i, òbviament, 1- π_{inf} = 1 – 0,109 = 0,891. En el cas del límit superior, π_{sup} = 0,191, i 1- π_{sup} = 1 – 0,191 = 0,809. Per tant, la comprovació de les condicions d'aplicació és:

(0,109 · 300) = 32,7 , (0,891 · 300) = 267,3

(0,191 · 300) = 57,3 , (0,809 · 300) = 242,7

Tots els resultats obtinguts són molt superiors a 5, per tant es compleixen les condicions d'aplicació i podem considerar l'interval de confiança obtingut com a vàlid.

Com passava en el cas de l'estimació de la mitjana poblacional, també es pot determinar quina és la grandària de mostra necessària per obtenir un interval de confiança amb un nivell de precisió determinat. En el cas de les proporcions, l'hemi-interval *e* ve donat per l'expressió:

$$e = \ Z\alpha_{/2} \cdot \ \sqrt{\frac{p \cdot (1-p)}{n}}$$

Per tant, per a un hemi-interval donat, la grandària necessària de la mostra seria:

$$n = (Z_{\alpha/2})^2 \cdot \frac{p(1-p)}{e^2}$$

En el cas que s'ha presentat, l'hemi-interval resultant per a l'interval de confiança del 95%, amb una mostra de 300 enquestes, ha estat de 0,041, de manera que la proporció estimada ha estat: $0{,}15 \pm 0{,}041 = (0{,}109 - -0{,}191)$. Si es vol un major nivell de precisió caldrà incrementar la grandària de la mostra. Per obtenir, per exemple, un hemi-interval de 0,02, caldria disposar del número d'enquestes següent:

$$n = 1{,}96^2 \cdot \frac{0{,}15 \cdot 0{,}85}{0{,}02^2} = 1219{,}835$$

Per tant, necessitarem aproximadament unes 1220 enquestes per incrementar la precisió de l'estimació fins a un hemi-interval de 0,02.

4.3 Introducció a la prova d'hipòtesis

L'estadística ens permet contrastar o provar diferents tipus d'hipòtesis quan disposem de dades que mostren una variabilitat important. En apartats anteriors s'han plantejat alguns exemples, i se'n poden plantejar molts altres: Hi ha diferències en el nivell de colesterol en sang entre homes i dones?, hi ha diferència en el contingut en sal de dos productes alimentaris?, la proporció de persones amb sobrepès és diferent a Europa i a Amèrica?, etc, etc. És fàcil veure la gran diversitat de qüestions que es poden plantejar si disposem de les dades corresponents, sempre que aquestes dades s'hagin obtingut amb les garanties necessàries. Com veurem, la forma més habitual de respondre aquestes qüestions és mitjançant el que anomenem *proves estadístiques de significació*.

En línies generals, es pot parlar de tres grans tipus d'hipòtesis estadístiques, que porten associades les corresponents proves de significació:

. Hipòtesis de conformitat. S'ajusten els resultats a una teoria o suposició prèvia?. Per exemple: a partir d'estudis previs podem tenir la suposició que el percentatge d'obesos a la població general és superior al 20%. Això serà cert també per a una sub-població concreta? (per exemple, això és cert per a la població universitària de Catalunya?).

. Hipòtesis d'homogeneïtat. Es refereixen a la suposició que dues o més mostres provenen d'una mateixa població, que suposadament és homogènia pel que fa a l'aspecte que ens interessa. Per exemple: la proporció de persones amb Rh negatiu és la mateixa per homes i dones?. En altres paraules: homes i dones constitueixen una població homogènia pel que fa a la presència de Rh negatius, o bé hi ha diferències entre aquests dos grups i, per tant, constitueixen dues sub-poblacions diferenciades en relació amb això?.

. Hipòtesis d'independència: Es refereixen a l'existència o no de relació entre dues o més variables. Per exemple, hi ha relació entre el nivell d'estudis i els hàbits alimentaris?. Això també és aplicable quan es produeix una intervenció deliberada: per exemple, dos sistemes de control de qualitat són igualment eficaços per detectar productes defectuosos?; dues dietes diferents tenen la mateixa eficàcia per mantenir els nivells de glucosa de pacients amb diabetis?, etc.

Cal insistir, doncs, en que per contestar qualsevol pregunta mitjançant les tècniques estadístiques cal plantejar-se una hipòtesi de partida. Per exemple, suposem que volem comparar el contingut en sal de dues marques de galetes (A i B). Per esbrinar això, i considerant que el contingut en sal d'un mateix producte pot mostrar un cert nivell de variabilitat (en aquest cas segurament petit, però existent), podríem agafar i analitzar una mostra de productes A i una altra de productes B, i establir en cada cas el nivell de sal. Suposem que la mitjana del contingut de sal del producte A és de 1,70 grs per cada 100 grams de producte, i que la mitjana corresponent per al producte B és de 1,82 grs/100 grs de producte. Hi ha una diferència real en el contingut de sal dels dos productes, o això només és un resultat aleatori produït per la variabilitat de les dades?. Això ens porta a definir dues hipòtesis contraposades, que són fonamentals en el treball estadístic: La *hipòtesi nul·la (H_0)* i la *hipòtesi alternativa (H_1).* Les dues hipòtesis són fàcils d'enunciar:

H_0: La diferència en les mitjanes de sal (1,70 vs 1,82) és producte de l'atzar i no ens indica una diferència real entre els dos productes

H_1: Hi ha una diferència real i consistent entre els dos productes pel que fa al seu contingut de sal

Cal indicar que la hipòtesi alternativa pot ser *unilateral* o *bilateral*. Tal i com està enunciada, la hipòtesi alternativa que s'ha proposat en aquest cas és bilateral, ja que la pressuposició és que el contingut de sal entre els dos productes pot ser diferent, però sense suposar una direcció concreta per aquesta diferència (per tant, no pressuposem que hagi d'haver-hi més sal al producte A que al B, ni al contrari). Si donem una direcció definida a la hipòtesi, llavors estem en el cas unilateral: si, per exemple, suposem per alguna raó que el contingut de sal de A hauria de ser superior al de B, la nostra hipòtesi seria A > B; si fos al contrari, la nostra hipòtesi seria A < B. En qualsevol dels dos casos, la nostra hipòtesi té una *direcció* definida, i per tant és unilateral.

Les proves estadístiques de significació ens ofereixen una forma de contrastar si es compleix o no la hipòtesi nul·la, sempre en termes probabilístics. *El que fa una prova de significació és calcular la probabilitat de que els resultats obtinguts puguin ser producte de l'atzar*. Si aquesta probabilitat és molt petita, podem pensar que el resultat obtingut no és explicable per atzar i, per tant, rebutjar la hipòtesi nul·la i acceptar la hipòtesi alternativa que haguem proposat. Si, pel contrari, la probabilitat de que el resultat sigui explicable per atzar és alta, podem concloure que no hi ha una diferència o un efecte real i acceptar la hipòtesi nul·la.

És important veure que l'ús de l'estadística només té sentit quan es produeix variabilitat en els resultats. Si tots els productes A tinguessin *exactament* el mateix contingut en sal (suposem que sigui de 1,70), i tots els productes B, per la seva part, tinguessin també un contingut de sal totalment idèntic (diguem 1,82), llavors n'hi hauria prou amb agafar una galeta A i una galeta B i fer l'anàlisi del seu contingut en sal. Amb això hi hauria base suficient per dir que el contingut en sal del producte B és superior al del A, ja que hauríem de pensar que *totes* les galetes A tenen un contingut de sal de 1,70, i *totes* les galetes B tenen un contingut de 1,82. En la mesura en que hi hagi variabilitat, en canvi, no podem limitar-nos a agafar només una galeta de cada tipus, ja que la diferència que puguem observar entre elles podria ser simplement casual i no extrapolable a la resta d'unitats d'ambdós productes. El fet que *una* galeta A tingui un contingut de sal de 1,70 i *una* galeta B tingui un contingut de 1,82 no seria llavors suficient per assegurar que un producte té sistemàticament més sal que l'altre.

Tornant a la lògica de l'estadística, com que disposem de les mitjanes de resultats per a les mostres de cadascun dels dos productes, podríem utilitzar una *prova de comparació entre dues mitjanes*. Com veurem a l'apartat següent, aquesta prova està basada en la distribució t de Student, i ens permetrà calcular un estadístic t i una probabilitat associada. Aquesta probabilitat és la que hem d'interpretar, i la que ens servirà de fonament per prendre una *decisió estadística*.

Suposem que es realitza la prova t de Student corresponent, i que la probabilitat obtinguda és p=0.014. La interpretació d'aquest resultat ha de ser automàtica i clara: D'acord amb les dades disponibles, la probabilitat de que la diferència entre les dues mitjanes obtingudes sigui deguda purament a l'atzar és de 0,014, és a dir, una probabilitat molt baixa. Si el resultat fos aquest, la conclusió lògica seria pensar que la diferència entre les dues mitjanes no és deguda a l'atzar i, per tant, que aquesta diferència és *estadísticament significativa*. Això ens fa pensar que existeix una diferència real en el contingut de sal dels dos productes.

Un problema evident d'aquesta forma de procedir és que necessitem un criteri per prendre aquesta decisió. A partir de quin valor de p hem de considerar que acceptem o rebutgem la hipòtesi nul·la?. A la majoria dels camps de treball, s'utilitza convencionalment com a valor límit el de 0,05. Per tant, un criteri de decisió habitual és:

$p \leq 0,05$	Rebutgem la hipòtesi nul·la i considerem que existeix una diferència significativa
$p > 0,05$	Acceptem la hipòtesi nul·la i atribuïm a l'atzar la diferència observada

Qualsevol decisió estadística està subjecta a error. Per exemple, en el cas citat anteriorment, a partir de la probabilitat obtinguda (p = 0,014), i aplicant el criteri que s'acaba d'indicar, s'ha arribat a la conclusió de que existeix una diferència estadísticament significativa entre les dues mitjanes comparades. Com ja s'ha indicat anteriorment, 0,014 és la probabilitat de que la diferència observada entre les dues mitjanes sigui deguda a l'atzar. Si, com hem fet, considerem que aquesta probabilitat és petita i podem rebutjar la hipòtesi nul·la, la probabilitat de cometre un error amb aquesta decisió és justament de 0,014.

De forma esquemàtica, les situacions que es poden donar en prendre una decisió estadística són les següents:

		Diferència real	
		Sí	No
Decisió estadística	Acceptar hipòtesi nul·la	Error tipus II (β)	Correcte
	Rebutjar hipòtesi nul·la	Correcte	Error tipus I (α)

Es poden cometre, doncs, dos tipus d'error:

. En un error tipus I, es declara com a significativa una diferència que no respon a una diferència real a la població. Podem dir, doncs, que es produeix una *falsa alarma*. La probabilitat de cometre aquest tipus d'error es representa com a α. A l'exemple anterior, α = 0,014.

. En un error tipus II, la prova estadística no és capaç de detectar una diferència real existent. Es comet, doncs, un error per *omissió*. La capacitat d'una prova estadística de fer correctament aquesta detecció s'anomena *potència* de la prova. La probabilitat de cometre un error tipus II es representa per β, i la potència de la prova és igual a (1-β).

La capacitat d'una prova estadística per detectar un determinat efecte depèn de diferents factors:

. De la magnitud de l'efecte que el vol detectar. Si un efecte és molt gran (per exemple, si la diferència en el contingut de sal dels productes A i B és molt gran), serà més fàcil de detectar

. De determinades característiques de la població i, molt especialment, de la variabilitat de les dades

. Del criteri utilitzat per a la decisió estadística. Ja s'ha indicat que un criteri habitual és considerar significatiu un resultat amb una probabilitat $p \leq 0,05$. Si s'utilitza un criteri més lax (per exemple, considerar significatiu un resultat amb una probabilitat associada $p \leq 0,10$), la probabilitat de declarar com a significativa una diferència real s'incrementa, tot i que, en contrapartida, també creix la probabilitat de cometre un error tipus I

. De la grandària de la mostra. Quan més gran sigui la mostra, més fàcil serà detectar un efecte real i evitar un error tipus II, per tant, més gran serà la potència de la prova. Això es deu al fet que amb mostres més grans l'error estàndard de la distribució mostral de l'estadístic disminueix

La forma més fàcil d'incrementar la potència d'una prova estadística és augmentar la grandària de la mostra o mostres utilitzades. Ara bé, és evident que no és possible que aquest increment sigui indefinit, per tant sempre s'ha d'assumir una probabilitat de cometre un error tipus II. Com a criteri habitual se sol considerar acceptable treballar amb una potència de 0,80. Si es coneix la magnitud de l'efecte que es vol detectar (cosa difícil si no es disposa d'informació prèvia) és possible determinar el número de casos necessari per realitzar la prova estadística amb un nivell de potència adient. Existeixen taules que permeten determinar el número de subjectes necessari per a diferents proves estadístiques de significació quan es vol assolir un determinat nivell de potència.

5.- Contrast d'hipòtesis sobre mitjanes

En aquest apartat s'estudiarà la forma d'abordar diferents tipus de problemes que impliquen, d'una o altra forma, la comparació entre dues o més mitjanes aritmètiques.

Identificarem bàsicament els següents tipus de situacions:

. Comparació mitjana mostral – mitjana poblacional

. Comparació entre dues mitjanes mostrals. Relació entre una variable qualitativa i una variable quantitativa. Distinció entre dades independents i dades aparellades

. Comparació de més de dues mitjanes mostrals. Relació entre una variable qualitativa i una variable quantitativa

. Correlació i regressió lineals: Relació entre dues (o més) variables quantitatives

5.1 Comparació mitjana mostral-mitjana poblacional

A l'apartat 4.2 s'ha plantejat una situació en la qual, a partir de la mitjana d'una mostra, s'intentava fer una estimació del corresponent paràmetre poblacional, mitjançant un interval de confiança.

Es poden trobar situacions on interessi el problema invers. Coneixent la mitjana de la població, podem intentar determinar si una determinada mostra prové o no d'aquesta població. Certament en molt casos no es coneix *realment* quina és la mitjana poblacional, ja que això implicaria haver mesurat la variable corresponent a *tota* la població, però de vegades es disposa d'estudis molt amplis (amb mostres molt grans) i repetits en el temps, que ens podem donar una molt bona aproximació a la mitjana poblacional, i que podem utilitzar com a referència.

Exemple: Diferents estudis mostren que la mitjana de consum de llet a Catalunya és de 186,5 grams/dia, amb variància desconeguda (en realitat, aquest valor s'ha obtingut a partir de diversos estudis amb una mostra molt àmplia, però a efectes pràctics suposarem que aquesta és la mitjana poblacional, a la qual sense dubte s'aproxima molt). Suposem que un investigador sospita que el consum de llet en determinats col·lectius, com seria el cas dels estudiants universitaris, pot ser diferent al de la població general. Per esbrinar això, agafa una mostra aleatòria de 35 estudiants universitaris i mesura la mateixa variable, obtenint una mitjana de 179,4 grams/dia, amb una variància de 287,1. Es pot afirmar que la mostra d'estudiants presenta un consum de llet similar al de la població general, o bé es produeix realment un consum diferencial?.

Partirem de la base que la distribució de resultats a la població segueix una llei normal, i que el mateix passa a la mostra escollida, cosa que podem comprovar a partir de la corresponent prova de normalitat.

Les hipòtesis a contrastar són, doncs, les següent:

H_0: $\bar{y} = \mu$

H_1: $\bar{y} \neq \mu$

Recordant el que s'indicava anteriorment, aquesta és una hipòtesi *bilateral*, és a dir, la hipòtesi alternativa és que la mitjana mostral difereix de la poblacional, ja sigui per sobre o per sota. L'investigador podria haver plantejat, si tingués fonament per fer-ho, una hipòtesi *unilateral*, per exemple, podria suposar que el consum de llet a la mostra universitària és *inferior* o *superior* a la poblacional.

Coneixent la mitjana de la població, si coneixem o podem estimar la variància poblacional estarem en condicions de calcular el que anomenem un *interval de probabilitat*, és a dir, un interval en el qual, amb una certa probabilitat, s'hauria de trobar la mitjana de qualsevol mostra que obtinguem d'aquesta població.

És una situació inversa a la que es produïa amb els intervals de confiança: Quan calculem un interval de confiança, partim de la mitjana d'una mostra i intentem calcular dins de quin interval, amb una certa probabilitat, es pot trobar la mitjana poblacional. El que fem ara és el procés invers: Coneixent la mitjana poblacional, establim un interval en el qual s'hauria de trobar, amb una certa probabilitat, la mitjana de qualsevol mostra que extraguem d'aquesta població. Si la mitjana d'una certa mostra està dins d'aquest interval, podem concloure que no difereix significativament del paràmetre poblacional, i que la mostra es comporta igual que la població general. En cas contrari, hem de pensar que hi ha diferència significativa i que, per alguna raó, la mostra presenta un comportament diferent al de la població.

Si la mostra és gran (n>30) i/o coneixem la variància poblacional (i, per tant, la corresponent desviació típica) , l'interval de probabilitat s'obté de la forma següent (similar al que ja hem vist amb l'interval de confiança):

$$I.P.= \mu \pm Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Si no es coneix la desviació típica poblacional, es pot utilitzar l'estimador corresponent, és dir, la desviació típica de la mostra. A l'exemple, la variància de la mostra és de 287,1, per tant, la desviació típica és $\sqrt{287,1} = 16,94$.

Per tant, si volem calcular l'interval en el qual s'hauria de trobar qualsevol mitjana mostral amb una probabilitat p = 0,95 (per tant, amb α=0,05), el resultat seria:

$$I.P.= 186,5 \pm 1,96 \cdot \frac{16,94}{\sqrt{35}} = 186,5 \pm 5,612 = (180,89 - -192,11)$$

Vist l'interval obtingut, i sabent que la mitjana de consum de la mostra universitària és de 179,4, constatem que la mitjana mostral està fora de l'interval de probabilitat establert. Per tant, amb un risc d'error α=0,05, podem rebutjar la hipòtesi nul·la i concloure que el consum de llet entre els universitaris és significativament diferent (en aquest cas inferior) al de la població general.

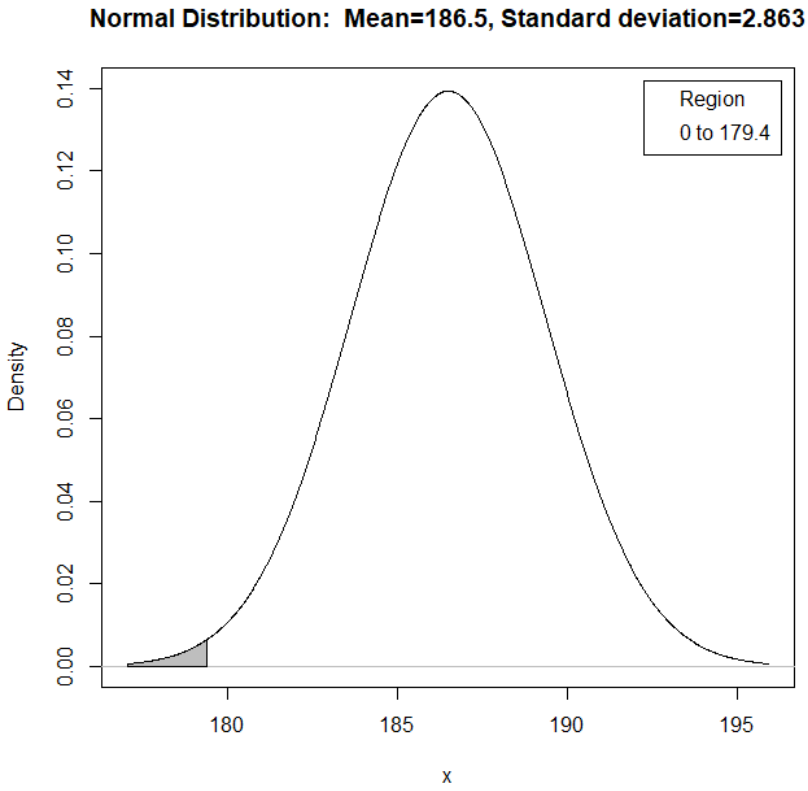
Les aplicacions estadístiques ens permeten anar més enllà, i calcular directament la probabilitat d'obtenir el valor que hem trobat. Si a una distribució normal calculem la probabilitat d'obtenir un valor de 179,4 o inferior, considerant que la mitjana poblacional és de 186,5 i que l'error estàndard de la distribució mostral de mitjanes (atenció: no la desviació típica de la mostra, sinó de la distribució mostral de mitjanes) és $\frac{16,94}{\sqrt{35}} = 2,863$, la probabilitat d'obtenir un valor de 179,4 o inferior és p = 0,00657, per tant, una probabilitat molt petita i, en tot cas, inferior al criteri de significació habitual de p=0,05. Això ens referma en la idea de que el comportament de consum de la mostra universitària s'allunya del de la població general.

També es pot arribar a un mateix resultat per altra via. En primer lloc caldria tipificar la mitjana mostral (és a dir, convertir-la en una puntuació Z), de la forma següent:

$$Z = \frac{\bar{y} - \mu}{\sigma/\sqrt{n}} = \frac{179{,}4 - 186{,}5}{16{,}94/\sqrt{35}} = -2{,}4799$$

A partir d'aquest valor Z es pot recórrer a les taules de la distribució normal tipificada, o demanar directament la probabilitat a qualsevol aplicació estadística. En els dos casos conclourem que la probabilitat d'obtenir una puntuació igual o inferior a -2,4799 és aproximadament de 0,0066 (en el cas de les taules no es pot precisar tant), és a dir, molt aproximadament el valor que ja coneixíem.

Gràficament, és fàcil veure que la puntuació 179,4 és a l'extrem esquerre de la distribució mostral de mitjanes (àrea grisa), d'acord amb la mitjana poblacional i l'error estàndard que coneixem:



L'àrea marcada en gris no és altra que la probabilitat d'obtenir un valor de 179,4 o inferior, i correspon a una probabilitat de 0,00657, com hem vist abans.

Com ja s'ha indicat, la hipòtesi plantejada és bilateral, és a dir, suposàvem que el consum de llet dels universitaris era diferent al de la població general, sense especificar si s'esperava que fos inferior o superior. Si l'investigador hagués tingut prèviament elements suficients per concretar millor aquesta hipòtesi, podria haver optat per plantejar una hipòtesi unilateral. Per exemple, a partir de la literatura prèvia sobre el tema, l'investigador podia haver pensat que el consum de llet dels universitaris seria inferior al de la població general. En aquest cas, les hipòtesis que es plantejarien són:

H_0: $\bar{y} = \mu$

H_1: $\bar{y} < \mu$

El càlcul de l'interval de probabilitat per valorar aquesta hipòtesi és similar al que ja s'ha vist en el cas bilateral, amb una única diferència: En una hipòtesi unilateral no s'ha de treballar amb la puntuació tipificada Z corresponent a α/2, sinó directament amb el valor α. Si volem calcular l'interval amb probabilitat p=0,95, llavors la puntuació Z a utilitzar no ha de ser $Z_{0,025}$, sinó $Z_{0,05}$, per tant, en lloc del valor 1,96 caldria fer ús del valor Z = 1,64, i a partir d'això fer el càlcul normalment. Això es deu al fet que el risc d'error que assumim està acumulat, en el cas unilateral, a un dels costats de la distribució, mentre que en el cas bilateral està repartit entre els dos extrems, ja que no tenim una expectativa prèvia sobre la direcció de la diferència que es pugui produir.

En tot cas, cal insistir en que disposant de l'ajut de les aplicacions estadístiques podem prescindir, a la pràctica, dels càlculs manuals i de les taules, ja que les aplicacions ens ofereixin directament la probabilitat d'obtenir una determinada puntuació o qualsevol inferior (o superior).

Si la mostra és petita (n<30) i la variància de la població és desconeguda, és més ajustat utilitzat la distribució *t* de Student que la distribució normal. Això passava també quan intentàvem establir intervals de confiança i, com en aquell cas, per comparar una mitjana mostral i una mitjana poblacional podem utilitzar sempre la distribució *t*, fins i tot quan la grandària de la mostra sigui gran. Això ho fan per defecte moltes aplicacions estadístiques.

Naturalment, en aquest cas hem de distingir també entre hipòtesis bilaterals o unilaterals, i haurem d'utilitzar el valor de *t* correcte en cada cas: $t_{\alpha/2}$ en el cas bilateral, i t_{α} en el cas unilateral.

En el cas bilateral, l'interval de probabilitat corresponent es calcula amb l'expressió:

$$I.P. = \mu \pm t_{(\nu,\alpha/2)} \cdot \frac{s}{\sqrt{n}}$$

on els graus de llibertat ν són (n – 1). Es pot observar que, suposant que la desviació típica poblacional és desconeguda, la substituïm directament per la seva millor estimació, que és la desviació típica de la mostra.

Si apliquem aquest procediment al cas que s'acaba d'exposar (tot i que la mostra sigui de 35 persones), i plantegem una hipòtesi bilateral amb α=0,05 (és a dir, un interval de probabilitat de 0,95), el valor $t_{0,025}$ amb 34 graus de llibertat és, d'acord amb les taules de la distribució *t* de Student, t = 2,0322. Per tant:

$$I.P. = 186{,}5 \pm 2{,}0322 \cdot \frac{16{,}94}{\sqrt{35}} = (180{,}18 - -192{,}32)$$

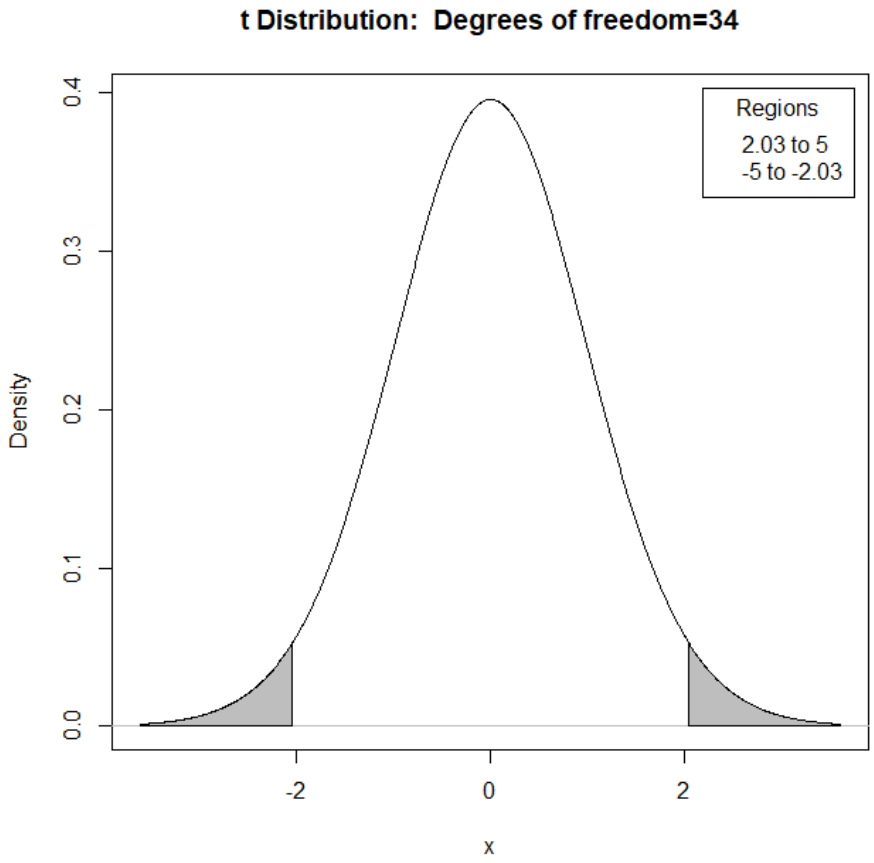
Novament el valor de 179,4, corresponent a la mostra d'universitaris, es troba fora de l'interval de probabilitat obtingut, per tant podem pensar que el comportament dels universitaris pel que fa al consum de llet difereix significativament del de la població general.

Si es vol obtenir la probabilitat exacta de trobar una puntuació igual o inferior a 179,4, primer caldria obtenir el valor *t* de Student corresponent a aquesta puntuació:

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{179{,}4 - 186{,}5}{16{,}94/\sqrt{35}} = -2{,}4799$$

Aquest valor *t* coincideix amb el valor Z obtingut abans. Això es deu a que en els dos casos hem utilitzat la desviació típica de la mostra, però en realitat per fer l'aproximació basada en la distribució normal hauríem d'haver fet servir la desviació típica de la població, que desconeixem. En tot cas, recorrent a les taules de la distribució de Student o demanant directament la probabilitat de t=2,4799 a una aplicació estadística, s'obté un valor de probabilitat aproximat de 0,0091. Tot i que el valor és una mica diferent de l'obtingut a la distribució Z (ens estem basant en dues distribucions de probabilitat diferents), la conclusió és la mateixa: la probabilitat d'obtenir per atzar un resultat igual o inferior a 179,4 és molt baixa (inferior al criteri de referència de 0,05), per tant podem concloure que el consum de llet entre els universitaris és diferent al de la població general.

Gràficament, la situació trobada es pot representar de la forma següent:



Podem veure també que la distribució t de Student-Fisher té la mateixa forma que la normal i també és simètrica, però els valors de referència són diferents i depenen del número de graus de llibertat. Al gràfic de la distribució t s'ha representat a esquerra i dreta la regió corresponent a α/2=0,025. Els límits d'aquestes regions venen marcats pels valors t de -2,0322 (a l'esquerra) i 2,0322 (a la dreta). És fàcil veure que el valor t=-24799 es troba fora de l'interval central, concretament a la zona ombrejada de l'esquerra. Aquesta és una altra forma de dir que el valor corresponent de consum de llet a la mostra d'estudiants (179,4) està fora de l'interval de probabilitat, com ja havíem vist.

Un cas particular (potser poc freqüent) d'ús de la comparació entre una mitjana mostral i una mitjana de referència és la comparació amb una mitjana teòrica. Imaginem, per exemple, que el nostre coneixement de la fisiologia humana ens permet determinar quin és el contingut ideal de sucres d'una determinada fruita per al seu consum humà Ens podem preguntar si la fruita recollida a una determinada zona del Mediterrani s'ajusta o no a aquest contingut "ideal". En aquest cas, no es tracta de comparar una mitjana mostral amb la mitjana poblacional corresponent, sinó amb un valor teòric. Tot i que conceptualment és diferent, el procediment per realitzar-ho és similar al que s'ha descrit, amb la diferència que el punt de partida no és una mitjana poblacional coneguda, sinó un valor teòric que establim a partir de les consideracions que sigui. El que voldrem saber és si el comportament d'una mostra o d'una sèrie de mostres s'ajusta o no al que s'esperaria a partir d'aquest valor teòric de referència, i això es pot fer novament a partir de l'establiment de l'interval de probabilitat corresponent.

5.2 Comparació entre dues mitjanes mostrals

En relació amb aquest tema cal fer dues distincions inicials importants:

En primer lloc, entre dades independents i dades emparellades o aparellades:

- En el cas de grups o dades independents, es tracta de comparar les mitjanes de dues mostres diferents pel que fa a una determinada variable. Exemple: Comparar la mitjana de presència de greixos insaturats (en grams per 100 grams de producte) en dos productes diferents, o en dues mostres diferents del mateix producte
- En el cas de dades emparellades, es compara la mitjana obtinguda en dues mesures diferents de la mateixa variable obtingudes per a un mateix grup de casos. Exemple: Comparació de pes d'un grup determinat de persones abans i després de la realització d'una certa dieta

La segona distinció és entre situacions experimentals i no experimentals:

- A les situacions experimentals existeix algun tipus de manipulació per part de l'investigador, i el que es vol determinar és si aquesta manipulació ha tingut algun efecte en els resultats. Exemples: Efecte d'un nou procediment de fabricació en el contingut d'aigua d'un producte alimentari. Podem treballar amb dues factories, introduir en una d'elles el nou procediment i mantenir a l'altra el procediment tradicional. Posteriorment es contrastaran les mitjanes de contingut d'aigua en mostres de productes de cadascuna de les dues fàbriques
- A les situacions no experimentals no es produeix manipulació per part de l'investigador, només es volen constatar possibles diferències que no són producte de cap actuació concreta per la seva part. Exemple: Comparació d'IMC entre homes i dones

Una altra forma d'entendre el que suposa la comparació entre dues mitjanes mostrals és considerar que el que fem realment és establir la relació entre dues variables: la que configura els grups o mesures i la variable quantitativa que mesurem. Molt habitualment la variable que configura els grups és qualitativa, tot i que pot haver-hi excepcions. Això es pot veure fàcilment en els exemples anteriors, on podem apreciar que relacionem una variable qualitativa i una quantitativa:

- Relació entre tipus de producte (quali) i quantitat de greixos insaturats (quanti)
- Relació entre la realització o no de dieta (quali) i pes (quanti)

- Relació entre procediment de fabricació (quali) i contingut d'aigua del producte (quanti)
- Relació entre gènere (quali) i IMC (quanti, si no es categoritza)

Quan parlem de la comparació entre dues mitjanes mostrals és fonamental tenir sempre present que som a l'àmbit de l'estadística inferencial. Per tant, encara que treballem amb les dades de mostres concretes, el nostre objectiu és generalitzar els resultats obtinguts a la població o poblacions d'origen d'aquestes mostres. La lògica i els procediments de l'estadística inferencial estan pensats per aconseguir això.

. Comparació de mitjanes amb dades independents

En aquestes situacions trobarem dos grups de subjectes o casos en els quals s'hauran registrat els resultats d'una variable quantitativa. S'obtindrà la mitjana d'aquesta variable a cadascun dels grups i a continuació es procedirà a comparar aquestes dues mitjanes amb la finalitat d'establir si són o no significativament diferents.

Com en qualsevol prova estadística de significació, es tracta de contrastar les dues hipòtesis possibles, nul·la i alternativa. Si ho expressem en termes de les mitjanes poblacionals corresponents a cadascun dels grups, les hipòtesis nul·la i alternativa serien:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

En aquest cas la hipòtesi alternativa és bilateral, és a dir, la nostra suposició és que les dues mitjanes comparades són diferents, sense especificar quina pot ser la més gran i quina la més petita. Naturalment, també es pot plantejar una hipòtesi alternativa unilateral, si esperem que una de les mitjanes sigui superior a l'altra.

La manera concreta de realitzar la comparació entre dues mitjanes en grups independents depèn sobretot de dos elements: la grandària de la mostra i l'homogeneïtat o no de les variàncies poblacionals.

Pel que fa al primer element (grandària de la mostra), com en casos anteriors es pot distingir entre mostres grans (amb n igual o superior a 30) i mostres petites (n inferior a 30).

El segon element no s'havia tractat fins ara, i es refereix al supòsit que les variàncies poblacionals corresponents als dos grups no són significativament diferents. Aquesta condició s'anomena *homogeneïtat de variàncies* o *homocedasticitat*, i el seu compliment és important per a un bon funcionament de les proves estadístiques. Com que normalment no coneixerem les variàncies poblacionals, com és habitual es treballarà amb les corresponents variàncies mostrals, és a dir, les variàncies de cadascun dels dos grups que es volen comparar.

La comprovació de la condició d'homocedasticitat es pot fer de forma senzilla realitzant el seu quocient, posant normalment la variància més gran al numerador i la més petita al denominador. El quocient de dues variàncies segueix una distribució de probabilitat coneguda, però que no havíem vist fins ara: la distribució F de Fisher-Snedecor. Com en altres casos, el coneixement d'aquesta distribució ens permetrà decidir, amb els risc α que vulguem assumir, si les dues variàncies comparades són o no diferents:

$$F = \frac{s^2_{max}}{s^2_{min}}$$

En aquesta expressió, s^2_{max} és la més gran de les dues variàncies comparades, i s^2_{min} és la més petita. A partir d'aquí, com sempre, es poden fer dues coses:

. Recórrer a les taules de la distribució F de Fisher-Snedecor, amb (n_1-1) i (n_2-1) graus de llibertat, on n_1 és el número de casos del grup 1 i n_2 el número de casos del grup 2. A partir d'aquí, com s'ha fet en casos anteriors, es pot establir de forma aproximada la probabilitat de que la diferència entre les dues variàncies (expressada en aquest cas en forma del seu quocient) sigui deguda a l'atzar. Com sempre, si aquesta probabilitat és molt baixa (inferior a 0,05) podem concloure que hi ha una diferència significativa entre les dues variàncies, i que no es compleix la condició d'homocedasticitat. En cas contrari, podem assumir que la condició es compleix

. Calcular directament, normalment a través de les aplicacions estadístiques, la probabilitat de que el valor F obtingut es pugui haver produït per atzar. Novament, la comparació d'aquesta probabilitat amb el valor de referència α = 0,05 ens permetrà determinar si la diferència entre les dues variàncies és o no estadísticament significativa i, per tant, si es compleix o no la condició d'homogeneïtat de variàncies

Per realitzar la comparació entre les mitjanes dels dos grups, la tendència habitual és utilitzar sempre l'estadístic *t* de Student, que ja coneixem. Efectivament, es pot demostrar que la diferència entre dues mitjanes obtingudes en grups de grandària n_1 i n_2 es distribueix segons una distribució *t* de Student-Fisher amb (n_1+n_2-2) graus de llibertat. Per tant, al conèixer quina distribució és aplicable estem en disposició de calcular la probabilitat associada a qualsevol diferència empírica que puguem trobar i, en conseqüència, podem prendre la decisió estadística corresponent.

En cas que es compleixi la condició d'homocedasticitat, el valor *t* per a la comparació de dues mitjanes $\bar{y}_1$ i $\bar{y}_2$ s'obté de la forma següent:

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

L'expressió s_p^2 és una variància ponderada entre les variàncies dels dos grups, que es calcula de la forma següent:

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

En cas que no es compleixi la condició d'homocedasticitat, no es realitza la ponderació de les dues variàncies, sinó que s'utilitzen per separat:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

A més, en aquest darrer cas (i especialment quan les mostres són petites) és molt recomanable utilitzar una correcció de la prova, anomenada *correcció de Welch* (o de Welch-Stterhwaite o de Welch-Aspin). La correcció es refereix als graus de llibertat associats a la prova. Aquests autors van demostrar que en cas d'incompliment de la condició d'homocedasticitat, la distribució de la diferència entre dues mitjanes no s'ajusta exactament a una distribució *t* de Student amb (n_1+n_2-2) graus de llibertat, sinó a la mateixa distribució amb uns graus de llibertat corregits.

Concretament, els graus de llibertat de Welch es calculen de la forma següent (s'utilitza l'expressió més habitual, tot i que hi ha petites variants):

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Una altra alternativa quan no es compleix la condició d'homogeneïtat de variàncies i les mostres són petites és utilitzar una *prova no paramètrica*. Les proves no paramètriques són menys potents que les habituals (o paramètriques), però tenen l'avantatge de no basar-se en supòsits previs sobre la forma de distribució de les dades, l'existència d'homocedasticitat o altres. En el cas de la comparació entre dues mitjanes en grups independents la prova no paramètrica corresponent és la U de Mann-Whitney. De tota manera, si es realitza la correcció de Welch normalment no serà necessari recórrer a aquesta alternativa, sempre que suposem que la distribució de la variable a la població segueix una llei normal.

En tot cas, la majoria d'aplicacions estadístiques treballen sempre per defecte amb la *t* de Student, i apliquen en cas necessari la correcció de Welch (ja sigui per defecte o a petició de l'usuari).

Exemple: Es vol valorar si existeixen diferències en el consum de verdures entre homes i dones en una població universitària. Per fer-ho es treballa amb un grup de 41 dones i un de 30 homes, escollits a l'atzar dins del conjunt dels estudiants universitaris, i es fa un seguiment via enquestes del seu consum de verdures durant un determinat període de temps. Els resultats mostren que la mitjana de consum de verdures en el grup de dones és de 125,6 grs/dia, mentre que en el grup d'homes és de 115 grs/dia. Les variàncies corresponents són de 203 i 196, respectivament. No existeix cap hipòtesi prèvia sobre la direcció de les possibles diferències entre els grups, per tant es plantejarà un contrast bilateral.

En primer lloc cal establir si es compleix o no la condició d'homogeneïtat de variàncies. Sabem que la variància més gran és l'obtinguda en el grup de dones. Per tant:

$$F = \frac{s_{max}^2}{s_{min}^2} = \frac{203}{196} = 1{,}036$$

Els graus de llibertat a aplicar són 40 per al numerador i 29 per al denominador, ja que la grandària de les mostres és de 41 per a dones i 30 per a homes. Si, a partir d'aquí, calculem mitjançant qualsevol aplicació estadística quina és la probabilitat exacta d'obtenir aquest resultat per atzar, el resultat és p = 0,467, molt per sobre del valor α=0,05. Per tant podem afirmar que la probabilitat de que la diferència entre les dues variàncies sigui producte de l'atzar és gran (concretament de 0,467), i en tot cas superior al criteri de risc assumit, i per tant, d'acord amb el procediment habitual, acceptem la hipòtesi nul·la i considerem que no ha una diferència significativa entre elles. És a dir, la condició d'homocedasticitat es compleix.

Vist el compliment de la condició d'homogeneïtat de variàncies, es procedeix a continuació a calcular la variància ponderada que s'utilitzarà en el càlcul de l'estadístic *t*.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(41 - 1) \cdot 203 + (30 - 1) \cdot 196}{41 + 30 - 2} = 200{,}06$$

A partir d'aquí es pot calcular el valor de l'estadístic:

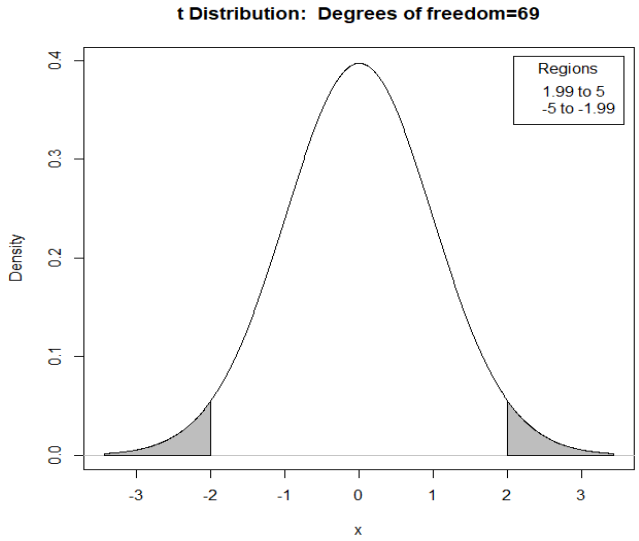
$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{125{,}6 - 115}{\sqrt{\frac{200{,}06}{41} + \frac{200{,}06}{30}}} = 3{,}118$$

Finalment, cal establir el nivell de significació per tal prendre la corresponent decisió estadística. Cal recordar que s'està provant una hipòtesi bilateral, per tant si s'assumeix el nivell de risc habitual (α=0,05) cal utilitzar com a referència el valor α/2=0,025. A partir d'aquí es pot prendre la decisió estadística a partir d'algun dels procediments ja coneguts:

. Si es consulta la taula de la distribució teòrica *t* de Student amb (41+30-2)=69 graus de llibertat, el valor de referència per a α/2=0,025 és de 1,9949. Això vol dir que el valor màxim que es pot produir com a resultat de l'atzar és de 1,9949. Ja que 3,118 > 1,9949, es pot rebutjar la hipòtesi nul·la i podem concloure que hi ha una diferència significativa en el consum de verdures entre homes i dones

. Novament, la forma més pràctica de procedir és fer els càlculs mitjançant una aplicació estadística, que ens donarà directament el nivell de significació que busquem. Cal tenir en compte, però, una qüestió: Normalment les aplicacions estadístiques demanen si la hipòtesi és unilateral o bilateral, i fan els càlculs d'acord amb això. En el nostre cas, si per exemple utilitzem l'aplicació R, caldria indicar que la hipòtesi és bilateral, i l'aplicació ens donaria el resultat t=3,118 i un valor p = 0,026. Seria un error comparar aquest valor amb la referència α/2= 0,025. Si la prova és bilateral, això vol dir que el risc α = 0,05, que assumim normalment, queda distribuït entre els dos extrems (0,025 a l'esquerra i 0,025 a la dreta), tal i com hem considerat al mirar les taules. El que fa l'aplicació estadística és calcular el valor p considerant que el resultat obtingut podria estar tant a la dreta (que és realment el cas) com a l'esquerra. Per tant, el valor p obtingut (0,026) no s'ha de comparar amb 0,025, com fèiem al mirar les taules, sinó amb α = 0,05. Ja que 0,026 < 0,05, concloem que la diferència obtinguda és estadísticament significativa, com ja havíem trobat al mirar les taules. Per tant, entenem que la diferència obtinguda no és deguda a l'atzar i podem concloure que, efectivament, hi ha diferència en el consum de verdures entre els dos grups.

Una forma de veure la situació de forma gràfica seria a partir de la representació següent:



Aquí es representen les regions de la distribució t (amb 69 graus de llibertat) corresponents a α/2=0,025, que tenen com a límit els valors t=1,9949 a la dreta i -1,9949 a l'esquerra. Com que la t obtinguda empíricament és positiva, cal fixar-se només a la regió de la dreta. Es pot veure fàcilment que el valor obtingut a la comparació empírica (t=3,118) es troba dins de la regió marcada, i molt per sobre del valor de referència de 1,9949.

. Comparació de dues mitjanes amb dades emparellades

Com s'ha indicat anteriorment, el cas de dades aparellades es planteja quan un mateix grup de subjectes o casos son mesurats dues vegades pel que fa a una mateixa variable. Per tant, per a cada cas tindrem dos resultats, un per a cadascun dels registres. Hi ha diverses situacions on es poden trobar dades aparellades: Pes d'un grup de persones abans i després de realitzar una dieta, IMC d'un mateix grup de persones als 15 i als 20 anys, nivell de glucosa en sang al matí o a la tarda en un determinat grup, nivell d'acidesa d'una mostra d'un aliment en dos moments diferents separats per un cert interval, etc.

A partir dels resultats de qualsevol estudi d'aquest tipus disposarem de dues dades per a cada cas. La forma d'abordar l'anàlisi de les dades és a partir de les diferències entre les puntuacions de cada subjecte o cas als dos registres. A continuació caldrà calcular la mitjana d'aquestes diferències i la seva variància i desviació típica. Esquemàticament:

Casos	Resultat 1	Resultat 2	Diferència
1	Y_{11}	Y_{12}	$Y_{11} - Y_{12}$
2	Y_{21}	Y_{22}	$Y_{21} - Y_{22}$
3	Y_{31}	Y_{32}	$Y_{31} - Y_{32}$
...	....	....	.....
...	....	....	.....
n	Y_{n1}	Y_{n2}	$Y_{31} - Y_{32}$
			$\bar{y}_d$, s^2_d, s_d

$\bar{y}_d$: Mitjana de les diferències

s^2_d : Variància de les diferències

s_d : Desviació típica de les diferències

Per valorar si les diferències són significatives o no es pot utilitzar l'estadístic *t* de Student segons l'expressió següent:

$$t = \frac{\bar{y}_d}{s_d / \sqrt{n}}$$

Aquest quocient segueix una distribució *t* de Student-Fisher amb (n-1) de graus de llibertat, cosa que ens permet valorar de forma adient la hipòtesi sobre la possible diferència en els resultats dels dos registres, d'acord amb els procediments habituals.

Exemple: Suposem que volem valorar l'eficàcia comparada de dos mètodes analítics a l'hora de valorar el contingut de sulfits en vins d'una determinada marca. Per fer-ho s'escull a l'atzar una mostra de 15 botelles del producte, i es sotmeten a anàlisi mitjançant cadascun dels dos procediments. Es registra el contingut d'anhídrid sulfurós en mg/litre. Per a cada ampolla s'obtenen, per tant, dos resultats, un amb cadascun dels mètodes d'anàlisi. Suposem que els resultats obtinguts són els següents:

	Contingut anhídrid sulfurós	
Ampolles	Mètode 1	Mètode 2
1	21,7	20,4
2	22,3	22,5
3	21,3	20,8
.....	.....	.....
.....	.....	.....
15	19,9	20,1

Per tal de procedir a l'anàlisi estadística, cal identificar en primer lloc que es tracta d'un cas de dades aparellades, ja que els dos mètodes d'anàlisi s'apliquen a la mateixa mostra de casos (n=15). Per tant, cal treballar amb les diferències entre els dos registres:

	Contingut anhídrid sulfurós		
Ampolles	Mètode 1	Mètode 2	Diferència
1	21,7	20,4	1,3
2	22,3	22,5	-0,2
3	21,3	20,8	1,5
.....	.....	.....	.....
.....	.....	.....	.....
15	19,9	20,6	-0,7
			$\bar{y}_d$,=0,207 s_d=0,555

Els resultats per a la columna de les diferències són:

$\bar{y}_d = 0{,}207$ $s_d = 0{,}555$

En conseqüència:

$$t = \frac{\bar{y}_d}{s_d/\sqrt{n}} = \frac{0{,}207}{0{,}555/\sqrt{15}} = 1{,}44$$

Significació estadística:

. Valor *t* teòric per (n-1) = 14 g.ll., i α/2 = 0,025 (hipòtesi bilateral): 2,1448

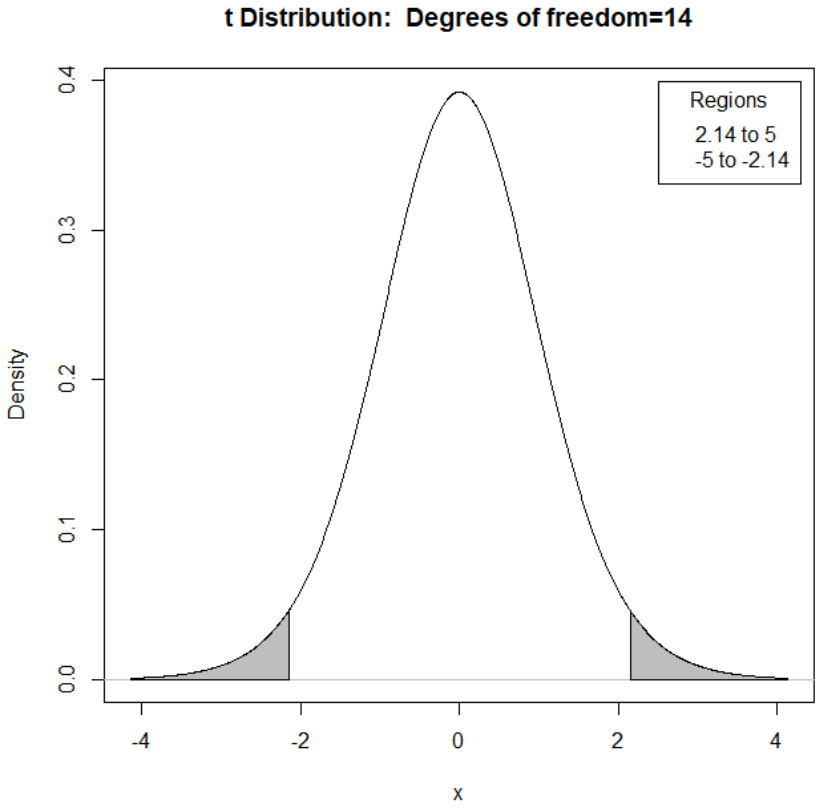
Ja que 1,44 < 2,1448 , no podem rebutjar la hipòtesi nul·la

. Càlcul probabilitat prova estadística: p = 0,171 (comparació amb 0,05)

Aquesta probabilitat és superior a 0,05, que és nivell de risc que assumim habitualment. Per tant, la probabilitat de que la diferència obtinguda sigui deguda a l'atzar és massa gran com per rebutjar la hipòtesi nul·la.

Queda clar, a partir dels resultats obtinguts, que hem d'acceptar la hipòtesi nul·la i, per tant, podem pensar que no hi ha una diferència significativa entre els dos mètodes d'anàlisi i, en conseqüència, els dos mètodes es podrien utilitzar indistintament.

Gràficament, la distribució *t* de Student amb 14 graus de llibertat és la següent (es marquen les regions corresponents a α/2 = 0,025 a dreta i esquerra):



Com sempre, les zones corresponents a α/2=0,025 venen limitades per valor *t* que podem trobar a les taules (t=2,1448). És obvi que el valor empíric obtingut a l'anàlisi (t=1,44) es troba fora de les zones marcades, cosa que significa que podem pensar que la diferència observada és deguda a l'atzar.

5.3 Comparació entre més de dues mitjanes: Introducció a l'anàlisi de la variància

És freqüent que es donin situacions a les quals calgui fer una comparació entre més de dues mitjanes: Per exemple, pot interessar comparar l'impacte de tres sistemes de fabricació diferents sobre la qualitat del producte final; o comparar l'IMC obtingut en quatre grups demogràfics diferents, etc.

Una manera d'abordar l'anàlisi dels resultats d'aquest tipus d'estudis seria a partir de la comparació dos a dos entre les mitjanes obtingudes. Per exemple, en el cas dels tres sistemes de producció, podríem comparar mitjançant la prova *t* de Student els resultats del sistema 1 amb els del sistema 2, els del 2 amb el 3, i els de l'1 amb el 3. Ara bé, hi ha una prova alternativa que ens informa sobre l'existència de diferencies *globals* entre les mitjanes, no dos a dos sinó

com a conjunt. Aquesta prova és l'*anàlisi de la variància* (AVAR o ANOVA), i aquí només s'exposaran les idees més bàsiques sobre ella.

A l'anàlisi de la variància, si tenim *k* mitjanes, la hipòtesi nul·la s'expressaria de la forma següent:

H_0: $\bar{y}_1 = \bar{y}_2 = \ldots\ldots = \bar{y}_k$

Naturalment, la hipòtesi alternativa és que hi ha hagi desigualtat entre algunes de les mitjanes, és a dir, que com a mínim per a algun parell de mitjanes *i j* es doni que : $\bar{y}_i \neq \bar{y}_j$.

Suposem, per exemple, que per als tres sistemes de fabricació citats els resultats mitjans de l'índex de qualitat dels productes siguin de 14.3, 13.6 i 16.4, respectivament (més puntuació significaria més qualitat). Òbviament les tres mitjanes són diferents, però el que cal és establir si aquesta diferència és deguda a l'atzar o bé té a veure amb diferències reals entre els sistemes.

El fet que les tres mitjanes obtingudes siguin diferents ens permet dir que existeix una certa variabilitat entre elles, de fet podem parlar d'una "variància de les mitjanes". El que fa l'anàlisi de la variància és comparar aquesta variància de les mitjanes amb la variància que es produeix dins de cadascun dels grups. La lògica d'aquesta comparació és la següent: la variabilitat que es pugui observar dins de cada grup és aleatòria i no es relaciona amb el sistema de fabricació, ja que tots els productes del grup han estat fabricats amb el mateix sistema. Per tant, aquesta "variància dins dels grups" és deguda a l'atzar. El que fa l'anàlisi de la variància és comparar la "variància de les mitjanes", o "variància entre els grups" (que no sabem si és deguda o no a l'atzar) amb la "variància dins dels grups" o variància intra-grups (que en principi atribuïm a l'atzar). Si la variància entre-grups és significativament més gran que la variància intra-grups, podem pensar que no és deguda a l'atzar i que realment els tres sistemes de fabricació difereixen en els seus resultats de qualitat.

Per tant, l'anàlisi de la variància deu el seu nom al fet que es comparen dues variàncies. En el cas proposat, la comparació és entre la variància entre-grups i la variància intra-grups. Com hem vist anteriorment, tot i que amb un objectiu força diferent, dues variàncies es poden comparar fent el seu quocient, i el quocient de dues variàncies segueix una distribució F de Fisher-Snedecor. Per tant, la prova estadística a utilitzar seria:

$$F = \frac{s^2_{entre}}{s^2_{intra}}$$

Per tant, la hipòtesi nul·la i la hipòtesi alternativa originals queden transformades de la forma següent:

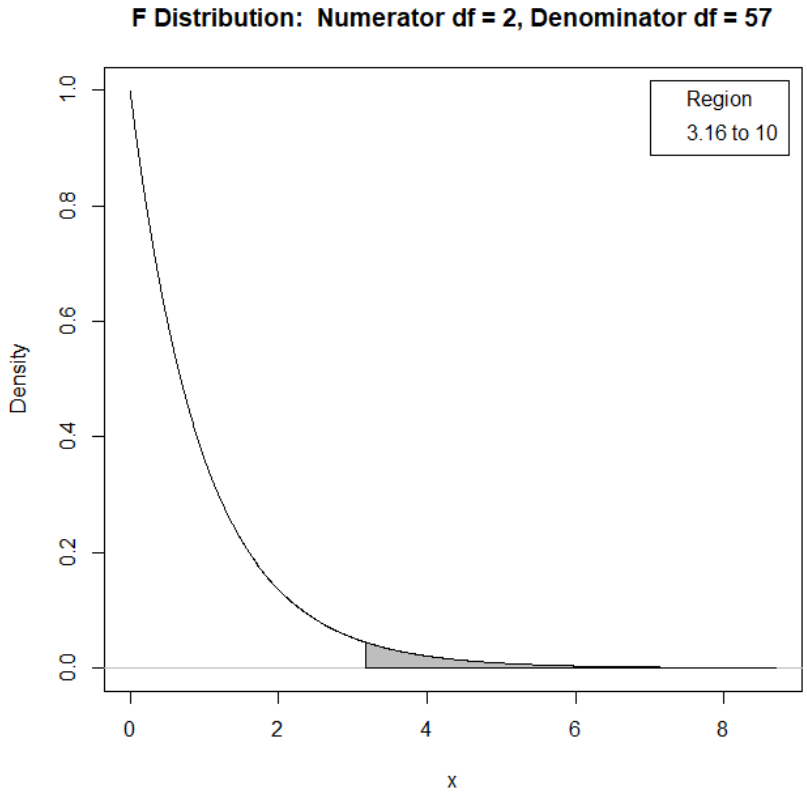
H_0: $s^2_{entre} = s^2_{intra}$

H_1: $s^2_{entre} > s^2_{intra}$

Aquesta hipòtesi és unilateral, ja que el que volem saber és la variància entre-grups és *més gran* que la variància intra-grups, no solament si són diferents.

El quocient indicat es distribueix segons una llei de Fisher-Snedecor amb (k-1) i (N-k) graus de llibertat, on k és el número de grups i N és el número total de casos per al conjunt dels grups. A partir d'aquí, segons els criteris habituals és possible establir si la diferència entre el conjunt de les mitjanes és o no significativa.

Si, en el cas exposat, suposem que per a cadascun dels tres sistemes de fabricació hem recollit una mostra de 20 productes, el número total de casos serà de 60 i, per tant, els graus de llibertat seran (k-1)=3-1=2 per al numerador, i (N-k)=(60-3)=57 per al denominador. Gràficament, la distribució F per aquest cas és:



La zona marcada és la corresponent a α=0,05, i el seu límit és el valor F=3,16 (que es pot trobar a les taules de la distribució). Per tant, si el quocient que obtinguem és igual o superior a 3,16, quedarà dins de la zona marcada i rebutjarem la hipòtesi nul·la. Naturalment, qualsevol aplicació estadística ens pot fer els càlculs corresponents i evita la necessitat de consultar les taules.

Com es pot veure, la distribució *F* és molt diferent de la normal o de la distribució *t* de Student-Fisher i, en particular, no és simètrica. Al fer l'anàlisi de la variància, plantegem la hipòtesi que la variància del numerador serà superior a la del denominador, en la magnitud que sigui. Fem, doncs, un plantejament unilateral. L'opció contrària no tindria interès ni sentit per a nosaltres.

L'anàlisi de la variància es pot utilitzar tant per a dades independents com per a dades aparellades, tot i que en aquest segon cas hi ha diferents variacions en l'anàlisi, que no seran estudiades aquí.

Finalment, cal indicar que la realització de l'anàlisi de la variància no és contradictòria amb la possibilitat de fer comparacions parcials entre les diferents mitjanes obtingudes, en realitat les dues anàlisis es complementen entre sí, tot i que per qüestions tècniques, que no tractarem aquí, la millor opció és començar amb l'anàlisi de la variància i posteriorment fer les comparacions parcials, per a les quals fins i tot hi ha proves específiques, més enllà de l'estadístic t de Student convencional, que té alguns inconvenients en aquest cas.

5.4 Relació entre variables quantitatives: Correlació i regressió lineal

Als exemples anteriors, d'una o altra forma es plantejava la relació entre una variable qualitativa (o tractada com a qualitativa) i una variable quantitativa. És possible també abordar l'anàlisi de la relació entre dues variables quantitatives, a través de dos instruments fonamentals:

. El coeficient de correlació: És un índex que ens permet valorar la intensitat de la relació entre dues variables quantitatives i si aquesta intensitat és prou gran com per ser considerada estadísticament significativa

. Els models de regressió: Ens permeten escriure una equació que ens relacioni les variables implicades i ens permeti, entre d'altres coses, fer prediccions sobre el resultat d'una variable a partir del valor d'una altra

Imaginem que per a una mostra de subjectes disposem de registres per a dues variables quantitatives: la ingestió mitjana diària de calories durant dues setmanes (variable x) i el pes de cada persona (variable y). Per valorar la intensitat de la relació que pugui existir entre les dues variables, l'instrument utilitzat més habitualment és el *coeficient de correlació lineal de Pearson*. Aquest coeficient s'obté de la forma següent:

En primer lloc cal establir la suma de productes creuats de les variables *x* i *y* i dividir-la entre el número de casos menys un. Si calculem prèviament la mitjana de cadascuna de les variables a la mostra, i si aquesta mostra consta de n casos, l'expressió és la següent:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

El numerador d'aquesta equació és la suma de productes creuats. Suposem, per exemple, que la mitjana d'ingesta de calories per als *n* casos és de 2690 i el pes mitjà de la mostra és de 65,8 quilos. Si una persona concreta presenta una ingestió de 2860 calories i un pes de 74,6 quilos, el producte creuat corresponent seria: (2860-2690)(74,6-65,8)=1496. Es procediria igual amb els *n* casos de la mostra i es faria la suma de tots els productes creuats obtinguts. Finalment, aquest sumatori es dividiria per (n-1).

Una vegada disposem del càlcul de s_{xy} només cal aplicar la fórmula del coeficient de correlació lineal de Pearson, que és:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

on s_x i s_y són, respectivament, les desviacions típiques de les distribucions de les variables x i y dins de la mostra.

El coeficient de correlació es pot interpretar tenint en compte el següent:

. El valor del coeficient de correlació lineal de Pearson està sempre entre -1 i 1

. Un coeficient positiu ens indicaria una relació directa o positiva entre les dues variables. En el nostre cas, això voldria dir que a més ingesta de calories, més pes

. Un coeficient negatiu ens indicaria una relació inversa o negativa entre les dues variables, de forma que quan una d'elles s'incrementa, l'altra tendeix a baixar

. Un coeficient proper a zero ens indicaria que no hi ha relació entre les dues variables

Per esbrinar si el coeficient de correlació és estadísticament significatiu és habitual utilitzar la distribució *t* de Student-Fisher, ja que es pot demostrar que la distribució mostral de correlacions segueix una llei de Student-Fisher amb (n-2) graus de llibertat. Com sempre, les aplicacions estadístiques ens permeten calcular la probabilitat exacta del coeficient de correlació obtingut i determinar si és significatiu o no comparant-lo amb el valor habitual de α=0,05. En realitat el que es fa és comparar el coeficient de correlació obtingut amb un valor teòric de zero, i veure si el coeficient real és o no significativament diferent de zero.

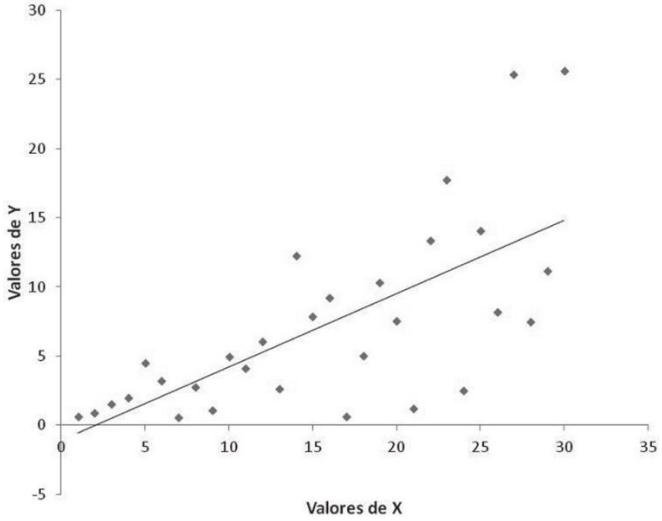
Cal dir que hi ha altres coeficients de correlació que poden ser útils quan es treballa amb variables ordinals o d'altre tipus (coeficients de Spearman, Kendall, Goodman i Kruskall, coeficient de correlació biserial i biserial-puntual, etc.). Aquests coeficients no seran tractats aquí, però es posen com a referència per si cal cercar-ne algun per a alguna anàlisi especial.

Tampoc s'entrarà en el detall dels models de regressió lineal, un tema per altra banda força extens i que pot arribar a tenir un cert nivell de complexitat. Només es farà una petita referència al cas més senzill, on tenim únicament dues variables quantitatives (x i y), i constatem (per exemple, a través del coeficient de correlació lineal de Pearson) que existeix una correlació significativa entre una i l'altra. En aquest cas, podríem escriure una equació que ens permeti, conegut un valor real o hipotètic de x, fer una predicció sobre el valor corresponent de y. Aquesta equació tindria la forma següent:

$$\hat{y}_i = b_0 + b_1 x_i$$

Aquest és un *model lineal,* cosa que es pot veure fàcilment observant que no és altra cosa que l'equació d'una recta, a la qual b_0 és l'element constant, és a dir, el valor de y quan x=0, i b_1 és el *coeficient de regressió*, que indica el pendent (positiu o negatiu) de la recta. És, per tant, un *model lineal*, representable mitjançant una línia recta.

Gràficament, es pot observar que el que intentem és definir la línia recta que millor s'ajusti al conjunt de punts (resultats de *x* i *y* per a cada subjecte o cas):



En aquest gràfic, la relació entre les dues variables és positiva o directa, és a dir, a major valor de x, normalment hi ha també un major valor de y, i per tant el pendent de la recta és positiu. Lògicament, el coeficient de correlació entre aquestes dues variables serà també positiu.

A l'exemple plantejat abans, si a partir de les dades podem calcular els valors de b_0 i de b_1, llavors estarem en condicions, coneixent la ingesta de calories d'una persona (x_i), de fer una predicció sobre el seu pes ($\hat{y}_i$). Naturalment, aquesta predicció no serà exacta, és a dir, inclourà gairebé sempre una desviació o error respecte de la puntuació real de pes d'aquella persona. Lògicament, quan més potent sigui la relació entre les variables *x* i *y* més precises seran les prediccions del model i menor serà el component d'error. El cas extrem seria el d'una correlació perfecta, ja sigui positiva (r=1) o negativa (r=-1). En aquest cas a partir de *x* es podrien fer prediccions perfectes sobre *y*, però naturalment això no es produeix mai a la pràctica.

L'estimació de les incògnites del model es pot a partir de les expressions següents, que inclouen termes ja coneguts:

Pel que fa al coeficient de regressió:

$$b_1 = \frac{s_{xy}}{s_x^2}$$

I el terme constant:

$$b_0 = \bar{y} - b_1 \bar{x}$$

Si el coeficient de correlació entre x i y és significatiu, llavors cal pensar que el model de regressió ens permetrà fer prediccions amb una certa precisió. Alternativament, el model de regressió pot ser validat directament a partir del *coeficient de determinació*. El coeficient de determinació (representat per R^2) no és altra cosa que el coeficient de correlació elevat al quadrat, i ens indica quina proporció de la variabilitat de la variable *y* es pot explicar a partir de la variable *x*. Si, per exemple, el coeficient de correlació entre les dues variables és de 0,64, el coeficient de determinació seria: $R^2 = 0{,}64^2 = 0{,}41$. Això significaria que el 41% de la variabilitat de la variable y està explicada per la variable x. L'avantatge del coeficient de determinació és que pot ser aproximat a una distribució F de Snedecor, i això ens ofereix una segona prova de significació (a més de la ja realitzada amb el coeficient de correlació) que ens permet validar el model.

Com ja s'ha indicat, s'han exposat només els elements més bàsics dels models de regressió lineal. Aquests models poden complicar-se considerablement introduint noves variables predictores, i permet fins i tot treballar amb variables qualitatives (cosa que exigiria una recodificació de les variables), entre altres possibilitats.

6.- Contrast d'hipòtesis sobre proporcions

A l'apartat anterior, s'han plantejat diferents situacions relatives a comparacions entre mitjanes, és a dir, per a variables quantitatives. A l'apartat actual es farà un plantejament paral·lel per a variables qualitatives (o tractades qualitativament), on treballarem habitualment a partir de proporcions o percentatges: Proporció de persones a cada categoria de IMC (l'índex de massa corporal és una variable quantitativa, però moltes vegades s'utilitza de forma qualitativa, amb categories); proporció de persones diabètiques; proporció de persones a cada grup sanguini; proporció de productes defectuosos en una factoria, etc, etc.

Les situacions principals que es poden plantejar són:

. Comparació entre una proporció mostral i una proporció poblacional

. Comparació entre proporcions mostrals

6.1 Comparació entre una proporció mostral i una proporció poblacional

. Cas de dues proporcions

Establint un paral·lelisme amb el que passava amb les mitjanes, en aquest cas es tracta de valorar si les proporcions observades en una determinada mostra difereixen o no significativament d'una proporció poblacional (o teòrica).

Per exemple, tant les diferents edicions de l'Enquesta de Salut de Catalunya com altres estudis estableixen de forma bastant aproximada que gairebé la meitat de la població catalana major de 18 anys té excés de pes, és a dir, es troben a les categories de sobrepès o obesitat. En concret, el 35,2% té sobrepès i el 13,8% obesitat, de forma que podem dir que el 49% de la població (o una proporció de 0,49) presenten excés de pes. Suposem que realitzem un estudi similar a una ciutat determinada, agafant una mostra de 50 persones, i constatem que 21 d'elles mostren excés de pes, és a dir, un 42% de la mostra es troba a la zona de sobrepès o obesitat. Podem pensar que el resultat a aquesta ciutat difereix significativament del resultat a la població global?.

El resultat obtingut a la mostra és fàcil de resumir:

	Excés de Pes	No excés	Total
Nª casos	21	29	50
Proporcions	0,42	0,58	1

Es tracta de valorar si una determinada proporció *p*, obtinguda en una mostra concreta, difereix o no significativament del paràmetre poblacional π. A l'exemple proposat, si ens centrem en la proporció de persones amb pes superior al normatiu, els valors a considerar són p=0,42, π=0,49.

La distribució mostral de proporcions segueix una llei normal sempre que $\pi \cdot n \geq 5$ i $(1-\pi) \cdot n \geq 5$. En el cas proposat: $0,49 \cdot 50 = 24,5$, i $0,51 \cdot 50 = 25,5$, de forma que es compleix aquesta condició.

D'acord amb això, tal i com passava en el cas de les mitjanes, hi ha diverses formes equivalents de determinar si existeix o no una diferència significativa entre el valor mostral i el poblacional.

En primer lloc és possible establir quin és l'interval en el qual, amb la probabilitat que es determini, s'hauria de trobar la proporció mostral, coneguda la proporció a la població. És a dir, podem calcular un interval de probabilitat. L'expressió a utilitzar és la següent:

$$I.P.= \pi \pm Z_{\alpha/2}\sqrt{\frac{\pi\,(1-\pi)}{n}}$$

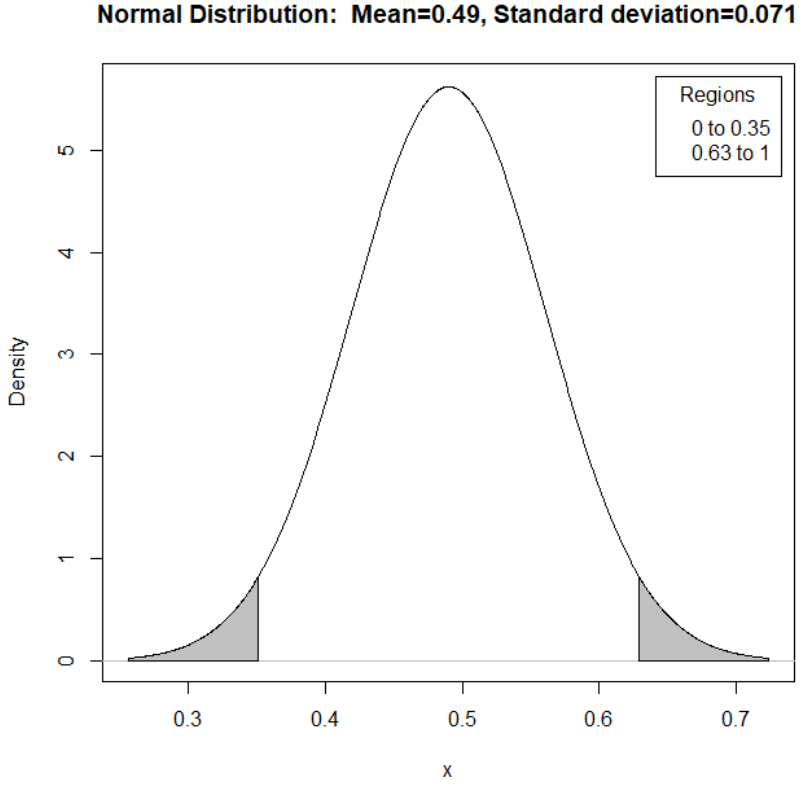
El valor central de la distribució mostral de proporcions és igual al paràmetre poblacional π, i l'error estàndard de la distribució és igual a $\sqrt{\frac{\pi\,(1-\pi)}{n}}$.

En el cas exposat, l'interval de probabilitat per a la proporció de persones amb pes superior al normatiu és, per a un valor α=0,05:

$$I.P.= 0{,}49 \pm 1{,}96\sqrt{\frac{0{,}49\,(1-0{,}49)}{50}} = (0{,}351 - -0{,}629)$$

El valor de 0,42 obtingut a la mostra es troba dins de l'interval de probabilitat 0,95, per tant, podem entendre que el resultat obtingut a la mostra no difereix significativament de la proporció poblacional, és a dir, que els resultats a la ciutat estudiada no són diferents dels de la població general.

Gràficament, la distribució mostral de proporcions (amb π=0,49, i error estàndard $\sqrt{\frac{0{,}49\,(1-0{,}49)}{50}} = 0{,}071$) té la forma següent:



Les zones en gris són les que queden fora de l'interval de probabilitat calculat. És obvi que el valor 0,42 es troba dins de l'interval, tal i com ja s'havia especificat.

Alternativament, podem calcular el valor z corresponent a la proporció mostral obtinguda. Això es faria de la forma següent:

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

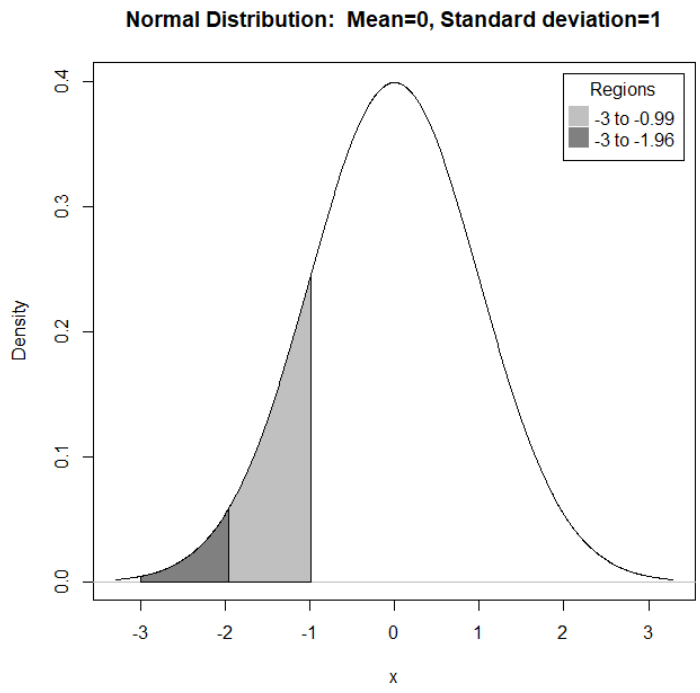
Com ja és conegut, l'obtenció d'una puntuació z ens permet establir la probabilitat del resultat obtingut. En el cas plantejat, el valor *z* seria:

$$z = \frac{0{,}42 - 0{,}49}{\sqrt{\frac{0{,}49(1-0{,}49)}{50}}} = -0{,}986$$

La probabilitat d'obtenir un valor *z*=-0,986 és, segons les taules de la distribució normal tipificada, d'aproximadament 0,16. Mitjançant una aplicació estadística podem calcular la probabilitat exacta, que és de 0,162.

A l'estudi indicat no s'ha fet cap predicció respecte de si la proporció de persones amb pes per sobre del normal a la mostra seria superior o inferior a la poblacional i, per tant, es planteja una hipòtesi bilateral. En conseqüència, si acceptem el nivell de risc habitual (α=0,05) caldrà utilitzar com a referència α/2=0,025. Ja que 0,162 > 0,025, podem concloure que no hi ha una diferència significativa entre els resultats de la mostra i el corresponent paràmetre poblacional.

Igualment, podem veure que la puntuació tipificada obtinguda (z= -0,986) és superior al valor de referència per a α/2, que és -1,96 (s'utilitza el valor negatiu ja que la puntuació obtinguda és negativa i, per tant, ens movem a la zona esquerra de la distribució). Gràficament:



La zona més fosca correspon a α/2=0,025, i el seu límit és *z*=-1,96. El valor obtingut a l'estudi és de -0.986 (zona marcada en gris clar). Queda clar que el valor obtingut queda fora de la "zona de significació", i ens indica que la diferència observada amb la proporció poblacional (que ocuparia el valor central de la distribució, z=0) no és significativa.

. Cas de més de dues proporcions (multinomial)

Imaginem que l'estudi anterior es realitza amb una major precisió, i que en lloc de classificar únicament les persones en dues categories (excés de pes o no excés de pes), ho fem en les quatre categories habituals de l'IMC: Inferior al normal (IN), Normal (N), Sobrepès (S) i Obesitat (O).

Suposem també que el conjunt d'estudis realitzats anteriorment ens permet tenir una bona aproximació de la situació al conjunt de la població, i que les proporcions poblacionals per a les diferents categories són les següents:

Categories	**IN**	**N**	**S**	**O**	**Total**
Proporcions	0,099	0,411	0,352	0,138	1

Imaginem també que novament es realitza un estudi a una ciutat determinada, en aquest cas amb una mostra de 125 casos, i que els resultats obtinguts són els següents:

Categories	**IN**	**N**	**S**	**O**	**Totals**
Nº de casos	9	55	39	22	125
Proporcions	0,072	0,440	0,312	0,176	1

La pregunta a realitzar a partir d'aquí és si la distribució de casos a les diferents categories a la mostra obtinguda a la ciutat estudiada és o no significativament diferent de les corresponents proporcions poblacionals.

Les hipòtesis nul·la i alternativa en aquest cas s'expressarien com:

H_0: $p_c = \pi$, per a totes les categories

H_1: $p_c \neq \pi$, com a mínim per a una de les categories

Tot i que seria possible utilitzar el mateix procediment que en el cas de 2 categories (fent la valoració categoria per categoria), en el cas multinomial és més senzill i pràctic recórrer a una nova distribució de probabilitat, concretament la distribució χ^2 (chi-quadrat, o khi-quadrat) de Pearson.

Pearson va demostrar que si comparem el número de casos obtingut a cada categoria amb els que s'haurien d'haver obtingut en cas de compliment de H_0, el resultat d'aquesta comparació segueix una distribució χ^2 amb v=(k-1) graus de llibertat, on k és el número de categories.

Per fer la comparació necessitem doncs saber quants casos hauríem d'haver trobat a cada categoria si la distribució en categories fos exactament igual a la poblacional. Això és fàcil de calcular, ja que disposem de les proporcions poblacionals:

Categories	IN	N	S	O	Total
Proporcions	0,099	0,411	0,352	0,138	1
Casos teòrics	12,375	51,375	44	17,25	125
Casos reals	9	55	39	22	125

Coneixent doncs les freqüències reals observades (n_o) i les teòriques o esperades (n_e) per a cada categoria, es pot calcular l'estadístic χ^2 mitjançant l'expressió següent:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_{oi} - n_{ei})^2}{n_{ei}}$$

$n_{oi} =$ Número de casos observat a la categoria i

$n_{ei} =$ Número de casos esperat a la categoria i

En el cas plantejat:

	IN	N	S	O	Total
n observat	9	55	39	22	125
n esperat	12,375	51,375	44	17,25	125
$\mathbf{n_o - n_e}$	-3,375	3,625	-5	4,75	
$\mathbf{(n_o - n_e)^2}$	11,391	13,141	25	22,562	
$\mathbf{(n_o - n_e)^2/n_e}$	0,920	0,256	0,568	1,308	**3,052**

En conseqüència, l'estadístic de contrast chi-quadrat és χ^2 = 3,052. Els graus de llibertat són v=(4-1)=3.

Cal indicar que per utilitzar correctament aquesta prova cal que les freqüències esperades per a totes les categories siguin superiors a 5, cosa que es compleix clarament en aquest cas.

A partir d'aquí, com en tots els casos anteriors, tenim diferents formes de determinar la significació (o no) de l'estadístic obtingut:

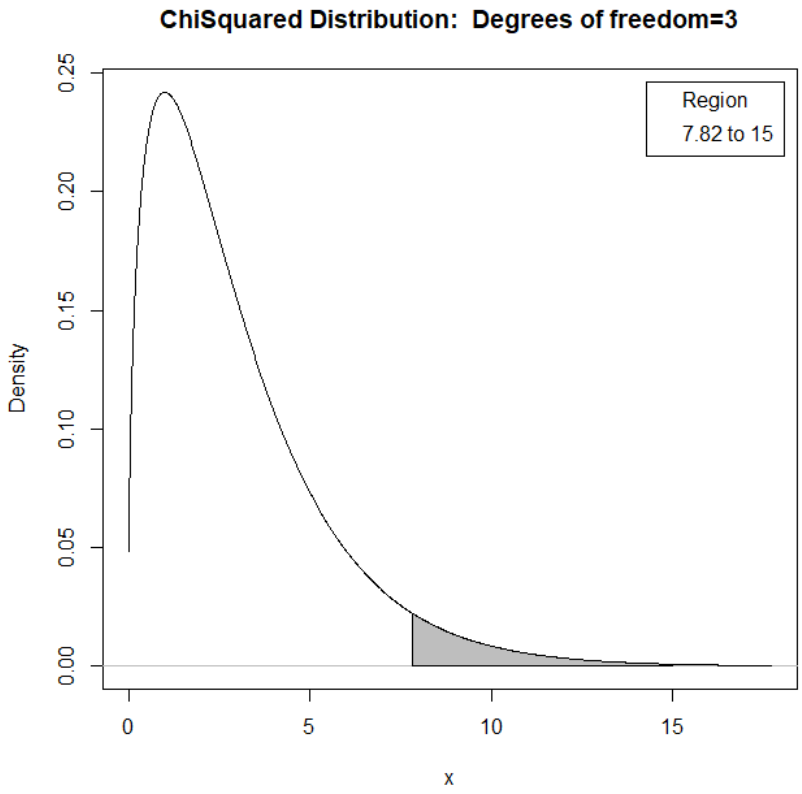
. Taules de la distribució χ^2

Amb 3 graus de llibertat i α=0,05, el valor de les taules és de 7,815. Ja que 3,052 < 7,815, podem concloure que la distribució en categories a la mostra no és significativament diferent a la que trobem a la població.

. Càlcul de la probabilitat exacta

D'acord amb l'aplicació estadística R, la probabilitat exacta d'obtenir per atzar un valor χ^2 de 3,052 o superior és p=0,384. Es tracta d'una probabilitat molt alta, i en tot cas molt superior al valor de referència α=0,05, que marcaria la zona de significació. Això ens indica que és molt probable que el resultat de 3,052 s'hagi obtingut per atzar i, per tant, que no hi hagi diferència real entre els resultats de la mostra i els de la població general

Gràficament:



La zona grisa és la delimitada per valor 7,815, és a dir, la corresponent al risc α=0'05. Si el valor obtingut a la prova estigués dins d'aquesta zona, hauríem de considerar el resultat com a significatiu, ja que la probabilitat de que sigui degut a l'atzar seria molt baixa. Com que no és el cas, ja que el valor 3,052 està clarament fora de la zona indicada, podem concloure que el resultat obtingut no és estadísticament significatiu, d'acord amb els criteris habituals.

6.2 Comparació entre proporcions mostrals

En aquesta situació el que es vol fer és comparar per a diverses mostres les proporcions segons les quals es distribueixen els casos en un conjunt de categories comú, o bé comparar els resultats per a una mateixa mostra en dues situacions o moments diferents. Com passava en el cas de les mitjanes, podem distingir doncs entre el cas de dades independents i de dades aparellades.

. Cas de dades independents

Implica, com ja s'ha dit, la comparació dels resultats de dues o més mostres. Normalment, les mostres que es comparen difereixen en alguna característica que podria relacionar-se amb distribucions diferents de casos a les diverses categories. Exemples: Distribució de grups sanguinis en homes i dones; distribució en categories d'IMC en diferents grups socio-demogràfics; distribució de tipologies d'hàbits alimentaris en grups de diferent nivell d'estudis, acceptació de dos productes diferents mesurada mitjançant una enquesta, etc.

Com es pot veure, molt habitualment la comparació entre proporcions mostrals el que ens permet fer és establir si existeix o no una relació entre dues variables categòriques, i aquesta és precisament la seva principal utilitat.

Exemple: Suposem que es vol analitzar si la valoració d'un determinat producte alimentari entre el públic difereix entre homes i dones. Per fer-ho s'administra una enquesta on es demana classificar el producte en una de quatre categories: Deficient – Correcte – Bo – Excel·lent. Es recullen els resultats per a 165 homes i 175 dones, amb els resultats següents:

	Deficient	Correcte	Bo	Excel·lent	Totals
Homes	50	58	39	18	165
Dones	33	53	59	30	175
Totals	83	111	98	48	340

Com ja s'ha indicat a l'apartat 2.1, aquesta taula de doble entrada, que mostra les freqüències per a les diverses caselles configurades per les dues variables categòriques, s'anomena *taula de contingència*, i els totals de fila i de columna s'anomenen *marginals*.

La hipòtesi nul·la en aquest cas és que les dues variables categòriques són independents entre sí. Això voldria dir que la distribució d'opinions a les diferents categories de l'enquesta no seria significativament diferent entre homes i dones.

La forma de valorar l'acceptació o rebuig de la hipòtesi nul·la és novament a través de la distribució de probabilitat χ^2. Com passava en el cas de la comparació entre proporcions a una mostra i proporcions poblacionals, el càlcul de l'estadístic es basa en la comparació entre les freqüències o número de casos observats i els que podríem esperar en cas de compliment de la hipòtesi nul·la.

Per tal d'arribar a obtenir el valor de l'estadístic, cal realitzar un procés en el qual la informació bàsica és la que trobem a la taula de contingències. El número de casos esperat a una determinada casella, situada a la fila *i* i columna *j*, és el següent:

$$n_{eij} = \frac{F_i \cdot C_j}{N}$$

En aquesta expressió, F_i és el marginal de la fila *i*, C_j és el marginal de la columna *j*, i N és el número total de casos.

Per exemple, per a la primera casella (Homes-Deficient), la freqüència esperada és:

$$n_{e11} = \frac{F_1 \cdot C_1}{N} = \frac{165 \cdot 83}{340} = 40{,}28$$

De la mateixa forma podem obtenir els valors esperats per a la resta de caselles. Per a més facilitat, s'inclouran entre parèntesi els resultats de les freqüències esperades a la taula original:

	Deficient	Correcte	Bo	Excel·lent	Totals
Homes	50 (40,28)	58 (53,87)	39 (47,56)	18 (23,29)	165
Dones	33 (42,72)	53 (57,13)	59 (50,44)	30 (24,71)	175
Totals	83	111	98	48	340

Per tal de continuar amb l'aplicació de la prova, cal comprovar que totes les freqüències esperades són iguals o superiors 5, cosa que es compleix clarament en el cas exposat. En cas contrari, es poden plantejar algunes correccions a la prova χ^2 o bé recórrer a la prova exacta de Fisher. Aquestes alternatives no seran exposades aquí, ja que en els estudis empírics en el camp de l'alimentació no és habitual trobar situacions amb freqüències tan baixes.

Una vegada calculades les freqüències esperades, l'estadístic chi-quadrat es calcularà de la forma següent:

$$\chi^2 = \sum_{i=1}^{i} \sum_{j=1}^{j} \frac{(n_{oij} - n_{eij})^2}{n_{eij}}$$

Com es pot veure, es tracta de la mateixa fórmula utilitzada en el cas de comparació de proporcions mostrals amb poblacionals, amb l'única diferència que aquí es treballa amb dues variables i tenim un conjunt de (*i x j)* caselles, definides per les *i* files i *j* columnes. Per a cada casella cal fer la diferència entre freqüència observada i freqüència teòrica, elevar-la al quadrat i dividir-la per la freqüència teòrica. Posteriorment es fa el sumatori dels resultats per a totes les caselles.

Per tant, en el cas exposat:

$$\chi^2 = \frac{(50 - 40{,}28)^2}{40{,}28} + \frac{(58 - 53{,}87)^2}{53{,}87} + \cdots\cdots + \frac{(30 - 24{,}71)^2}{24{,}71} = \ 10{,}504$$

El resultat d'aquesta operació es distribueix segons una distribució χ^2amb graus de llibertat iguals a ν=(i-1)·(j-1), on *i* és el número de files i *j* és el número de columnes.

En el nostre cas: ν = (2-1)·(4-1) = 3

A partir d'aquí, es pot establir la significació o no de l'estadístic mitjançant els procediments habituals:

. Les taules de la distribució χ^2 amb 3 graus de llibertat i α=0,05 ens donen un valor límit de 7,815. Ja que 10,504 > 7,815, podem rebutjar la hipòtesi nul·la. Podem pensar, doncs, que les dues variables estan relacionades o que existeix alguna forma de dependència entre elles

. La probabilitat exacta del resultat obtingut és, d'acord amb les aplicacions estadístiques, p=0,0147. Això ens confirma que la probabilitat d'error en rebutjar la hipòtesi nul·la és molt baixa

Cal tenir sempre present quina és la hipòtesi que s'està contrastant. La hipòtesi nul·la ens indicava que les dues variables analitzades eren independents. Al rebutjar-la, assumim com a hipòtesi alternativa que existeix algun tipus de relació entre les dues variables. La inspecció del quadre de resultats ens indica que les dones tendeixen a valorar millor aquest producte que els homes, i la prova estadística ens mostra que aquesta diferència no és fruit de l'atzar sinó que, molt probablement, és una diferencia real, les causes de la qual s'haurien d'analitzar

. Cas de dades aparellades

En aquesta situació, un mateix grup de persones o casos és mesurat en dos moments o situacions diferents, i el que es vol valorar és si els resultats són o no diferents en aquestes dues mesures, és a dir, si s'han produït canvis entre els dos registres.

Exemple: Suposem que es vol valorar si un determinat programa educatiu és eficaç a l'hora de modificar els hàbits alimentaris en estudiants de secundària. Imaginem també que es comença administrant una enquesta sobre hàbits d'alimentació a una mostra d'estudiants, i que a partir dels resultats obtinguts se'ls classifica en dues categories, en funció de si mostren o no hàbits alimentaris saludables. Posteriorment s'introdueix un programa informatiu i educatiu durant 3 mesos, i una vegada finalitzat es torna a administrar l'enquesta al mateix grup de persones, i es tornen a classificar en les dues categories citades.

Com és evident, en aquest cas la hipòtesi nul·la és que no hi ha canvis entre la primera mesura i la segona, cosa que indicaria que el programa educatiu utilitzat no ha produït canvis en el comportament alimentari dels subjectes.

Suposem que els resultats obtinguts són els següents:

		Després de la intervenció	
		Saludable	No saludable
Abans de la intervenció	Saludable	25	5
	No saludable	18	42

La lectura d'aquesta taula és: Hi ha 25 persones que estaven a la categoria d'hàbits saludables abans de la intervenció i que hi continuen estant després: hi ha 5 persones que han passat d'hàbits saludables a no saludables, etc.

La forma més habitual d'abordar aquest tipus de casos (dades aparellades amb dos moments o ocasions de registre) és a partir del *test o prova de McNemar*, una versió simplificada de les proves chi-quadrat vistes anteriorment.

La lògica de la prova de McNemar és la següent, en el cas de l'exemple proposat: Si la intervenció no ha tingut cap efecte, els canvis que s'hagin produït entre la primera mesura i la segona seran producte de l'atzar; per tant, hauríem d'observar aproximadament tants canvis en una direcció (No saludable-Saludable) com en l'altra (Saludable-No saludable). Per tant, en el test de McNemar només s'utilitza la informació inclosa a les caselles que impliquen canvi. En el nostre cas, 5 persones han passat d'hàbits saludables a no saludables, mentre que 18 han passat d'hàbits no saludables a saludables. En total, doncs, 23 persones han canviat de categoria. Si la hipòtesi nul·la fos certa, aquests canvis s'haurien produït aleatòriament en una o altra direcció, per tant, estrictament hi hauria d'haver la meitat de canvis en la direcció No saludable – Saludable i l'altra meitat en direcció contrària. És a dir, la meitat dels 23 canvis (o sigui, 11,5) haurien de ser en una direcció i 11,5 en l'altra (òbviament, l'ús d'un número decimal en aquest cas és fictici, però és necessari en haver-hi un número de canvis imparell). En altres paraules, aquesta xifra teòrica de canvis és la freqüència esperada en cas que es compleixi la hipòtesi nul·la.

Si anomenem A al número de canvis en una direcció (per exemple, de Saludable a No Saludable) i B al número de canvis en direcció contrària, llavors les hipòtesis nul·la i alternativa es poden re-expressar de forma simple:

H_0: $A = B$

H_1: $A \neq B$ (bilateral). En el nostre cas, seria més lògic pensar que $B > A$, si el tractament és eficaç.

La informació disponible es pot representar en una taula molt senzilla:

	Canvis A	Canvis B	Total
Observats	5	18	23
Esperats	11,5	11,5	

En aquest cas és aplicable l'estadístic χ^2 que ja coneixem, a partir de la comparació entre freqüències observades i freqüències teòriques. Ara bé, en els casos (com el que tenim) amb una taula 2 x 2 i amb freqüències esperades iguals, l'expressió per obtenir l'estadístic es pot simplificar molt:

$$\chi^2 = \frac{(A-B)^2}{A+B}$$

En el nostre cas:

$$\chi^2 = \frac{(5-18)^2}{5+18} = 7{,}35$$

Aquest quocient es distribueix segons una llei χ^2 amb 1 grau de llibertat (ja que en qualsevol taula 2 x 2, els graus de llibertat són v = (2-1)·(2-1) = 1).

A partir d'aquí, l'establiment de la significació o no del canvi es fa de la forma habitual:

. El valor crític de la distribució χ^2 per a 1 grau de llibertat i α=0,05 és 3,841. Ja que 7,35 > 3,841, hem de rebutjar la hipòtesi nul·la i podem pensar que hi ha hagut un impacte del programa educatiu utilitzat. Les dades ens indiquen que hi ha més casos de canvi No saludable – Saludable que en sentit contrari, i que aquesta diferència és estadísticament significativa, per tant constatem un efecte positiu del programa

. La probabilitat exacta associada al valor χ^2 = 7,35 amb 1 grau de llibertat és p=0,0067. Per tant, aquesta és la (petita) probabilitat d'error que acceptem si rebutgem la hipòtesi nul·la. Això ens indica novament que la diferència obtinguda és significativa i que podem considerar que el programa ha tingut un efecte real sobre els resultats