07
Eina-e

Estadística pràctica pas a pas

Josep Maria Mateo Sanz

[publicacions] urv

Estadística pràctica pas a pas

Eina-e, 7

Edita:
Publicacions URV

1a edició: Agost de 2011
ISBN: 978-84-694-6149-5
Dipòsit legal: T-1012-2011

Publicacions de la Universitat Rovira i Virgili:
Av. Catalunya, 35 - 43002 Tarragona
Tel. 977 558 474 - Fax: 977 558 393
www.publicacionsurv.cat
publicacions@urv.cat

Estadística pràctica pas a pas

Josep Maria Mateo Sanz

[publicacions] urv

Tarragona, 2011

Índex de continguts

Presentació

L'objectiu principal d'aquest material és proporcionar als estudiants una eina, estructurada en format de llibre, que els permeti consultar clarament i ràpidament com es pot aplicar de manera pràctica la teoria relacionada amb qualsevol dels continguts que formen part d'un curs bàsic d'estadística de nivell universitari.

En aquest material es defuig tant com és possible un llenguatge massa tècnic i teòric, el qual és difícil d'entendre pels estudiants, i s'usa un llenguatge més planer i en què es fa èmfasi especialment en l'aplicació de les diverses tècniques estadístiques. S'explica quan i com es poden aplicar aquestes tècniques pas a pas, de manera que es guia l'alumnat en el procediment que cal seguir en cada situació. El material està farcit d'exemples on també s'ha desenvolupat la solució seguint els passos marcats a la part teòrica. S'utilitza aquesta estructura per afavorir l'aprenentatge; s'assisteix l'estudiant en la resolució dels exercicis que es proposen en l'assignatura, ja que es marca el camí que ha de seguir per aplicar cada tècnica.

En aquest material també s'explica quines són les funcions i les eines estadístiques d'Excel que es poden fer servir per automatitzar els càlculs que es requereixen per aplicar les diverses tècniques estadístiques.

1. Estadística descriptiva

1.1 Concepte d'*estadística*. Contingut de l'estadística

Definició. L'**estadística** és la ciència, el mètode, les tècniques, l'operació d'anàlisi matemàtica que permeten estudiar numèricament amb el màxim de precisió els fenòmens col·lectius incompletament coneguts.

El contingut de l'estadística es pot dividir en dos grans grups:

- Estadística descriptiva
- Estadística inferencial (o inductiva)

Definició. L'**estadística descriptiva** estudia la manera d'ordenar i analitzar totes les dades d'una població, amb l'objectiu d'obtenir conclusions sobre aquesta població.

Exemples. La direcció d'un centre escolar vol fer un estudi sobre els resultats acadèmics d'un curs determinat o es vol estudiar els resultats dels diferents equips de futbol de primera divisió durant els 10 últims anys. Aquests són problemes d'estadística descriptiva, ja que disposem de les dades de tots els elements que volem estudiar.

Definició. L'**estadística inferencial** té com a finalitat obtenir conclusions respecte d'una població, mitjançant l'anàlisi d'una mostra de la població.

Exemples. Es vol estudiar l'alçada de tots els catalans i només disposem de l'alçada de 1.000 persones o es vol fer un estudi sobre la durada de les bombetes d'una determinada marca i només disposem de la durada de 100 bombetes d'aquesta marca. Aquests són problemes d'estadística inferencial, ja que no disposem de les dades de tots els elements que volem estudiar, sinó d'una mostra.

1.2 Concepte de *població, mostra, individu i variable estadística*

Definició. S'anomena **població** el conjunt sobre el qual es vol portar a terme l'estudi estadístic.

Definició. S'anomena **mostra** qualsevol subconjunt de la població.

Definició. S'anomena **individu** qualsevol element del conjunt de la població.

Definició. S'anomena **variable estadística** la característica que es vol estudiar d'una població.

Exemple. Agafant l'exemple anterior sobre l'estudi de les bombetes, la població la formen totes les bombetes d'aquella marca: cada bombeta és un individu, les 100 bombetes de les quals sabem la durada formen una mostra de la població i la variable estadística que estem estudiant és la durada de les bombetes.

Observació. Depenent del que vulguem estudiar, un mateix conjunt pot ser una mostra o una població. Per exemple, si només tenim les notes dels alumnes que han seguit certs estudis a la universitat URV i volem treure conclusions sobre aquest grup en concret, aquest grup d'alumnes de la URV és la població; en canvi, si volem treure conclusions sobre els estudiants universitaris de Catalunya, el grup d'alumnes de la URV és una mostra i tots els estudiants universitaris de Catalunya són la població.

1.3 Classificació de les variables estadístiques

Les variables estadístiques les classificarem segons el tipus de valors que poden prendre. Una primera divisió és:

- Variables estadístiques qualitatives (o nominals)
- Variables estadístiques quantitatives (o numèriques)

Definició. Les **variables estadístiques qualitatives** són aquelles que no prenen valors numèrics.

Exemples. El color dels ulls, el sexe, el tipus de distracció preferit.

Definició. Les **variables estadístiques quantitatives** són aquelles que prenen valors numèrics. Dintre d'aquestes últimes encara podem fer una subdivisió:

- Variables estadístiques quantitatives discretes
- Variables estadístiques quantitatives contínues

Definició. Una variable estadística és **quantitativa discreta** quan entre dos valors qualssevol que pot prendre la variable només hi ha un nombre finit de possibles valors de la variable. També es pot dir que una variable estadística és quantitativa discreta

quan el nombre de possibles valors que pot prendre la variable és una quantitat numerable (finita o infinita), és a dir que els possibles valors es poden numerar i que després d'un possible valor sempre sé quin és el següent.

Exemples. El nombre de germans, el nombre de cotxes que passen en un dia per un cert punt, el nombre de trucades que es reben cada hora en una ciutat.

Definició. Una variable estadística és **quantitativa contínua** quan entre dos valors qualssevol que pot prendre la variable hi pot haver un nombre infinit de possibles valors de la variable. Aquesta definició és equivalent a dir que, si agafem dos valors que pot prendre la variable, tan propers com vulguem, sempre és possible trobar un altre valor de la variable que estigui entre els dos valors anteriors. Els possibles valors d'una variable quantitativa contínua són infinits i no numerables i després d'un possible valor no es pot concretar quin és el següent.

Exemples. L'alçada de les persones, el pes de les taronges, el temps que es tarda a fer un examen. S'ha d'observar, per exemple en el primer cas, que, si agafem les alçades de dues persones, sempre és possible trobar-ne una altra que tingui una alçada entre les dues anteriors. En aquest sentit, cal pensar que l'alçada exacta d'una persona és una quantitat amb infinites xifres decimals però que, per problemes amb la precisió dels aparells de mesura, la majoria de vegades només donem l'alçada en centímetres, és a dir, generalment discretitzem variables de naturalesa contínua.

Observació. Cal remarcar que una variable estadística quantitativa discreta no és només aquella que pot prendre un nombre finit de resultats, sinó que de vegades podrà prendre un nombre infinit de possibles resultats. Per exemple, en el cas del nombre de cotxes que passen per cert punt en un dia, la variable és discreta, però els possibles valors que pot prendre són infinits.

1.4 Distribució de freqüències. Representacions gràfiques

1.4.1 Taula de freqüències

Si observem una variable estadística sobre un conjunt d'individus, obtindrem una sèrie de dades (que poden estar repetides o no). Aquestes dades, si són quantitatives, les podem ordenar. Si són qualitatives, l'ordenació és arbitrària. Els valors de les dades ordenades els notarem com a $x_1, x_2, ..., x_k$, on x_1 és el valor de la dada més petita, x_2 el valor de la segona i així successivament.

Definició. La **freqüència absoluta** d'un valor és el nombre de vegades que apareix aquest valor a la sèrie. La freqüència absoluta del valor x_i la notarem amb n_i.

Definició. La **freqüència relativa** d'un valor x_i és el quocient entre la freqüència absoluta del valor i el nombre total de dades de la sèrie. Notarem amb N el nombre total de dades i amb f_i la freqüència relativa del valor x_i. Així obtenim: $f_i = \frac{n_i}{N}$.

Definició. La **freqüència absoluta acumulada** d'un valor x_i és la suma de totes les freqüències absolutes dels valors de la sèrie des del principi fins al valor x_i. Notarem amb N_i la freqüència absoluta acumulada del valor x_i. Així obtenim: $N_i = n_1 + n_2 + \ldots + n_i$. També es pot obtenir N_i fent $N_i = N_{i-1} + n_i$.

Definició. La **freqüència relativa acumulada** d'un valor x_i és la suma de totes les freqüències relatives dels valors de la sèrie des del principi fins al valor x_i. Notarem amb F_i la freqüència relativa acumulada del valor x_i. Així obtenim: $F_i = f_1 + f_2 + \ldots + f_i$. També es pot obtenir F_i fent $F_i = F_{i-1} + f_i$ o bé fent $F_i = \frac{N_i}{N}$.

Propietats de les freqüències

- La suma de les freqüències absolutes és igual al nombre total de dades de la sèrie.
- La freqüència relativa d'un valor sempre estarà entre 0 i 1.
- La suma de totes les freqüències relatives és igual a 1.
- Si multipliquem la freqüència relativa d'un valor per 100 obtindrem el tant per cent de vegades que es repeteix el valor dins de la sèrie.

Donada una sèrie de dades podem crear una taula de freqüències on apareguin els valors de les dades i tots els tipus de freqüències esmentades.

Exemple. S'ha llançat un dau 20 vegades i s'ha obtingut el resultat següent: 2, 3, 6, 6, 2, 4, 4, 4, 2, 5, 2, 4, 2, 4, 2, 5, 1, 6, 2 i 2. Fem la taula de freqüències:

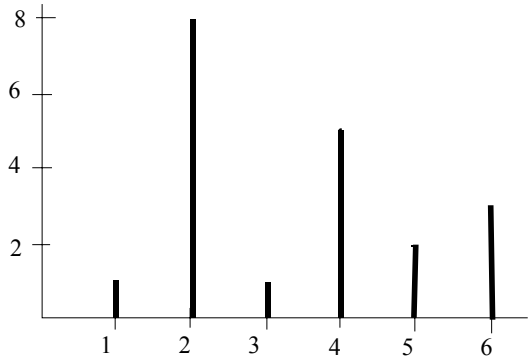
Valor (x_i)	*Freq. abs.* (n_i)	*Freq. rel.* (f_i)	*Freq. abs. ac.* (N_i)	*Freq. rel. ac.* (F_i)
1	1	0.05	1	0.05
2	8	0.4	9	0.45
3	1	0.05	10	0.5
4	5	0.25	15	0.75
5	2	0.1	17	0.85
6	3	0.15	20	1

1.4.2 Representació gràfica

Diagrama de barres. L'usarem quan els valors de la variable estadística siguin donats de manera individual. Normalment representarem les freqüències absolutes, les freqüències relatives o el tant per cent. Per fer-ho hem de dibuixar dos eixos. A l'eix horitzontal hem de posar-hi els valors de la variable de manera ordenada. A l'eix vertical posarem

una escala adient segons el que vulguem representar. El gràfic es construeix aixecant sobre cada valor de la variable un segment vertical de llargada igual a la freqüència que vulguem representar.

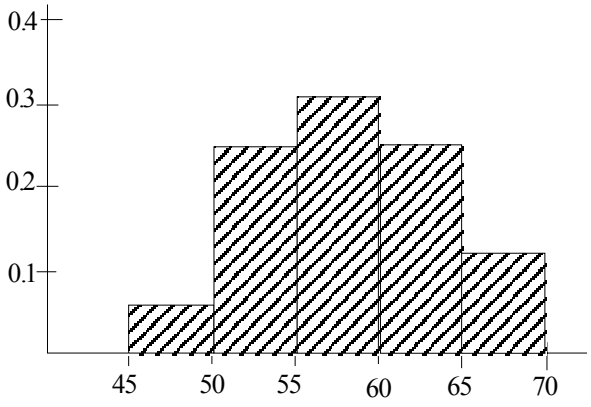
Exemple. Fem un diagrama de barres per representar les freqüències absolutes amb les dades de l'exemple anterior:



Histograma. L'usarem quan els valors de la variable estadística estiguin donats en forma d'interval. En aquest cas suposarem que els diferents intervals tenen la mateixa amplada. L'única diferència amb el diagrama de barres és que a l'eix horitzontal hem de posar-hi els valors dels extrems dels intervals de manera ordenada i el gràfic es construeix aixecant sobre cada interval un rectangle vertical d'altura igual a la freqüència que vulguem representar.

Exemple. Fem un histograma per representar les freqüències relatives de les dades següents agrupades en intervals corresponents als pesos (en kg) de 16 alumnes d'una classe de secundària.

Valor (x_i)	*Freq. abs.* (n_i)	*Freq. rel.* (f_i)	*Freq. abs. ac.* (N_i)	*Freq. rel. ac.* (F_i)
45-50	1	0.0625	1	0.0625
50-55	4	0.25	5	0.3125
55-60	5	0.3125	10	0.625
60-65	4	0.25	14	0.875
65-70	2	0.125	16	1



Observació. Si els intervals no tenen la mateixa amplada, l'altura de cada rectangle es calcula dividint la freqüència de l'interval que vulguem representar entre la seva amplada.

1.5 Agrupació de dades en intervals

De vegades, quan el nombre de valors diferents que hi ha en una sèrie de dades estadístiques és gran, convé agrupar les dades en intervals. En aquest cas, només estudiarem el cas en què els intervals tinguin la mateixa amplada. Per agrupar les dades distingirem dos casos segons que la variable estadística que estem estudiant sigui quantitativa contínua o quantitativa discreta. De vegades hi haurà variables discretes que tractarem com a contínues si el nombre de valors que comprèn és molt gran.

Quan agrupem les dades en intervals, apareix un nou concepte que és la marca de classe.

Definició. La **marca de classe** d'un interval és el nombre que representa l'interval. Aquest nombre pot ser qualsevol que estigui dins de l'interval, però normalment s'agafa el punt mitjà de l'interval (que és el que farem a partir d'ara). El punt mitjà d'un interval el podem obtenir sumant els extrems de l'interval i dividint entre 2. La marca de classe la notarem amb x_i.

1.5.1 Variable quantitativa contínua

El primer que cal fer és determinar el nombre d'intervals que volem fer amb les dades de la sèrie estadística. Aleshores, l'amplada dels intervals la podem calcular segons la fórmula següent:

$$\text{Amplada} = \frac{\text{Valor màxim - Valor mínim}}{\text{Nombre d'intervals}}$$

Els intervals els obtindrem sumant successivament l'amplada a partir del valor mínim. D'aquesta manera, l'extrem inferior d'un interval coincidirà amb l'extrem superior de l'interval anterior.

Observacions

- A l'hora d'assignar les freqüències absolutes a cada interval, podem tenir dubtes amb les dades que coincideixen amb els extrems dels intervals, ja que no sabem en quin interval posar-les. Per aquest motiu cal indicar en quin interval

inclourem aquestes dades dubtoses. Això ho farem mitjançant els claudàtors [,], per indicar que l'extrem és inclòs a l'interval, i els parèntesis (,), per indicar que l'extrem no és inclòs a l'interval. Cal que tots els extrems superiors estiguin inclosos (i els inferiors exclosos) o que tots els extrems inferiors estiguin inclosos (i els superiors exclosos). El valor més petit i el més gran han d'estar sempre inclosos.

- Si l'amplada no dóna un nombre exacte, podem arrodonir-la per excés a una quantitat adient i començar a fer intervals abans del valor mínim i acabar després del valor màxim.

Exemple. Les dades següents corresponen als pesos (en kg) de 16 alumnes d'una classe de secundària: 55, 55.5, 70, 60, 54.5, 54, 63, 54, 70, 64, 56, 52, 62.5, 57, 45 i 55. Amb aquestes dades farem 5 intervals de la mateixa amplada i inclourem els extrems superiors en els intervals.

$$\text{Amplada} = \frac{\text{Valor màxim - Valor mínim}}{\text{Nombre d'intervals}} = \frac{70 - 45}{5} = 5$$

Pesos	*Marca de classe* (x_i)	*Freq. abs.* (n_i)
[45,50]	47.5	1
(50,55]	52.5	6
(55,60]	57.5	4
(60,65]	62.5	3
(65,70]	67.5	2

1.5.2 Variable quantitativa discreta

L'amplada dels intervals la podem calcular segons la fórmula següent:

$$\text{Amplada} = \frac{\text{Valor màxim - Valor mínim} + 1}{\text{Nombre d'intervals}}$$

En aquest cas, l'amplada indicarà el nombre de valors que s'inclouran a cada interval. El primer interval començarà pel valor més petit i acabarà al valor obtingut de sumar el valor mínim amb l'amplada menys 1. El següent interval començarà al valor següent del valor amb què ha acabat l'interval anterior i acabarà al valor obtingut de sumar l'extrem inferior amb l'amplada menys 1. I així successivament.

Observacions

- En aquest cas, no hi pot haver dubtes en l'assignació de freqüències absolutes ja que l'extrem inferior d'un interval no coincideix amb l'extrem superior de l'interval anterior.
- Si l'amplada no dóna un nombre exacte, podem arrodonir-la per excés a una quantitat adient i començar a fer intervals abans del valor mínim i acabar després del valor màxim.

Exemple. Les dades següents corresponen al mes en què van néixer 20 alumnes d'una classe d'universitat: 8, 4, 8, 5, 2, 11, 2, 3, 9, 10, 12, 11, 5, 5, 12, 4, 6, 1, 7 i 2. Amb aquestes dades, farem 4 intervals de la mateixa amplada.

$$\text{Amplada} = \frac{\text{Valor màxim - Valor mínim} + 1}{\text{Nombre d'intervals}} = \frac{12 - 1 + 1}{4} = 3$$

Mesos	*Marca de classe* (x_i)	*Freq. abs.* (n_i)
1-3	2	5
4-6	5	6
7-9	8	4
10-12	11	5

1.6 Mesures de posició

En els següents subapartats estudiarem les mesures de posició. La majoria d'aquestes mesures només té sentit aplicar-les sobre variables quantitatives. En els tres primers seran mesures de posició central, i el quart subapartat es dedicarà a mesures de posició no central. Les mesures de **posició central** tenen per objectiu resumir una sèrie de dades estadístiques en un sol nombre. Les mesures de **posició no central** són valors que divideixen la sèrie en parts iguals. Totes les mesures de posició agafen valors que estan entre el valor mínim i el valor màxim. Les mesures de posició central que veurem són: la mitjana, la mediana i la moda.

1.6.1 Mitjana aritmètica

Definició. La **mitjana aritmètica** d'una sèrie estadística és la suma de totes les dades de la sèrie dividida pel nombre total de dades. Simplificant, l'anomenarem *mitjana* i la notarem amb $\bar{x}$ (o μ). La fórmula per trobar la mitjana és:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{N} = \frac{\sum_{i=1}^{k} x_i \; n_i}{N}$$

Observació. Si les dades estan donades en intervals, hem de treballar amb les marques de classe.

Exemple. Agafem les dades de l'exemple sobre el resultat del llançament d'un dau:

Valor (x_i)	*Freq. abs.* (n_i)
1	1
2	8
3	1
4	5
5	2
6	3

La mitjana serà:

$$\bar{x} = \frac{1 \cdot 1 + 2 \cdot 8 + 3 \cdot 1 + 4 \cdot 5 + 5 \cdot 2 + 6 \cdot 3}{20} = 3.4$$

Observació. Un altre concepte relacionat és el de **mitjana ponderada,** que es produeix quan a cada valor de la sèrie li donem una importància diferent. Aquesta importància es mesura segons una ponderació de cada valor. La fórmula per trobar la mitjana ponderada és:

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + \ldots + x_k w_k}{w_1 + w_2 + \ldots + w_k} = \frac{\sum_{i=1}^{k} x_i \; w_i}{\sum_{i=1}^{k} w_i}$$

Exemple. En una assignatura s'ha de presentar un treball que es valora en 1, s'han de lliurar uns problemes que es valoren en 2 i s'ha de fer un examen que es valora en 3. La nota del treball és 5, la dels problemes és 8 i la de l'examen és 7. Per trobar la nota final hem de fer la mitjana ponderada segons la importància que s'ha donat a cada part.

Valor (x_i)	*Ponderació* (w_i)
5	1
8	2
7	3

La mitjana ponderada serà:

$$\bar{x} = \frac{5\cdot 1 + 8\cdot 2 + 7\cdot 3}{1+2+3} = 7$$

1.6.2 Mediana

Definició. La **mediana** és la dada que ocupa la posició del mig a la sèrie una vegada s'han ordenat les dades, de més petita a més gran, tenint en compte les repeticions.

Observacions

- Si hi ha un nombre senar de dades, hi haurà una única dada que estigui al mig. Aquesta dada és la que ocupa la posició $\frac{N+1}{2}$
- Si hi ha un nombre parell de dades, n'hi haurà dues que seran al mig; en aquest cas, la mediana la trobarem fent la mitjana d'aquestes dues dades. Les dues dades ocupen la posició $\frac{N}{2}$ i $\frac{N}{2}+1$.
- També es pot calcular la mediana quan les dades estan agrupades en intervals.

Exemple. Agafem les dades de l'exemple sobre el resultat del llançament d'un dau:

Valor (x_i)	*Freq. abs.* (n_i)	*Freq. abs. ac.* (N_i)
1	1	1
2	8	9
3	1	10
4	5	15
5	2	17
6	3	20

La mediana serà la mitjana dels valors que són a les posicions 10 i 11. Aquests valors són el 3 i el 4; per tant, la mediana és $Me = 3.5$.

1.6.3 Moda

Definició. La **moda** és el valor de la dada que es repeteix més vegades a la sèrie. Aquesta mesura es pot aplicar tant a variables quantitatives com qualitatives.

Observacions

- Si les dades estan agrupades en intervals, parlarem de l'*interval modal*, que serà l'interval on hi ha més dades.

- Si hi ha empat en el nombre de repeticions entre dos o més valors, hi haurà més d'una moda.

Exemple. Agafem les dades de l'exemple sobre el resultat del llançament d'un dau:

Valor (x_i)	*Freq. abs.* (n_i)
1	1
2	8
3	1
4	5
5	2
6	3

La moda és el 2, ja que es repeteix 8 vegades. $Mo = 2$.

1.6.4 Mesures de posició no central: percentils

Com s'ha comentat abans, les mesures de posició no central són les que divideixen la sèrie de dades en parts iguals. Entre aquestes trobem:

- Els **quartils,** que són els tres valors que divideixen la sèrie en quatre parts iguals (a cada part hi haurà el 25% de la sèrie).
- Els **decils,** que són els nou valors que divideixen la sèrie en 10 parts iguals (a cada part hi haurà el 10% de la sèrie).
- Els **centils o percentils,** que són els 99 valors que divideixen la sèrie en 100 parts iguals (a cada part hi haurà l'1% de la sèrie).

De fet, tot es pot reduir a percentils, ja que el primer quartil es correspon amb el percentil 25 o el quart decil es correspon amb el percentil 40 i així es pot fer amb els altres quartils i decils. La mediana es correspon amb el percentil 50. Per això, l'estudi més complet el farem sobre els percentils.

Procediment i (per a dades no agrupades en intervals)

El percentil 1 (P_1) és el valor que supera l'1% de les dades d'una sèrie i és superat pel 99% restant de les dades d'aquella sèrie. El percentil 2 (P_2) és el valor que supera el 2% de les dades d'una sèrie i és superat pel 98% restant de les dades d'aquella sèrie. En general, el percentil i (P_i) és el valor que supera l'i% de les dades d'una sèrie i és superat pel (100 – i)% restant de les dades d'aquella sèrie. Per tant, si el percentil i ocupa la posició x cal que:

$$100\frac{x-1}{N-1}=i$$

ja que les $x-1$ dades que són superades per la dada que ocupa la posició x han de representar un i% de les dades de la sèrie sense comptar la mateixa dada que ocupa la posició x (per això es divideix entre $N-1$).

De l'expressió anterior, obtenim que la posició que ocupa el percentil i serà:

$$x=\frac{i\cdot(N-1)}{100}+1$$

Observació. Si la posició marcada per algun percentil no coincideix exactament amb la posició d'algun valor de la sèrie de dades, es pot realitzar una interpolació per determinar exactament el valor del percentil.

Exemple. Agafem les dades de l'exemple sobre el resultat del llançament d'un dau:

Valor (x_i)	*Freq. abs.* (n_i)
1	1
2	8
3	1
4	5
5	2
6	3

Si ens demanen el quartil 1, podem buscar el percentil 25, P_{25}, ja que és el mateix. El P_{25} ocupa la posició $x=\frac{25\cdot 19}{100}+1=5.75$ (entre la posició 5 i la posició 6). Amb la sèrie de dades ordenada veiem que la posició 5 correspon a un 2 i la posició 6 també correspon a un 2. Per tant, el percentil 25 serà igual a 2 ($P_{25}=2$).

Si ens demanen el percentil 45, P_{45}, primer hem de mirar quina posició ocupa. El P_{45} ocupa la posició $x=\frac{45\cdot 19}{100}+1=9.55$ (entre la posició 9 i la posició 10). Amb la sèrie de dades ordenada veiem que la posició 9 correspon a un 2 i la posició 10 correspon a un 3. Interpolant, agafem la diferència entre la posició x i l'enter més petit que x ($9.55-9=0.55$) i la multipliquem per la diferència entre els dos valors marcats per les dues posicions enteres al voltant de la posició x ($3-2=1$). Al resultat de la multiplicació

$(0.55 \cdot 1 = 0.55)$ se li ha d'afegir el valor corresponent a la posició entera més petita que x per tal de trobar el P_{45}. És a dir, $P_{45} = 2 + 0.55 = 2.55$.

Procediment 2 (per a dades agrupades en intervals)

Si les dades estan agrupades en intervals, el percentil i (P_i) el trobarem mitjançant la fórmula:

$$P_i = L_{j-1} + \frac{\frac{i(N-1)}{100} + 1 - N_{j-1}}{n_j} \cdot a_j$$

on L_{j-1} és l'extrem inferior de l'interval on és la dada $\frac{i \cdot (N-1)}{100} + 1$, n_j és la freqüència absoluta d'aquest interval, a_j és l'amplada d'aquest interval i N_{j-1} és la freqüència absoluta acumulada de l'interval anterior.

Exemple. Les dades següents corresponen a l'exemple dels pesos d'una classe de secundària. Busquem el P_{70}.

Valor (x_i)	*Freq. abs.* (n_i)	*Freq. abs. ac.* (N_i)
45-50	1	1
50-55	4	5
55-60	5	10
60-65	4	14
65-70	2	16

$$P_{70} = 60 + \frac{11.5 - 10}{4} \cdot 5 = 61.875$$

1.7 Mesures de dispersió

Les **mesures de dispersió** ens indiquen si les dades d'una sèrie estadística estan més o menys juntes. Aquestes mesures ens ajuden a especificar millor com és la sèrie de dades, ja que completen la informació facilitada per la mitjana o qualsevol altra mesura de posició central. Com més gran sigui el valor d'aquestes mesures, més dispersió hi haurà entre les dades. Totes aquestes mesures sempre prenen valors positius.

Exemple. Si calculem la nota mitjana de dos alumnes que han fet dos exàmens cadascú, en els quals el primer alumne ha tret dos 5 i el segon alumne ha tret un 0 i un 10, veiem que la nota mitjana és 5 en els dos casos. Però la dispersió de les notes és molt més gran en el segon alumne.

1.7.1 Recorregut

Definició. El **recorregut** és la diferència entre el valor màxim i el valor mínim d'una sèrie estadística.

1.7.2 Desviació mitjana

Definició. La **desviació mitjana** és la mitjana dels valors absoluts de les diferències entre els valors de la sèrie i la mitjana. La fórmula per trobar-la és:

$$D_{\bar{x}} = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| \cdot n_i}{N}$$

Observació. Si les dades estan agrupades en intervals, agafarem com a x_i la marca de classe.

1.7.3 Variància

Definició. La **variància** és la mitjana dels quadrats de les diferències entre els valors de la sèrie i la mitjana. La fórmula per trobar-la és:

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{\sum_{i=1}^{k} x_i^2 \cdot n_i}{N} - \bar{x}^2$$

Observacions

Si les dades estan agrupades en intervals, agafarem com a x_i la marca de classe.

Si el conjunt de dades sobre el qual es treballa correspon a una mostra de la població que es vol estudiar, la variància rep el nom de **variància mostral** i es nota i calcula segons la fórmula següent:

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{N-1} = \frac{N}{N-1}\sigma^2$$

1.7.4 Desviació típica o estàndard

Definició. La **desviació típica o estàndard** és l'arrel quadrada de la variància. D'aquesta manera s'obté una mesura en les mateixes unitats que les dades. La notarem amb σ.

Observació. Si el conjunt de dades sobre el qual es treballa correspon a una mostra de la població que es vol estudiar, la desviació estàndard rep el nom de **desviació estàndard mostral** i es nota amb la lletra *s* i es calcula fent l'arrel quadrada de la variància mostral.

1.7.5 Coeficient de variació

Definició. El **coeficient de variació** és el quocient entre la desviació típica i la mitjana. A diferència de les anteriors, aquesta és una mesura de dispersió relativa; les anteriors eren mesures de dispersió absolutes. La fórmula per trobar-lo és:

$$V = \frac{\sigma}{\bar{x}}$$

Exemple. Calculem totes les mesures de dispersió de les dades de l'exemple sobre el resultat del llançament d'un dau:

Valor (x_i)	*Freq. abs.* (n_i)
1	1
2	8
3	1
4	5
5	2
6	3

Recorregut = 6 – 1 = 5

$$\bar{x} = \frac{1\cdot 1 + 2\cdot 8 + 3\cdot 1 + 4\cdot 5 + 5\cdot 2 + 6\cdot 3}{20} = \frac{68}{20} = 3.4$$

$$D_{\bar{x}} = \frac{|1-3.4|\cdot 1+|2-3.4|\cdot 8+\ldots+|6-3.4|\cdot 3}{20} = \frac{28}{20} = 1.4$$

$$\sigma^2 = \frac{(1-3.4)^2\cdot 1+(2-3.4)^2\cdot 8+\ldots+(6-3.4)^2\cdot 3}{20} = \frac{48.8}{20} = 2.44$$

$$\sigma = 1.56$$

$$s^2 = \frac{(1-3.4)^2\cdot 1+(2-3.4)^2\cdot 8+\ldots+(6-3.4)^2\cdot 3}{19} = \frac{48.8}{19} = 2.57$$

$$s = 1.6$$

$$V = \frac{1.56}{3.4} = 0.46$$

Exemple. Calculem totes les mesures de dispersió de les dades de l'exemple corresponent als pesos d'una classe de secundària.

Pesos	*Marca de classe* (x_i)	*Freq. abs.* (n_i)
45-50	47.5	1
50-55	52.5	4
55-60	57.5	5
60-65	62.5	4
65-70	67.5	2

Recorregut = 70 – 45 = 25

$$\bar{x} = \frac{47.5\cdot 1+52.5\cdot 4+57.5\cdot 5+62.5\cdot 4+67.5\cdot 2}{16} = \frac{930}{16} = 58.125$$

$$D_{\bar{x}} = \frac{|47.5-58.125|\cdot 1+|52.5-58.125|\cdot 4+\ldots+|67.5-58.125|\cdot 2}{16} = \frac{72.5}{16} = 4.53$$

$$\sigma^2 = \frac{(47.5-58.125)^2\cdot 1+(52.5-58.125)^2\cdot 4+\ldots+(67.5-58.125)^2\cdot 2}{16} = \frac{493.75}{16} = 30.86$$

$$\sigma = 5.555$$

$$s^2 = \frac{(47.5 - 58.125)^2 \cdot 1 + (52.5 - 58.125)^2 \cdot 4 + \ldots + (67.5 - 58.125)^2 \cdot 2}{15} = \frac{493.75}{15} = 32.92$$

$$s = 5.74$$

$$V = \frac{5.555}{58.125} = 0.09557$$

1.8 Funcions d'Excel per calcular mesures estadístiques

Farem unes consideracions generals sobre quines funcions té Excel que facin operacions estadístiques i com s'han d'introduir aquestes funcions.

Abans d'usar una funció, generalment haurem d'haver posat les dades sobre les quals aplicarem la funció; per exemple, abans d'usar la funció que calcula la mitjana d'una sèrie de dades haurem d'haver posat les dades a Excel.

Per introduir una funció a Excel cal que ens trobem en una casella en blanc i anem a "Insertar –> Función". Veurem que les funcions es troben agrupades en categories. En aquesta assignatura usarem bàsicament les funcions de la categoria "Estadísticas".

Cada funció necessita uns arguments a partir dels quals donarà el resultat; per exemple, a la funció que calcula la mitjana d'una sèrie de dades cal introduir-hi com a argument sobre quines dades cal que calculi la mitjana.

Una vegada s'han introduït els arguments d'una funció, el programa retorna el resultat del càlcul demanat.

MESURES ESTADÍSTIQUES	*FUNCIONS D'EXCEL*
Mitjana, $\bar{x}$	PROMEDIO
Mediana	MEDIANA
Moda	MODA
Desviació mitjana, $D_{\bar{x}}$	DESVPROM
Variància poblacional, σ^2	VARP
Desviació estàndard poblacional, σ	DESVESTP
Variància mostral, s^2	VAR
Desviació estàndard mostral, s	DESVEST
Percentil, P_k	PERCENTIL

2. Variables aleatòries

2.1 Experiments aleatoris. Espai mostral. Successos

Definició. Es diu que un **experiment és aleatori, estocàstic o estadístic** si, quan es repeteix indefinidament en les mateixes condicions, no és possible predir el resultat, encara que coneguem les condicions inicials. En un experiment aleatori no coneixem el resultat fins que s'ha realitzat la prova.

Exemples. Són experiments aleatoris:

- L'extracció d'una carta de la baralla.
- El llançament d'un dau.
- L'extracció d'una bola de la loteria.
- El llançament d'una moneda.

No són experiments aleatoris:

- El resultat d'una reacció química.
- La velocitat d'arribada d'un cos a terra quan el deixem caure des d'una torre.

Definició. El conjunt de tots els resultats possibles que es poden obtenir amb un experiment aleatori s'anomena **espai mostral**. El notarem amb Ω.

Exemples. Considerarem ara diversos experiments i definirem els corresponents espais mostrals:

- Llançar una moneda i observar el costat que apareix.

$$\Omega = \{C, X\}$$

- Llançar 2 monedes i observar els costats que apareixen.

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

Observem que en aquest experiment el conjunt de resultats possibles consta de quatre elements. Així, CC és un sol resultat que té dos components, el mateix amb CX, etc.

- Llançar 3 monedes i observar els costats que surten.
$$\Omega = \{CCC,CCX,CXC,XCC,CXX,XCX,XXC,XXX\}$$
- Llançar un dau i observar la puntuació que apareix.
$$\Omega = \{1,2,3,4,5,6\}$$
- Llançar dos daus i observar la suma de les puntuacions que apareixen.
$$\Omega = \{2,3,4,5,6,7,8,9,10,11,12\}$$
- Comptar el nombre de persones que baixen d'un autobús en una parada determinada.
$$\Omega = \{0,1,2,3,...,N\}$$
- Extreure un nombre a l'atzar de l'interval $(0,1)$.
$$\Omega = \{\, x \,/\, x \in (0,1) \,\}$$

Definició. Un **succés** és qualsevol subconjunt del conjunt de resultats possibles Ω. Un subconjunt que conté un sol punt mostral s'anomena **succés elemental**.

2.2 Concepte de probabilitat

Donats uns successos, cal assignar valors numèrics a les diferents possibilitats d'ocurrència dels distints successos. Aquests valors numèrics seran la probabilitat de tals successos.

Operacions elementals amb la probabilitat. Notarem amb $P(A)$ la **probabilitat d'un succés A.** A continuació mostrem un conjunt de resultats referents a les operacions que es poden realitzar amb les probabilitats.

a) $0 \leq P(A) \leq 1$.

b) $P(\Omega) = 1$.

c) Sent $\emptyset$ el conjunt buit o succés impossible, llavors $P(\emptyset) = 0$.

d) Si A^c és el conjunt (succés) complementari d'A, llavors: $P(A) + P(A^c) = 1$.

e) Si A i B són successos qualssevol, aleshores $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

2.3 Concepte de variable aleatòria

Definició. Sigui Ω un espai mostral. Una **variable aleatòria** X és una funció definida sobre Ω de manera que a cada succés elemental de Ω li fa correspondre un nombre real.

Exemples. En cada cas, Ω serà l'espai mostral que s'obté en fer l'experiment aleatori corresponent:

- Llançar una moneda i observar el costat que apareix.

$$\begin{array}{lccc} X: & \Omega & \to & \Re \\ & C & \to & 1 \\ & X & \to & 0 \end{array}$$

- Llançar 2 monedes i observar els costats que apareixen.

$$\begin{array}{lccc} X: & \Omega & \to & \Re \\ & CC & \to & 1 \\ & CX & \to & 2 \\ & XC & \to & 3 \\ & XX & \to & 4 \end{array}$$

- Llançar un dau i observar la puntuació que apareix.

$$\begin{array}{lccc} X: & \Omega & \to & \Re \\ & 1 & \to & 1 \\ & 2 & \to & 2 \\ & 3 & \to & 3 \\ & 4 & \to & 4 \\ & 5 & \to & 5 \\ & 6 & \to & 6 \end{array}$$

- Llançar dos daus i observar la suma de les puntuacions que apareixen.

$$\begin{array}{lccc} X: & \Omega & \to & \Re \\ & (1,1) & \to & 2 \\ & & \dots & \\ & (6,6) & \to & 12 \end{array}$$

- Comptar el nombre de persones que baixen d'un autobús en una parada determinada.

$$\begin{array}{lccc} X: & \Omega & \to & \Re \\ & 0 & \to & 0 \\ & 1 & \to & 1 \\ & & \dots & \end{array}$$

- Extreure un nombre a l'atzar de l'interval (0,1).

$$\begin{array}{lccc} X: & \Omega & \to & \Re \\ & x & \to & x \end{array}$$

Observació. Normalment, si l'espai mostral és quantitatiu, la variable aleatòria assigna els valors que es corresponen amb el succés. Si l'espai mostral és qualitatiu, l'assignació que fa la variable aleatòria és més arbitrària.

Definició. Una variable aleatòria X és **discreta** si el nombre de valors que pot agafar és numerable (pot ser finit o infinit). Això vol dir que els valors es poden comptar i que sempre sabem quin valor va després d'un altre. En els cinc primers exemples anteriors ens trobem amb variables aleatòries discretes.

Definició. Una variable aleatòria X és **contínua** si els seus valors són un o més intervals de la recta real. L'últim exemple anterior és una variable aleatòria contínua.

2.4 Variables aleatòries discretes: funció de probabilitat

Cada valor d'una variable aleatòria discreta té associada una probabilitat. En el primer exemple de l'apartat anterior, $P(X = 1)$ és la probabilitat que la variable aleatòria agafi el valor 1 o, dit d'una altra manera, la probabilitat que surti cara quan llancem la moneda.

Definició. Si X és una v. a. discreta, $P(X = x)$ és una **funció de probabilitat** de la v. a. X si es compleixen les propietats següents:

- $P(X = x) \geq 0$ per a tots els valors x que pren la v. a. X.
- $\sum_x P(X = x) = 1$ (la suma de les probabilitats per a tots els valors que pren la v. a. X és 1).

Exemple. Considerem el llançament de dos daus i la v. a. X que representa la suma de les cares dels dos daus. Aquesta v. a. és discreta. Aleshores, l'assignació de la v. a. i la funció de probabilitat queden:

Succés elemental	Valor de la v. a.	Funció de probabilitat
(1,1)	2	1 / 36
(1,2), (2,1)	3	2 / 36
(1,3), (2,2), (3,1)	4	3 / 36
(1,4), (2,3), (3,2), (4,1)	5	4 / 36
(1,5), (2,4), (3,3), (4,2), (5,1)	6	5 / 36
(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)	7	6 / 36
(2,6), (3,5), (4,4), (5,3), (6,2)	8	5 / 36
(3,6), (4,5), (5,4), (6,3)	9	4 / 36
(4,6), (5,5), (6,4)	10	3 / 36
(5,6), (6,5)	11	2 / 36
(6,6)	12	1 / 36

D'aquesta manera $P(X = 2)$ és la probabilitat que la v. a. prengui el valor 2, que la suma de les cares dels daus sigui 2, i val 1/36 ja que hi ha un cas favorable d'entre els 36 possibles (si els daus són perfectes). El mateix podríem comentar sobre els altres valors de la v. a. X. Observem que la suma de les probabilitats de tots els valors és 1.

Definició. Si X és una v. a. discreta quantitativa, la **funció de distribució** de X en un punt x és la probabilitat acumulada fins a x.

2.5 Variables aleatòries contínues: funció de densitat

En el cas discret, assignàvem una probabilitat concreta a cada valor de la v. a. En el cas de v. a. contínues, el nombre de possibles valors és infinit no numerable i la probabilitat que la v. a. prengui un valor determinat és 0. Per tant, en aquest cas no podem assignar probabilitats a valors individuals de la v. a. i hem de treballar amb intervals. Això ho farem mitjançant la funció de densitat.

Definició. La **funció de densitat**, $f(x)$, d'una v. a. contínua X és la funció que compleix:

a) $f(x) \geq 0, -\infty < x < \infty$

b) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

c) $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$

Exemple. La v. a. X representa el temps (en minuts) que hi ha entre dues arribades consecutives a una botiga i la seva funció de densitat és donada per:

$$f(x) = \begin{cases} ke^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Ens demanen:

a) El valor de k perquè $f(x)$ sigui una funció de densitat.

b) La probabilitat que una persona arribi entre 2 i 6 minuts després de l'anterior.

c) La probabilitat que una persona arribi abans que passin 8 minuts des que ha arribat l'anterior.

Solució

a) $f(x)$ compleix la primera propietat de les funcions de densitat. Per complir la segona ha de passar que $\int_{-\infty}^{\infty} f(x)\,dx = 1$, és a dir, $\int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} ke^{-x/2}\,dx = 1$. La primera integral és 0, per tant s'ha de complir que $\int_{0}^{\infty} ke^{-x/2}\,dx = 1 \Rightarrow -2ke^{-x/2}\Big|_{0}^{\infty} = 0 + 2k = 2k = 1 \Rightarrow \mathrm{k} = 1/2$.

b) Aquest apartat ens demana

$$\mathrm{P}(2<\mathrm{X}<6) = \int_{2}^{6} \frac{1}{2} e^{-x/2}\,dx = -\frac{1}{2} 2e^{-x/2}\Big|_{2}^{6} = -e^{-x/2}\Big|_{2}^{6} = -e^{-3}+e^{-1} = 0.3181.$$

c) Busquem $\mathrm{P}(\mathrm{X}<8) = \int_{0}^{8} \frac{1}{2} e^{-x/2}\,dx = -e^{-x/2}\Big|_{0}^{8} = -e^{-4}+e^{0} = -e^{-4}+1 = 0.9817.$

Definició. Si X és una v. a. contínua, la **funció de distribució** de X en un punt x és la probabilitat acumulada fins a x.

2.6 Esperança matemàtica

L'*esperança matemàtica* és un concepte que es correspon amb el concepte de *mitjana* estudiat al tema d'estadística descriptiva. En aquest cas, treballem amb variables aleatòries i l'esperança matemàtica serà el valor mitjà teòric de tots els valors que pot prendre la v. a. L'esperança matemàtica no ha de coincidir necessàriament amb la mitjana d'una sèrie de dades obtingudes a partir de la v. a. que estem estudiant, encara que, si el nombre de vegades que es fa l'experiment és cada vegada més gran, la mitjana de les dades tendirà al valor de l'esperança.

Definició. L'**esperança matemàtica**, $E(X)$, d'una variable aleatòria X és el valor mitjà teòric de X i es calcula mitjançant les fórmules següents:

$$E(X) = \sum_{x} x \cdot p(x) \qquad \text{si } X \text{ és discreta}$$

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx \qquad \text{si } X \text{ és contínua}$$

on $p(x)$ i $f(x)$ són les funcions de probabilitat i de densitat, respectivament.

Notació. L'esperança matemàtica també la notarem amb la lletra grega μ.

Exemple 1. Agafem la v. a. X, que representa la suma de les cares de dos daus, i busquem la seva esperança. En aquest cas, X és discreta i hem d'aplicar la fórmula corresponent:

$$\mu = \sum_{x=2}^{12} x \cdot p(x) = 2 \cdot (1/36) + 3 \cdot (2/36) + 4 \cdot (3/36) + \ldots + 11 \cdot (2/36) + 12 \cdot (1/36) = 7$$

Exemple 2. Agafem la v. a. X, que representa el temps entre dues arribades consecutives en una botiga, i busquem la seva esperança. En aquest cas, X és contínua i hem d'aplicar la fórmula corresponent:

$$\mu = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_{0}^{\infty} x \cdot \frac{1}{2} e^{-x/2}\, dx = (\text{per parts}) = -xe^{-x/2}\Big|_{0}^{\infty} + \int_{0}^{\infty} e^{-x/2}\, dx = 0 - 2e^{-x/2}\Big|_{0}^{\infty} = 2$$

2.7 Variància

La *variància d'una v. a.* és un concepte que es correspon amb el concepte de *variància* estudiat al tema d'estadística descriptiva. En aquest cas, treballem amb variables aleatòries i la variància d'una v. a. mesura la dispersió mitjana dels valors d'una v. a. respecte de la seva esperança. Igual que passava amb l'esperança, la variància d'una v. a. no ha de coincidir necessàriament amb la variància d'una sèrie de dades obtingudes a partir de la v. a. que estem estudiant, encara que, si el nombre de vegades que es fa l'experiment és cada vegada més gran, la variància de les dades tendirà al valor de la variància de la v. a.

Definició. La **variància d'una v. a.**, $Var(X)$, d'una variable aleatòria X és l'esperança de la nova v. a. $[X - E(X)]^2$ i es calcula mitjançant les fórmules:

$$Var(X) = \sum_{x} [x - E(X)]^2 \cdot p(x) \qquad \text{si } X \text{ és discreta}$$

$$Var(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot f(x)\, dx \qquad \text{si } X \text{ és contínua}$$

on $p(x)$ i $f(x)$ són les funcions de probabilitat i de densitat, respectivament.

Definició. La **desviació típica o estàndard d'una v. a.** és l'arrel quadrada de la variància.

Notació. La variància d'una v. a. també la notarem amb la lletra grega σ^2, i la desviació típica, amb σ.

Fórmula alternativa

La variància d'una v. a. X també la podem obtenir fent: $Var(X) = E(X^2) - [E(X)]^2$. En aquest cas, les fórmules per calcular la variància serien:

$$Var(X) = \sum_{x} [x^2 \cdot p(x)] - [E(X)]^2 \qquad \text{si } X \text{ és discreta}$$

$$Var(X) = \int_{-\infty}^{\infty} [x^2 \cdot f(x)]\, dx - [E(X)]^2 \qquad \text{si } X \text{ és contínua}$$

Exemple 1. Agafem la v. a. X, que representa la suma de les cares de dos daus, i busquem la seva variància. En aquest cas, X és discreta i hem d'aplicar la fórmula corresponent:

$$\sigma^2 = \sum_{x=2}^{12} [x\text{-}E(X)]^2 \cdot p(x) = E(X^2) - [E(X)]^2 = \sum_{x=2}^{12} x^2 \cdot p(x) - 7^2 =$$

$$= 2^2 \cdot (1/36) + 3^2 \cdot (2/36) + 4^2 \cdot (3/36) + \ldots + 11^2 \cdot (2/36) + 12^2 \cdot (1/36) - 49 = 54.83\text{-}49 = 5.83.$$

Exemple 2. Agafem la v. a. X, que representa el temps entre dues arribades consecutives a una botiga, i busquem la seva variància. En aquest cas, X és contínua i hem d'aplicar la fórmula corresponent:

$$\sigma^2 = E(X^2) - [E(X)]^2 = \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx - 2^2 = \int_{0}^{\infty} x^2 \cdot \frac{1}{2} e^{-x/2}\, dx - 4 = (\text{per parts}) =$$

$$= -x^2 e^{-x/2}\Big|_0^{\infty} + 2\int_0^{\infty} x e^{-x/2}\, dx - 4 = (\text{per parts}) = 0 - 4x e^{-x/2}\Big|_0^{\infty} + 4\int_0^{\infty} e^{-x/2}\, dx - 4 =$$

$$= 0 + 0 - 8e^{-x/2}\Big|_0^{\infty} - 4 = 8 - 4 = 4$$

3. Models de distribució de probabilitats

3.1 Distribucions discretes

Donarem en aquest apartat uns models de distribucions de probabilitats que tenen com a base variables aleatòries discretes.

3.1.1 Distribució de Bernouilli

Definició. Direm que una v. a. X segueix una distribució de **Bernouilli** quan la v. a. només pot prendre 2 valors diferents (0 i 1).

Exemples. En cada cas, Ω serà l'espai mostral que s'obté en fer l'experiment aleatori corresponent:

- Llançar una moneda i observar si surt cara o no:

$$\begin{array}{rccc} X: & \Omega & \to & \Re \\ & C & \to & 1 \\ & X & \to & 0 \end{array}$$

- Llançar un dau i observar si surt múltiple de 3 o no:

$$\begin{array}{rccc} X: & \Omega & \to & \Re \\ & 3, 6 & \to & 1 \\ & 1, 2, 4, 5 & \to & 0 \end{array}$$

- Agafar un iogurt i veure si està caducat o no:

$$\begin{array}{rccc} X: & \Omega & \to & \Re \\ & \text{caducat} & \to & 1 \\ & \text{no caducat} & \to & 0 \end{array}$$

- Mirar la nota d'un alumne i veure si ha aprovat o no:

$$\begin{array}{rccc} X: & \Omega & \to & \Re \\ & \text{aprovat} & \to & 1 \\ & \text{no aprovat} & \to & 0 \end{array}$$

Si p és la probabilitat que la v. a. X prengui el valor 1, la funció de probabilitat d'una v. a. que segueix una distribució de Bernouilli és:

$$p(1) = p \qquad p(0) = 1 - p = q$$

Propietats

1) L'esperança d'una Bernouilli és p ja que:

$$E(X) = 1 \cdot \mathrm{p} + 0 \cdot \mathrm{q} = \mathrm{p}$$

2) La variància d'una Bernouilli és pq ja que:

$$Var(X) = E(X^2) - [E(X)]^2 = 1^2 \cdot p + 0^2 \cdot q - p^2 = p - p^2 = p \cdot (1 - \mathrm{p}) = pq$$

3.1.2 Distribució binomial

Definició. Una v. a. X segueix una distribució **binomial** si es repeteix n vegades (de manera independent) una experiència de tipus Bernouilli. La v. a. X és el nombre de vegades que apareix el valor 1 en una Bernouilli i, per tant, els valors que pot prendre X estaran entre 0 i n.

Exemples. En cada cas, Ω serà l'espai mostral que s'obté en fer l'experiment aleatori corresponent:

- Llançar 10 monedes i observar quantes cares surten:

$$\begin{array}{rccc} X: & \Omega & \to & \Re \\ & 0 \text{ cares} & \to & 0 \\ & 1 \text{ cara} & \to & 1 \\ & & \dots & \\ & 10 \text{ cares} & \to & 10 \end{array}$$

- Llançar 6 daus i observar quants múltiples de 3 apareixen.
- Agafar 20 iogurts i veure quants estan caducats.
- Mirar les notes d'una classe de 30 alumnes i veure quants han aprovat.

La funció de probabilitat d'una v. a. que segueix una distribució binomial amb paràmetres n i p (n és el nombre de vegades que es repeteix l'experiència Bernouilli i p és la probabilitat d'obtenir el valor 1 en una sola experiència Bernouilli), és:

$$p(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, ..., n$$

Notació. Una v. a. X que segueixi una binomial amb paràmetres n i p la notarem amb $B(n,p)$.

Propietats

1) L'esperança d'una $B(n,p)$ és np ja que és la suma de n Bernouilli independents (cadascuna amb esperança p).

$$E(X) = np$$

2) La variància d'una $B(n,p)$ és npq ja que és la suma de n Bernouilli independents (cadascuna amb variància pq).

$$Var(X) = npq$$

Exemple. Agafem la situació del tercer exemple anterior on es miren 20 iogurts i observem quants estan caducats; suposem que històricament hi ha un 15% de iogurts caducats. Aquesta v. a. segueix una binomial de paràmetres $p = 0.15$ i $n = 20$, ja que 0.15 és la probabilitat que un iogurt estigui caducat. Si ens preguntem per la probabilitat que, entre els 20 iogurts observats, n'hi hagi 4 de caducats, apliquem la funció de probabilitat d'una $B(20,0.15)$ per a $k = 4$:

$$\mathrm{p(4) = P(X=4)} = \binom{20}{4} \cdot 0.15^4 \cdot 0.85^{20-4} = 0.1821.$$

L'esperança és: $E(X) = np = 20 \cdot 0.15 = 3$.

La variància és: $Var(X) = npq = 20 \cdot 0.15 \cdot 0.85 = 2.55$.

3.1.3 Distribució de Poisson

Definició. Una v. a. X segueix una distribució de **Poisson** amb paràmetre λ si la funció de probabilitat és donada per:

$$p(k) = P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \qquad k = 0, 1, 2, ...$$

Propietats. Si X és una Poisson de paràmetre λ:

1) $E(X) = \lambda$.

2) $Var(X) = \lambda$.

3) Si X_1 i X_2 són v. a. independents de Poisson de paràmetres λ_1 i λ_2, respectivament, la variable suma $X = X_1 + X_2$ segueix una distribució de Poisson de paràmetre $\lambda = \lambda_1 + \lambda_2$.

Exemple. Suposem que el nombre de cotxes que passen per una cruïlla en un minut segueix una Poisson de paràmetre $\lambda = 10$. Si ens preguntem per la probabilitat que passin 7 cotxes en un minut per la cruïlla, apliquem la funció de probabilitat per a $k = 7$:

$$p(7) = P(X=7) = e^{-10}\frac{10^7}{7!} = 0.09$$

L'esperança és: $E(X) = \lambda = 10$. La variància és: $Var(X) = \lambda = 10$.

Si ens demanen la probabilitat que passin per la cruïlla 19 cotxes en 2 minuts, hem d'aplicar la tercera propietat (ja que el nombre de cotxes que passen en 2 minuts és la suma de dues distribucions de Poisson de paràmetre 10 per a cadascuna) i obtenim que el nombre de cotxes que passen en 2 minuts segueix una Poisson de paràmetre 20:

$$p(19) = P(X=19) = e^{-20}\frac{20^{19}}{19!} = 0.0888.$$

3.1.4 Distribució uniforme discreta

Definició. Una v. a. X segueix una distribució **uniforme discreta** si la v. a. pot prendre n valors diferents amb la mateixa probabilitat. Els valors els podem ordenar des d'1 fins a n i obtenim la funció de probabilitat:

$$p(k) = P(X = k) = \frac{1}{n} \qquad k = 1, 2, ..., n.$$

Propietats. Si X és una uniforme discreta:

1) $E(X) = \frac{n+1}{2}$

2) $Var(X) = \frac{n^2 - 1}{12}$

3.2 Distribucions contínues

Donarem en aquest apartat uns models de distribucions de probabilitats que tenen com a base variables aleatòries contínues.

3.2.1 Distribució uniforme contínua

Definició. Una v. a. X segueix una distribució **uniforme contínua** si és el resultat d'escollir un nombre a l'atzar dins d'un interval (a,b). La funció de densitat és:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & x \in (a,b) \\ 0 & x \notin (a,b) \end{cases}$$

Propietats. Si X és una uniforme contínua:

1) $E(X) = \dfrac{a+b}{2}$

2) $Var(X) = \dfrac{(b-a)^2}{12}$

Exemple. Sóc a la parada d'un autobús i sé que passa un autobús, de manera regular, cada 20 minuts. Suposant que arribo a la parada d'autobús d'improvís i sense saber l'horari de pas, el temps d'espera és una variable aleatòria que segueix una distribució uniforme entre 0 i 20 minuts. Si vull saber la probabilitat que hagi d'esperar menys de 7 minuts he de fer:

$$P(X<7) = \int_0^7 \frac{1}{20-0}\,dx = \frac{x}{20}\Big|_0^7 = \frac{7}{20} = 0.35$$

Si agafo l'autobús moltes vegades, em puc preguntar pel temps mitjà d'espera; això és equivalent a buscar l'esperança de la v. a. temps d'espera i hem de fer:

$$E(X) = \frac{a+b}{2} = \frac{0+20}{2} = 10$$

3.2.2 Distribució exponencial

Definició. Una v. a. X segueix una distribució **exponencial** si la seva funció de densitat és:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Aquesta funció depèn d'un paràmetre λ.

Propietats. Si X és una v. a. exponencial:

1) $E(X) = \frac{1}{\lambda}$

2) $Var(X) = \frac{1}{\lambda^2}$

Exemple. Suposem que el temps que es tarda, en minuts, en canviar una roda segueix una v. a. exponencial amb mitjana 5 minuts. Llavors el paràmetre de la v. a. exponencial serà $\lambda = 1/5$. Si busquem la probabilitat que una persona tardi més de 6 minuts a canviar una roda, hem de fer:

$$P(X > 6) = \int_6^\infty \frac{1}{5} e^{-\frac{x}{5}} dx = -e^{-\frac{x}{5}} \Big|_6^\infty = 0 + e^{-\frac{6}{5}} = 0.3$$

3.3 Llei normal general: N(μ,σ)

Definició. Una variable aleatòria contínua X es diu que es distribueix normalment, o que segueix una **llei normal**, si la seva funció de densitat és de la forma:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0.$$

La notarem amb $N(\mu, \sigma)$.

Propietats

1) La distribució normal depèn de dos paràmetres μ i σ^2, que són la seva esperança i variància, respectivament.

$$E(X) = \mu \qquad Var(X) = \sigma^2$$

2) La distribució normal és simètrica respecte de la seva esperança μ.

3) Si X és normal $N(\mu, \sigma)$ i $a \neq 0$ és una constant, $a \cdot X$ és $N(a \cdot \mu, |a| \cdot \sigma)$.

4) Si X_1 i X_2 són normals $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$ i independents, llavors $X = X_1 + X_2$ també és normal: $N(\mu_1+\mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$. D'aquesta propietat podem deduir el teorema següent:

Teorema de l'addició. Si una variable aleatòria, X, és suma de n variables normals independents, X_i, amb mitjanes μ_i i variàncies σ_i^2, X es distribuirà també normalment amb mitjana igual a la suma de mitjanes i variància igual a la suma de variàncies, és a dir,

$$\left.\begin{array}{r} X = X_1 + X_2 + \ldots + X_n \\ X_i \sim N(\mu_i, \sigma_i) \end{array}\right\} \quad \Rightarrow \quad X \sim N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

Exemple. Suposem que Noè ja ha embarcat quasi tots els animals a la seva arca i que només li falten una parella d'elefants i una parella de girafes. Sabem que els pesos, en quilos, d'aquests animals segueixen distribucions normals amb els paràmetres següents:

$$X_{elefant\ mascle} \sim N(5150\ ,\ 400)$$
$$X_{elefant\ femella} \sim N(2950\ ,\ 200)$$
$$X_{girafa\ mascle} \sim N(1200\ ,\ 150)$$
$$X_{girafa\ femella} \sim N(850\ ,\ 100)$$

Si l'arca només té marge per augmentar en 11000 quilos el seu pes de càrrega, quina distribució té la variable que modela el pes conjunt dels quatre animals?

Solució

La variable que volem estudiar és el pes total dels quatre animals, X. La variable X és la suma de les 4 variables $X_{elefant\ mascle}$, $X_{elefant\ femella}$, $X_{girafa\ mascle}$ i $X_{girafa\ femella}$, les quals segueixen distribucions normals. Com que:

$$5150 + 2950 + 1200 + 850 = 10150 \text{ i}$$
$$\sqrt{400^2 + 200^2 + 150^2 + 100^2} = 482.18\ ,$$

la variable X seguirà una distribució normal amb els paràmetres següents: $X \sim N(10150, 482.18)$ i cal buscar $P(X < 11000)$.

3.3.1 La normal estàndard: N(0,1)

Definició. La **normal estàndard** és la normal $N(0,1)$ de mitjana 0 i desviació estàndard 1. També rep els noms de **normal tipificada** i de **normal reduïda**. Generalment és la normal que es troba en les taules estadístiques. Al llarg del text la notarem amb Z.

Qualsevol altra variable normal X que sigui $N(\mu,\sigma)$ pot passar a normal reduïda fent el canvi:

$$Z = \frac{X - \mu}{\sigma}$$

D'aquesta manera obtenim: $P(x_1 \leq X \leq x_2) = P(z_1 \leq Z \leq z_2)$ on $z_i = \frac{x_i - \mu}{\sigma}$

Exemple: suposem que el pes, en quilos, de les persones de certa població segueix una distribució normal $X \sim N(72 , 18)$ i volem trobar la probabilitat que una persona triada a l'atzar pesi entre 65 i 75 quilos. Com caldria buscar aquesta probabilitat si hem de treballar amb la taula de la normal estàndard?

Solució

Cal estandarditzar els valors 65 i 75 mitjançant la transformació $z = \frac{x - \mu}{\sigma}$. En aquest cas, tindrem:

$$z_1 = \frac{65 - 72}{18} = -0.39 \quad \text{i} \quad z_2 = \frac{75 - 72}{18} = 0.17$$

Per tant, és el mateix buscar $P(65 \leq X \leq 75)$ que buscar $P(-0.39 \leq Z \leq 0.17)$.

3.4 Distribucions deduïdes de la normal

3.4.1 Distribució khi quadrat

Definició. Siguin Z_1, ..., Z_m v. a. $N(0,1)$ independents i considerem la variable $\chi_m^2 = Z_1^2 + ... + Z_m^2$. Es diu que χ_m^2 és la v. a. amb **distribució khi quadrat amb *m* graus de llibertat.**

Propietats

1) $E(\chi_m^2) = m \quad Var(\chi_m^2) = 2m$

2) Si χ_m^2 i χ_n^2 són independents amb m i n graus de llibertat, llavors $\chi_{m+n}^2 = \chi_m^2 + \chi_n^2$, és a dir, la suma és una altra khi quadrat amb graus de llibertat la suma dels graus de llibertat.

3.4.2 Distribució t de Student

Definició. Sigui Z una v. a. amb distribució $N(0,1)$. Sigui χ_m^2 una altra variable amb distribució khi quadrat amb m graus de llibertat. Suposem Z i χ_m^2 independents. Es defineix la **distribució t de Student amb *m* graus de llibertat** com la que segueix la v. a.

$$t = \frac{Z}{\sqrt{\chi_m^2 / m}}$$

Propietats

1) L'esperança i variància existeixen per a $m > 1$ i $m > 2$, respectivament.

$$E(t) = 0 \quad \text{si} \quad m > 1$$
$$Var(t) = m / (m - 2) \quad \text{si} \quad m > 2$$

2) La distribució límit de t, per a $m \to \infty$, és $N(0,1)$.

3.4.3 Distribució F de Fisher-Snedecor

Definició. Sigui U una v. a. amb distribució χ_m^2 i V una altra v. a. amb distribució χ_n^2. Suposem que U i V són independents. Es defineix la **distribució F amb *m* graus de llibertat al numerador i *n* graus de llibertat al denominador** com a:

$$F = \frac{U / m}{V / n}$$

També es diu que F té m i n graus de llibertat o (m,n) g. l. i es nota com a $F_{m,n}$.

Propietats

1) L'esperança i variància existeixen per a $n > 2$ i $n > 4$, respectivament.

$$E(F) = n / (n - 2) \text{ si } n > 2$$

$$Var(F) = \frac{2\,n^2\,(m + n - 2)}{m(n - 2)^2\,(n - 4)} \text{ si } n > 4$$

2) Si F té distribució $F_{m,n}$, llavors $1/F$ té la distribució $F_{n,m}$. La relació entre $F_{m,n}$ i $F_{n,m}$ és:

$$P(F_{m,n} \leq x) = 1 - P(F_{n,m} \leq 1/x) = P(F_{n,m} > 1/x)$$

3) La distribució límit de $F_{m,n}$ quan $n \to \infty$ és χ_m^2.

3.5 Convergència a la llei normal: teorema del límit central

Enunciarem un teorema que, sota certes condicions, permet realitzar la substitució d'una variable aleatòria binomial per una de normal. Aquest teorema rep el nom de *teorema de Laplace - De Moivre*.

Teorema de Laplace - De Moivre. Sigui X_n una variable aleatòria binomial de paràmetres n i p (per tant, amb mitjana np i desviació típica $\sqrt{np(1-p)}$). La distribució de la variable aleatòria X_n tendeix a una normal $N(np, \sqrt{np(1-p)})$ a mesura que fem tendir n cap a infinit mantenint p constant o $X_n \xrightarrow[n\to\infty]{} N(np, \sqrt{np(1-p)})$.

Exemple. Suposem que la probabilitat que una peça sigui defectuosa és del 0.02. Si agafem 20000 peces, quina és la probabilitat que:

a) El nombre de peces defectuoses sigui 410?

b) El nombre de peces defectuoses estigui entre 400 i 450?

Solució:

a) La variable X, nombre de peces defectuoses, segueix una distribució binomial $B(20000, 0.02)$, però, com que $n = 20000$ és prou gran, podem aproximar la distribució binomial per una distribució normal X_{normal}, en aquest cas, de paràmetres $N(400, 19.8)$. Per tant, serà pràcticament el mateix buscar $P(X = 410)$ amb la distribució binomial que buscar $P(X_{normal} = 410)$ amb la distribució normal. El problema apareix quan busquem $P(X_{normal} = 410)$, ja que la probabilitat en un punt de qualsevol distribució contínua és 0. Per solucionar aquest problema, que apareix sempre que aproximem una distribució discreta per una distribució contínua, haurem d'usar la **correcció per continuïtat,** que implica que cada valor x de la variable discreta s'associa a l'interval $(x - 0.5\ ,\ x + 0.5)$ quan l'aproximem a una distribució contínua. En el nostre cas, serà aproximadament el mateix buscar $P(X = 410)$ que buscar $P(409.5 \leq X_{normal} \leq 410.5)$.

b) Actuant de la mateixa manera que en l'apartat anterior, serà aproximadament el mateix buscar $P(400 \leq X \leq 450)$ que buscar $P(399.5 \leq X_{normal} \leq 450.5)$.

3.5.1 Teorema del límit central

Una generalització del teorema de Laplace - De Moivre és la que es coneix amb el nom de *teorema del límit central*. Aquest teorema se sol aplicar quan $n \geq 30$.

Teorema 1. Si X_1, X_2, ..., X_n, són variables aleatòries independents, d'esperances $E(X_i) = \mu_i$ i variàncies $Var(X_i) = \sigma_i^2$ finites, $i = 1, ..., n$, llavors sota certes condicions generals, s'obté que:

$$X_1 + \ldots + X_n \xrightarrow[n\to\infty]{} N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

Teorema 2. Si X_1, X_2, ..., X_n són variables aleatòries independents, que provenen de la mateixa distribució, amb esperança $E(X_i) = \mu$ i variància $Var(X_i) = \sigma^2$ finites, $i = 1, ..., n$, llavors, sota certes condicions generals, s'obté que:

$$X_1 + \ldots + X_n \xrightarrow[n\to\infty]{} N\left(n\mu, \sqrt{n}\sigma\right)$$

Exemple. Suposem que es vol anar de Tarragona a Montserrat (100 quilòmetres) en una cursa de relleus on cada atleta participant fa un quilòmetre. Cada atleta tarda a fer un quilòmetre una mitjana de 4 minuts amb una desviació estàndard de 0.5 minuts. Quina distribució segueix la variable "temps total per fer els 100 quilòmetres de recorregut"?

Solució

La variable X que modela el temps total a fer els 100 quilòmetres de recorregut és la suma de les variables que modelen el temps que triga cada atleta a fer un quilòmetre, i encara que la distribució del temps de cada atleta no segueixi una normal, com que tenim que $n = 100$, podem fer l'aproximació del teorema anterior:

$$X = X_1 + \ldots + X_{100} \sim N(400,5)$$

Teorema 3. Si X_1, X_2, ..., X_n són variables aleatòries independents, que provenen de la mateixa distribució, amb esperança $E(X_i) = \mu$ i variància $Var(X_i) = \sigma^2$ finites, $i = 1, ..., n$, llavors, sota certes condicions generals, s'obté que:

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n} \xrightarrow[n\to\infty]{} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Exemple. Suposem que el temps mitjà diari que triga un estudiant a anar de casa seva a la facultat és de 23 minuts amb una desviació estàndard de 4 minuts. Si observem aquest

estudiant durant 60 dies, quina distribució segueix la variable "temps mitjà durant 60 dies a fer el trajecte de casa seva a la facultat"?

Solució

La variable $\overline{X_{60}}$ que modela el temps mitjà durant 60 dies a fer el trajecte de casa de l'estudiant a la facultat és la mitjana de les variables que modelen el temps que triga cada dia a fer aquest trajecte, i encara que la distribució del temps diari no segueixi una normal, com que tenim que $n = 60$, podem fer l'aproximació del teorema anterior:

$$\overline{X_{60}} = \frac{X_1 + \ldots + X_{60}}{60} \sim N(23,0.52)$$

Observacions

1) El teorema de Laplace - De Moivre és un cas particular del teorema central del límit, perquè una variable aleatòria binomial és suma de variables aleatòries de Bernouilli independents.

2) El teorema de l'addició es refereix a la distribució d'una variable suma de diverses variables normals independents que és exactament normal, mentre que el teorema central del límit es refereix a la d'una variable suma de diverses variables independents, però no necessàriament normals, que s'aproximarà a la normal d'una manera millor a mesura que augmentem el nombre de sumands.

3.6 Ús de les taules estadístiques

3.6.1 La taula normal estàndard

Aquesta taula la farem servir quan treballem amb v. a. normals. Si la normal amb la qual estem treballant no és estàndard, podem transformar-la perquè ho sigui.

La columna de l'esquerra correspon als punts z, amb un decimal, a partir dels quals volem trobar la probabilitat, i la fila de dalt correspon al segon decimal dels punts z. Els valors que hi ha al mig de la taula són les probabilitats associades a l'interval $[z,\infty)$.

Exemples

a) Suposem que $Z \sim N(0,1)$ i busquem $P(Z \geq 1.67)$. Amb la taula podem trobar aquesta probabilitat directament. A la columna de l'esquerra mirem la fila corresponent a 1.6 i dintre d'aquesta fila mirem quina àrea està associada quan a dalt hi ha el valor 7 (el qual és el segon decimal). Veiem que l'àrea associada és 0.0475. Per tant,

$$P(Z \geq 1.67) = 0.0475.$$

b) Suposem que $Z \sim N(0,1)$ i busquem $P(Z \leq -1.67)$. Com que la normal estàndard és simètrica respecte al 0, la probabilitat que hi ha entre $-\infty$ i -1.67 és la mateixa que hi ha entre 1.67 i ∞. Per tant, puc calcular aquesta última com hem fet a l'exemple *a*) i obtenim:

$$P(Z \leq -1.67) = P(1.67 \leq Z) = 0.0475.$$

c) Suposem que $Z \sim N(0,1)$ i busquem $P(0.54 \leq Z \leq 1.67)$. Amb la taula podem trobar les probabilitats que hi ha entre 0.54 i ∞ i entre 1.67 i ∞. A nosaltres ens interessa la probabilitat que hi ha entre 0.54 i 1.67 i aquesta probabilitat la podem aconseguir restant de l'àrea que hi ha entre 0.54 i ∞ l'àrea que hi ha entre 1.67 i ∞. Per tant:

$$P(0.54 \leq Z \leq 1.67) = P(Z \geq 0.54) - P(Z \geq 1.67) = 0.2946 - 0.0475 = 0.2471$$

d) Suposem que $Z \sim N(0,1)$ i busquem $P(-1.67 \leq Z \leq -0.54)$. Com que la normal estàndard és simètrica respecte al 0, la probabilitat que hi ha entre -1.67 i -0.54 és la mateixa que hi ha entre 0.54 i 1.67. Per tant, puc calcular aquesta última com hem fet a l'exemple *c*) i obtenim:

$$P(-1.67 \leq Z \leq -0.54) = P(0.54 \leq Z \leq 1.67) = 0.2471$$

e) Suposem que $Z \sim N(0,1)$ i busquem $P(-0.54 \leq Z \leq 1.67)$. Amb la taula podem trobar les probabilitats que hi ha entre $-\infty$ i -0.54 i entre 1.67 i ∞. A nosaltres ens interessa la probabilitat que hi ha entre -0.54 i 1.67 i aquesta probabilitat la podem aconseguir restant d'1 l'àrea que hi ha entre 1,67 i ∞ i l'àrea que hi ha entre $-\infty$ i -0.54. Per tant:

$$P(-0.54 \leq Z \leq 1.67) = 1 - (P(Z \leq -0.54) + P(Z \geq 1.67)) =$$
$$1 - (0.2946 + 0.0475) = 0.6579$$

f) Suposem que $Z \sim N(0,1)$ i busquem $P(0 \leq Z \leq 1.67)$. A la taula trobem la probabilitat que hi ha entre 1.67 i ∞ i sabem que la probabilitat que Z sigui més gran que 0 és 0.5. A nosaltres ens interessa la probabilitat que Z estigui entre 0 i 1.67 i aquesta probabilitat la podem aconseguir restant de 0.5 l'àrea que hi ha entre 1.67 i ∞. Per tant:

$$P(0 \leq Z \leq 1.67) = 0.5 - 0.0475 = 0.4525.$$

g) Suposem que $X \sim N(50,12)$ i busquem $P(44 \leq X \leq 58)$. La v. a. no és una normal estàndard i, per tant, hem de transformar les dades. Sabem que $Z = \frac{X - \mu}{\sigma}$ és una normal estàndard i hem de fer:

$$P(44 \leq X \leq 58) = P\left(\frac{44 - 50}{12} \leq \frac{X - \mu}{\sigma} \leq \frac{58 - 50}{12}\right) = P(-0.5 \leq Z \leq 0.67).$$

Una vegada hem passat les dades a una normal estàndard, ja podem treballar amb la taula:

$$\begin{aligned} P(-0.5 \leq Z \leq 0.67) &= 1 - (P(Z \leq -0.5) + P(Z \geq 0.67)) \\ &= 1 - (0.3085 + 0.2514) = 0.4401 \end{aligned}$$

h) Suposem que $Z \sim N(0,1)$ i busquem un punt z de manera que $P(Z \geq z) = 0.0918$. Ara ens donen la probabilitat i hem de trobar el punt. Per la dada que ens donen sabem que el punt z serà un nombre positiu. Hem de buscar a l'interior de la taula l'àrea 0.0918 i veure amb quin punt es correspon. El punt és $z = 1.33$.

i) Suposem que $Z \sim N(0,1)$ i busquem un punt z de manera que $P(Z \geq z) = 0.8238$. Ara sabem que el punt z serà un nombre negatiu. Com que la taula només treballa amb nombres positius, haurem d'usar la simetria de la normal. El punt z que compleix $P(Z \geq z) = 0.8238$ també compleix que $P(Z \leq z) = 0.1762$. Hem de buscar un punt positiu z' de manera que $P(Z \geq z') = 0.1762$ i després canviar-li el signe, ja que es complirà que $z = -z'$. Buscant a l'interior de la taula l'àrea 0.1762, obtenim el punt $z' = 0.93$ i, per tant, el punt z que es busca és $z = -0.93$.

j) Suposem que $Z \sim N(0,1)$ i busquem un punt z de manera que $P(Z \leq z) = 0.9207$. Ara sabem que el punt z serà un nombre positiu, però ens estan donant l'àrea acumulada fins al punt (l'àrea de l'esquerra) i la taula treballa amb l'àrea de la dreta. El punt z que compleix $P(Z \leq z) = 0.9207$ també compleix que $P(Z \geq z) = 0.0793$. Hem de buscar a l'interior de la taula l'àrea 0.0793 i veure amb quin punt es correspon. El punt és $z = 1.41$.

k) Suposem que $Z \sim N(0,1)$ i busquem un punt z de manera que $P(Z \leq z) = 0.3228$. Ara sabem que el punt z serà un nombre negatiu i haurem d'usar la simetria de la normal per trobar-lo. Hem de buscar un punt positiu z' de manera que $P(Z \geq z') = 0.3228$ i després canviar-li el signe, ja que es complirà que $z = -z'$. Buscant a l'interior de la taula l'àrea 0.3228 obtenim el punt $z' = 0.46$ i, per tant, el punt z que es busca és $z = -0.46$.

3.6.2 La taula khi quadrat

Aquesta taula la farem servir quan treballem amb una v. a. Khi quadrat. Aquesta taula té unes característiques diferents de la taula anterior.

La columna de l'esquerra correspon als graus de llibertat de la χ^2 amb la qual treballem. La fila de dalt correspon a les probabilitats que hi ha entre el punt amb el qual estem treballant i ∞. Els valors que hi ha al mig de la taula són els punts associats a les probabilitats de la fila superior.

Exemples

a) Suposem que $X \sim \chi^2_{14}$ i busquem $P(X \geq 7.8)$. Hem d'agafar la fila corresponent als 14 graus de llibertat. Si busquem el punt 7.8 en aquesta fila, veiem que té una probabilitat associada de 0.9:

$$P(X \geq 7.8) = 0.9$$

b) Suposem que $X \sim \chi^2_{14}$ i busquem $P(X \leq 7.8)$. Aquest succés és el complementari de l'anterior i, per tant, hem de restar d'1 la probabilitat anterior:

$$P(X \leq 7.8) = 1 - P(\mathrm{X} \geq 7.8) = 1 - 0.9 = 0.1$$

c) Suposem que $X \sim \chi^2_{14}$ i busquem un punt x, de manera que $P(X \geq x) = 0.1$. Ara ens donen la probabilitat i hem de trobar el punt i, per tant, hem de buscar a la fila superior l'àrea 0.1 i veure amb quin punt es correspon. El punt és $x = 21.1$.

d) Suposem que $X \sim \chi^2_{14}$ i busquem un punt x de manera que $P(X \leq x) = 0.75$. Ara ens donen la probabilitat que la v. a. sigui més petita que un punt. La taula ens dóna les probabilitats dels successos contraris i hem de trobar el punt de manera que $P(X \geq x) = 1 - P(X \leq x) = 1 - 0.75 = 0.25$ i el punt és $x = 17.1$.

3.6.3 La taula t de Student

Aquesta taula la farem servir quan treballem amb una v. a. t de Student. Aquesta taula té unes característiques semblants a la taula anterior, però cal tenir en compte que la distribució t de Student és simètrica respecte al zero.

La columna de l'esquerra correspon als graus de llibertat de la t de Student amb la qual treballem. La fila de dalt correspon a les probabilitats que la v. a. sigui més gran que el punt amb el qual estem treballant. Els valors que hi ha al mig de la taula són els punts associats a les probabilitats de la fila superior.

Exemples

a) Suposem que $X \sim$ t de Student amb 14 graus de llibertat i busquem $P(X \geq 2.14)$. Hem d'agafar la fila corresponent als 14 graus de llibertat. Si busquem el punt 2.14 en aquesta fila veiem que té una probabilitat associada de 0.025:

$$P(X \geq 2.14) = 0.025$$

b) Suposem que $X \sim$ t de Student amb 14 graus de llibertat i busquem $P(X \leq 2.14)$. Aquest succés és el complementari de l'anterior i, per tant, hem de restar d'1 la probabilitat anterior:

$$P(X \leq 2.14) = 1 - P(X \geq 2.14) = 1 - 0.025 = 0.975$$

c) Suposem que $X \sim$ t de Student amb 14 graus de llibertat i busquem un punt x de manera que $P(X \geq x) = 0.1$. Ara ens donen la probabilitat i hem de trobar el punt i, per tant, hem de buscar a la fila superior l'àrea 0.1 i veure amb quin punt es correspon. El punt és $x = 1.35$.

d) Suposem que $X \sim$ t de Student amb 14 graus de llibertat i busquem un punt x de manera que $P(X \leq x) = 0.75$. Ara ens donen la probabilitat que la v. a. sigui més petita que un punt. La taula ens dóna les probabilitats dels successos contraris i hem de trobar el punt de manera que:

$$P(X \geq x) = 1 - P(X \leq x) = 1 - 0.75 = 0.25 \text{ i el punt és } x = 0.69$$

e) Suposem que $X \sim$ t de Student amb 14 graus de llibertat i busquem un punt x de manera que $P(X \leq x) = 0.25$. Aquest punt serà a l'esquerra del 0 (serà negatiu), i com que la v. a. és simètrica, el punt buscat és l'oposat del punt que $P(X \geq x) = 0.25$. El punt buscat és –0.69.

3.6.4 La taula F de Fisher

Aquesta taula la farem servir quan treballem amb una v. a. F de Fisher. Aquesta taula té unes característiques semblants a les dues taules anteriors.

La fila de dalt correspon als graus de llibertat del numerador de la F de Fisher amb la qual treballem i la columna de l'esquerra correspon als graus de llibertat del denominador. Els valors que hi ha al mig de la taula són els punts associats a les probabilitats. En aquest cas, tenim una taula per a cada probabilitat diferent que es vulgui treballar. La primera taula correspon a la probabilitat 0.05 a la dreta del punt, la segona a 0.025 i la tercera a 0.01.

Exemples

a) Suposem que $X \sim F_{10,15}$ i busquem un punt x de manera que $P(X \leq x) = 0.95$. Hem d'agafar la taula corresponent a la probabilitat 0.05 i buscar el punt que es troba a la columna dels 10 graus de llibertat i la fila dels 15 graus de llibertat. El punt és $x = 2.54$.

b) Suposem que $X \sim F_{10,15}$ i busquem un punt x de manera que $P(X \leq x) = 0.99$. Hem d'agafar la taula corresponent a la probabilitat 0.01 i buscar el punt que es troba a la columna dels 10 graus de llibertat i la fila dels 15 graus de llibertat. El punt és $x = 3.8$.

c) Suposem que $X \sim F_{10,15}$ i busquem un punt x de manera que $P(X \geq x) = 0.99$. Hem d'agafar la taula corresponent a la probabilitat 0.01 i apliquem la propietat 2) de la distribució de Fisher, és a dir, s'ha d'intercanviar l'ordre dels graus de llibertat, buscar

el punt x' que deixa a la seva dreta una probabilitat de 0,01 (o que deixa a la seva esquerra una probabilitat de 0.99) i assignar $x = 1/x'$:

$$0.99 = P(F_{10,15} \geq x) = P(F_{15,10} < x')$$

El punt x' trobat és 4.56. Per tant, $x = 1 / 4.56 = 0.219$.

3.7 Funcions d'Excel per calcular probabilitats

3.7.1 Distribució binomial

Funció Excel: DISTR.BINOM.

Objectiu: buscar probabilitats corresponents als possibles valors d'una distribució binomial (n,p).

Paràmetres:

- Nombre_èxits: valor del qual volem calcular la probabilitat (ja sigui la probabilitat exacta o la probabilitat acumulada).
- Intents: paràmetre n de la binomial.
- Prob._èxit: probabilitat p d'aconseguir el que s'està estudiant quan agafem un sol element. És el paràmetre p de la binomial.
- Acumulat: posar 0 si només volem la probabilitat que el resultat sigui exactament "Nombre_èxits" i posar 1 si volem la probabilitat acumulada fins a "Nombre_èxits".

Enunciat exemple 1

Tenim 20 iogurts a la nevera. La probabilitat que un iogurt estigui caducat és del 15%. Quina és la probabilitat que, d'entre els iogurts que hi ha a la nevera, n'hi hagi 6 de caducats?

Solució exemple 1

Usant la funció d'Excel DISTR.BINOM, hem de posar:

- Nombre_èxits: 6.
- Intents: 20.
- Prob_èxit: 0.15.
- Acumulat: 0.

La probabilitat demanada és 0.0454 (un 4.54%).

Enunciat exemple 2

Tenim 20 iogurts a la nevera. La probabilitat que un iogurt estigui caducat és del 15%. Quina és la probabilitat que, d'entre els iogurts que hi ha a la nevera, n'hi hagi 6 o menys de caducats?

Solució exemple 2

Usant la funció d'Excel DISTR.BINOM, hem de posar:

- Nombre_èxits: 6.
- Intents: 20.
- Prob_èxit: 0.15.
- Acumulat: 1.

La probabilitat demanada és 0.9781 (un 97.81%).

3.7.2 Distribució de Poisson

Funció Excel: POISSON.

Objectiu: buscar probabilitats corresponents als possibles valors d'una distribució de Poiss(λ).

Paràmetres:

- X: valor del qual volem calcular la probabilitat (ja sigui la probabilitat exacta o la probabilitat acumulada).
- Mitjana: mitjana o paràmetre λ de la Poisson.
- Acumulat: posar 0 si només volem la probabilitat que el resultat sigui exactament *X* i posar 1 si volem la probabilitat acumulada fins a *X*.

Enunciat exemple 1

El nombre de clients que arriben a certa entitat bancària segueix una distribució de Poisson amb una mitjana de 15 clients per hora. Quina és la probabilitat que durant la propera hora arribin 12 clients a l'entitat bancària?

Solució exemple 1

Usant la funció d'Excel POISS, hem de posar:

- X: 12.
- Mitjana: 15.
- Acumulat: 0.

La probabilitat demanada és 0.0829 (un 8.29%).

Enunciat exemple 2

El nombre de clients que arriben a certa entitat bancària segueix una distribució de Poisson amb una mitjana de 15 clients per hora. Quina és la probabilitat que durant la propera hora arribin 12 clients o menys a l'entitat bancària?

Solució exemple 2

Usant la funció d'Excel POISS, hem de posar:

- X: 12.
- Mitjana: 15.
- Acumulat: 1.

La probabilitat demanada és 0.2676 (un 26.76%).

3.7.3 Distribució exponencial

Funció Excel: DISTR.EXP.

Objectiu: buscar probabilitats corresponents als possibles valors d'una distribució exponencial Exp(λ).

Paràmetres:

- X: valor del qual volem calcular la probabilitat acumulada.
- Lambda: paràmetre λ de l'exponencial.
- Acumulat: s'ha de posar 1 per calcular la probabilitat acumulada fins a *X*.

Enunciat exemple

El temps que es triga a canviar una roda segueix una distribució exponencial amb una mitjana de 5 minuts. Quina és la probabilitat que es tardin menys de 6 minuts per canviar la propera roda?

Solució exemple

Usant la funció d'Excel DISTR.EXP, hem de posar:

- X: 6.
- Lambda: 0.2 (és el mateix que 1/5).
- Acumulat: 1.

La probabilitat demanada és 0.6988 (un 69.88%).

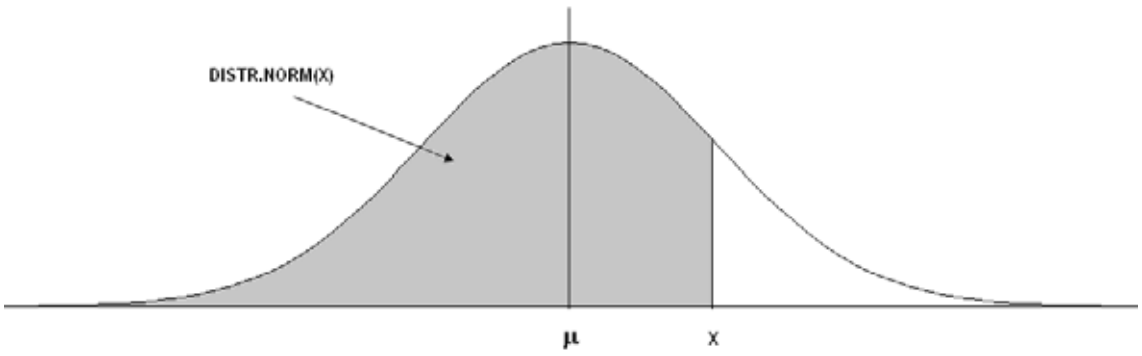
3.7.4 Distribució normal

Funció Excel: DISTR.NORM.

Objectiu: buscar probabilitats corresponents als possibles valors d'una distribució Normal(μ,σ).

Paràmetres:

- X: valor del qual volem calcular la probabilitat acumulada.
- Mitjana: mitjana de la distribució normal.
- Desv_estàndard: desviació estàndard de la distribució normal.
- Acumulat: hem de posar 1 per calcular la probabilitat acumulada fins a *X*.



Enunciat exemple

Suposem que una variable aleatòria *X* segueix una distribució *N*(40,8). Quina és la probabilitat $P(X \leq 34)$?

Solució exemple

Usant la funció d'Excel DISTR.NORM, hem de posar:

- X: 34.
- Mitjana: 40.
- Desv_estàndard: 8.
- Acumulat: 1.

La probabilitat demanada és 0.2266 (un 22.66%).

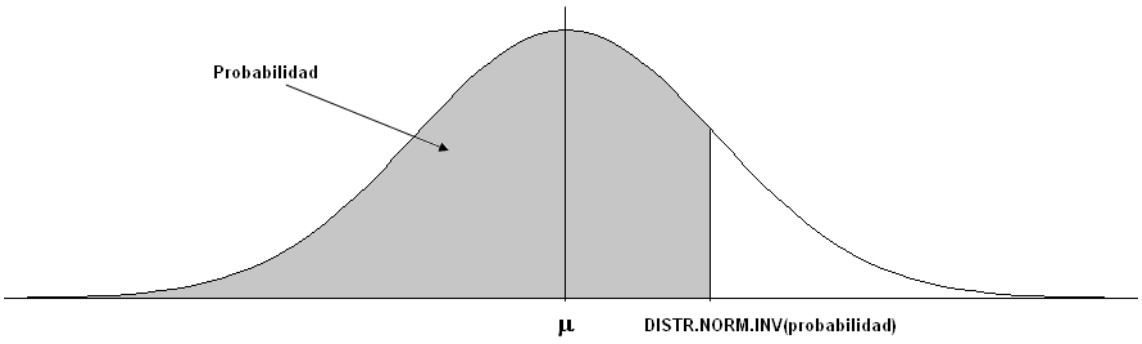
Funció Excel: DISTR.NORM.INV.

Objectiu: buscar valors d'una distribució Normal(μ,σ) corresponents a unes probabilitats donades.

Paràmetres:

- Probabilitat: probabilitat acumulada fins al punt que serà la resposta d'aquesta funció.

- Mitjana: mitjana de la distribució normal.
- Desv_estàndard: desviació estàndard de la distribució normal.



Enunciat exemple

Suposem que una variable aleatòria X segueix una distribució $N(40,8)$. Quin és el valor a que fa que $P(X \leq a) = 0.33$?

Solució exemple

Usant la funció d'Excel DISTR.NORM.INV, hem de posar:

- Probabilitat: 0.33.
- Mitjana: 40.
- Desv_estàndard: 8.

El valor a demanat és 36.48.

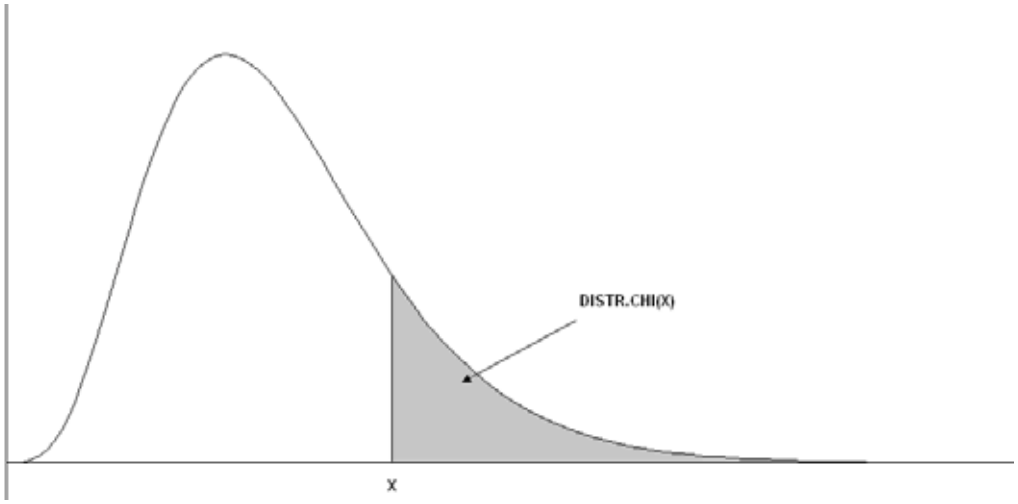
3.7.5 Distribució khi quadrat

Funció Excel: DISTR.CHI.

Objectiu: buscar probabilitats corresponents als possibles valors d'una distribució khi quadrat amb n graus de llibertat.

Paràmetres:

- X: valor del qual volem calcular la probabilitat corresponent a la cua de la dreta.
- Graus_de_llibertat: graus de llibertat de la khi quadrat.



Enunciat exemple

Suposem que una variable aleatòria X segueix una distribució khi quadrat amb 12 graus de llibertat. Quina és la probabilitat $P(X \geq 6.3)$?

Solució exemple

Usant la funció d'Excel DISTR.CHI, hem de posar:

- X: 6.3.
- Graus_de_llibertat: 12.

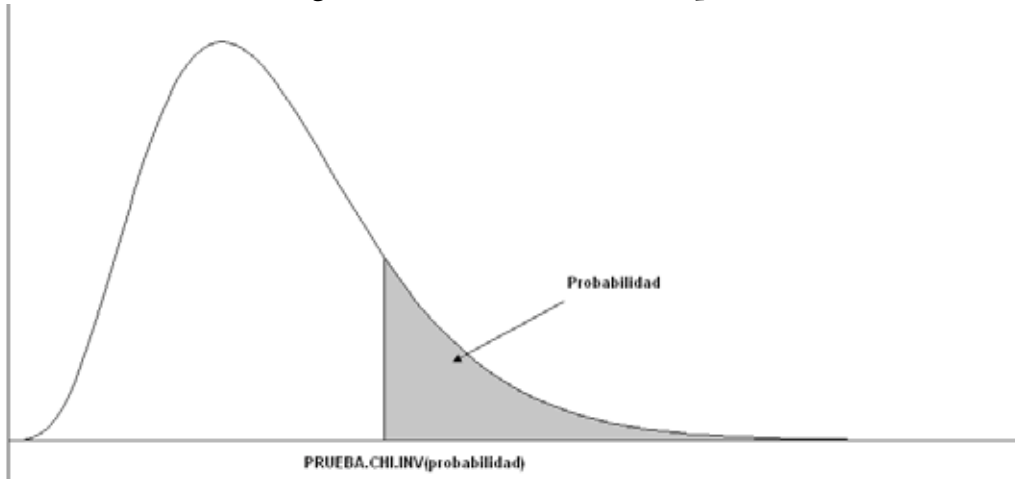
La probabilitat demanada és 0.9002 (un 90.02%).

Funció Excel: PRUEBA.CHI.INV.

Objectiu: buscar valors d'una distribució khi quadrat amb *n* graus de llibertat corresponents a unes probabilitats donades.

Paràmetres:

- Probabilitat: probabilitat corresponent a la cua de la dreta del punt que serà la resposta d'aquesta funció.
- Graus_de_llibertat: graus de llibertat de la khi quadrat.



Enunciat exemple

Suposem que una variable aleatòria X segueix una distribució khi quadrat amb 12 graus de llibertat. Quin és el valor a que fa que $P(X \geq a) = 0.75$?

Solució exemple

Usant la funció d'Excel PRUEBA.CHI.INV, hem de posar:

- Probabilitat: 0.75.
- Graus_de_llibertat: 12.

El valor a demanat és 8.4.

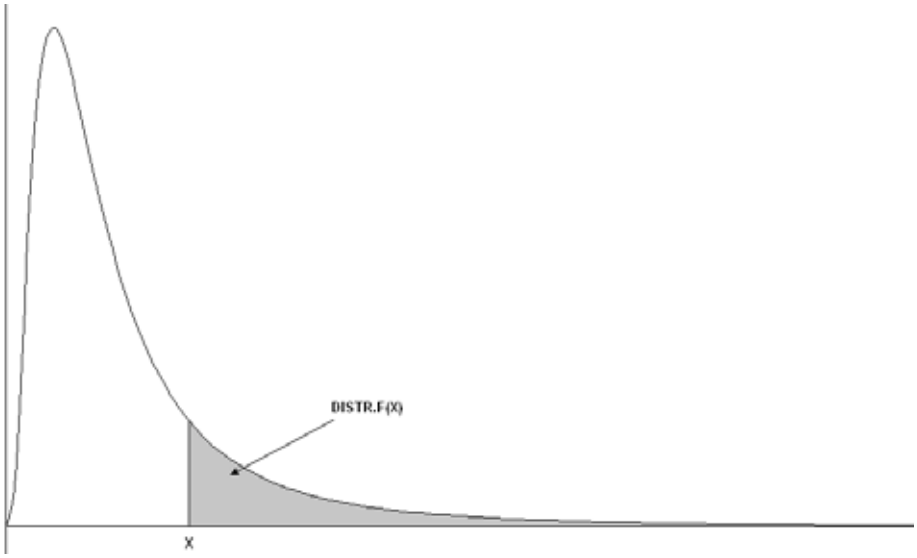
3.7.6 Distribució F de Fisher

Funció Excel: DISTR.F.

Objectiu: buscar probabilitats corresponents als possibles valors d'una distribució F de Fisher amb m i n graus de llibertat.

Paràmetres:

- X: valor del qual volem calcular la probabilitat corresponent a la cua de la dreta.
- Graus_de_llibertat: graus de llibertat del numerador de la F.
- Graus_de_llibertat2: graus de llibertat del denominador de la F.



Enunciat exemple

Suposem que una variable aleatòria X segueix una distribució F de Fisher amb 12 i 18 graus de llibertat. Quina és la probabilitat $P(X \geq 1.4)$?

Solució exemple

Usant la funció d'Excel DISTR.F, hem de posar:

- X: 1.4.
- Graus_de_llibertat: 12.
- Graus_de_llibertat2: 18.

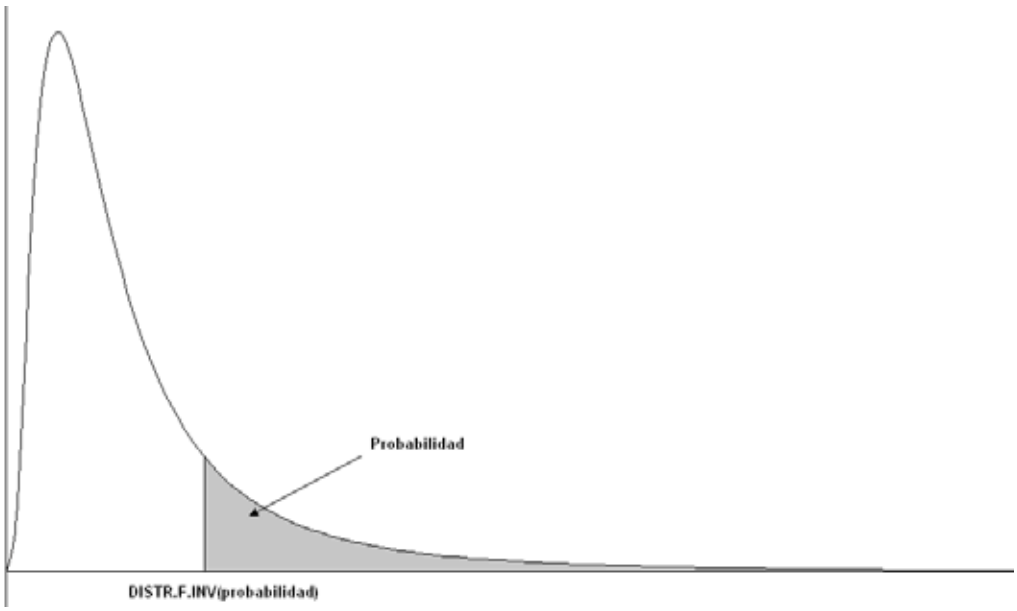
La probabilitat demanada és 0.2518 (un 25.18%).

Funció Excel: DISTR.F.INV.

Objectiu: buscar valors d'una distribució F de Fisher amb *m* i *n* graus de llibertat corresponents a unes probabilitats donades.

Paràmetres:

- Probabilitat: probabilitat corresponent a la cua de la dreta del punt que serà la resposta d'aquesta funció.
- Graus_de_llibertat1: graus de llibertat del numerador de la F.
- Graus_de_llibertat2: graus de llibertat del denominador de la F.



Enunciat exemple

Suposem que una variable aleatòria X segueix una distribució F de Fisher amb 12 i 18 graus de llibertat. Quin és el valor a que fa que $P(X \geq a) = 0.85$?

Solució exemple

Usant la funció d'Excel DISTR.F.INV, hem de posar:

- Probabilitat: 0.85.
- Graus_de_llibertat1: 12.

- Graus_de_llibertat2: 18.

El valor *a* demanat és 0.5546.

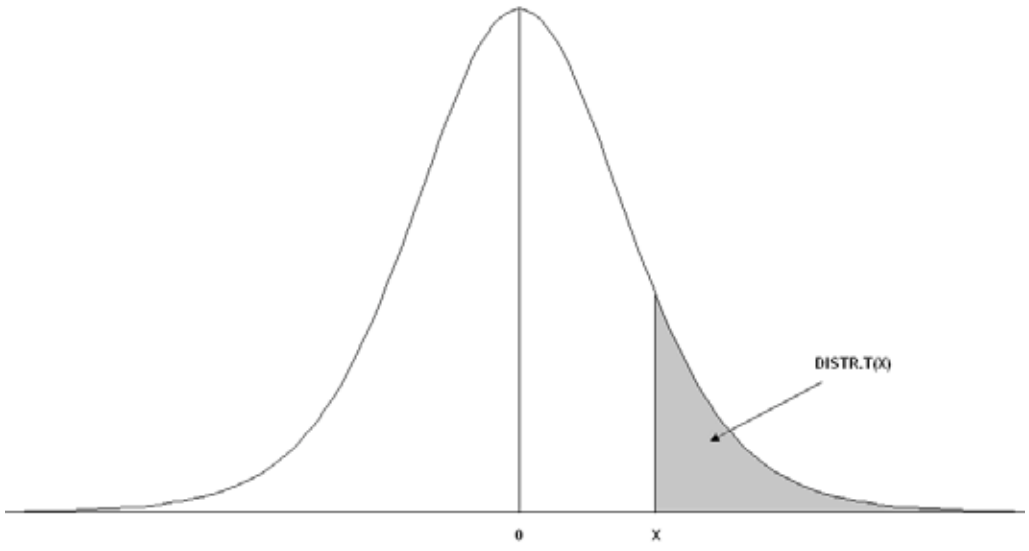
3.7.7 Distribució t de Student

Funció Excel: DISTR.T.

Objectiu: buscar probabilitats corresponents als possibles valors d'una distribució t de Student amb *n* graus de llibertat.

Paràmetres:

- X: valor del qual volem calcular la probabilitat corresponent a la cua de la dreta. NOMÉS s'admeten valors positius.
- Graus_de_llibertat: graus de llibertat de la t de Student.
- Cues: hem de posar 1 per calcular la probabilitat corresponent a la cua de la dreta.



Enunciat exemple

Suposem que una variable aleatòria X segueix una distribució t de Student amb 12 graus de llibertat. Quina és la probabilitat $P(X \geq 0.94)$?

Solució exemple

Usant la funció d'Excel DISTR.T, hem de posar:

- X: 0.94.
- Graus_de_llibertat: 12.
- Cues: 1.

La probabilitat demanada és 0.1829 (un 18.29%).

Funció Excel: DISTR.T.INV.

Objectiu: buscar valors d'una distribució t de Student amb *n* graus de llibertat corresponents a unes probabilitats donades. Aquesta funció NOMÉS busca valors positius.

Paràmetres:

- Probabilitat: probabilitat corresponent al DOBLE de la cua de la dreta del punt que serà la resposta d'aquesta funció. Per exemple, si volem trobar el punt que deixa a la cua de la seva dreta una probabilitat de 0.4, a "Probabilitat" hem de posar 0.8.
- Graus_de_llibertat: graus de llibertat de la t de Student.

Enunciat exemple

Suposem que una variable aleatòria X segueix una distribució t de Student amb 12 graus de llibertat. Quin és el valor a que fa que $P(X \geq a) = 0.35$?

Solució exemple

Usant la funció d'Excel DISTR.T.INV, hem de posar:

- Probabilitat: 0.7.
- Graus_de_llibertat: 12.

El valor a demanat és 0.3947.

4. Intervals de confiança

4.1 Nocions de mostra i mostreig

Definició. L'estadística té com a objectiu l'estudi de les **poblacions,** entenent per aquest terme un conjunt de persones, coses o, en general, elements amb alguna característica comuna a tots ells.

De l'observació del comportament individual de cada un dels elements que componen la població es poden obtenir unes lleis generals per a tots els elements de la població.

Sembla evident que per trobar aquestes lleis generals sigui necessària l'observació exhaustiva de tota la població. Inconvenients d'organització, de temps i, en definitiva, econòmics fan molt difícil estudiar tots els elements de la població si aquesta és molt gran. Per exemple, si volem fer un estudi de la vida de les bombetes que produeix una fàbrica, hem d'observar quant de temps passa fins que la bombeta es fon, i això no ho farem amb totes les bombetes que es fabriquen (si ho féssim, la fàbrica es quedaria sense bombetes per vendre).

Definició. En els casos que no puguem observar tots els elements de la població, seleccionarem un conjunt d'elements de la població, que anomenem **mostra.**

Perquè sigui correcta la substitució de l'observació exhaustiva de la població per la més limitada observació dels elements que formen una mostra, cal que la composició d'aquesta sigui representativa de la composició de la població.

Definició. S'anomena **mostreig** la tècnica emprada per a l'obtenció de mostres.

Definició. Direm que una mostra és **aleatòria simple** quan:

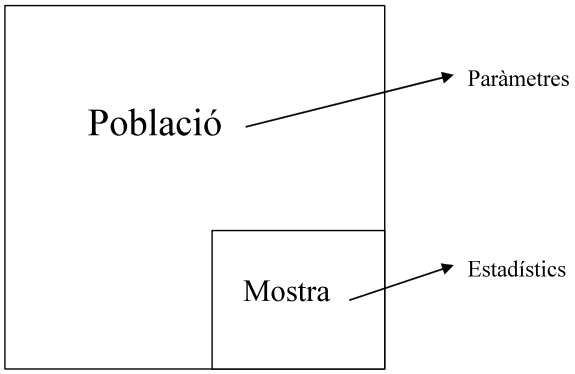
1. Cada element de la població té la mateixa probabilitat de ser escollit.
2. Les observacions es realitzen amb reemplaçament, de manera que la població és idèntica en totes les extraccions.

4.2 Concepte d'estadístic i de paràmetre

Definició. Donada una mostra $(X_1, ..., X_n)$, s'anomena **estadístic** tota variable que sigui funció de la mostra:

$$U_n = f(X_1, ..., X_n)$$

Definició. S'anomena **paràmetre** qualsevol valor obtingut d'una població.



Els diversos paràmetres poblacionals generalment són desconeguts, ja que no tenim disponibles les dades de tota la població. El valor dels estadístics sempre serà conegut, ja que sempre tindrem disponibles les dades d'alguna mostra. Els paràmetres poblacionals tenen relacionats certs estadístics mostrals i els valors d'aquests estadístics ens permetran fer una inferència o estimació sobre quin és el valor del paràmetre poblacional corresponent.

POBLACIÓ → PARÀMETRES (desconeguts)	MOSTRA → ESTADÍSTICS (coneguts)
Mitjana poblacional, μ	Mitjana mostral, $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Variància poblacional, σ^2	Variància mostral, $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$
Desviació estàndard poblacional, σ	Desviació estàndard mostral, $S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$
Proporció poblacional, p	Proporció mostral, $\hat{p}$
Mediana poblacional, Me	Mediana mostral, $\hat{Me}$
Percentil 35 poblacional, P_{35}	Percentil 35 mostral, $\hat{P}_{35}$
...	...

4.3 Estimació puntual i estimació per intervals

El procés d'estimació estadística es relaciona amb el conjunt de mètodes i procediments estadístics a partir dels quals, i de la informació donada per una mostra aleatòria obtinguda d'una determinada població, es pretén fer alguna afirmació en termes numèrics sobre el valor del paràmetre o paràmetres desconeguts que caracteritzen la població. Es tracta, en definitiva, de com usar de manera òptima la informació que ens dóna una mostra per poder assignar valors numèrics als paràmetres desconeguts d'una determinada població estadística.

La funció de probabilitat o de densitat d'una v. a. X depèn d'alguns paràmetres, com ara l'esperança i la variància. Els vertaders valors d'aquests paràmetres sovint són desconeguts a la pràctica i han de ser estimats a partir d'una mostra de X.

Definició. Sigui $f(x,\theta)$ la funció de densitat de X on θ és el paràmetre desconegut. S'anomena **estimador** de θ una variable aleatòria funció de la mostra

$$U = g(X_1, ..., X_n)$$

Donada una mostra aleatòria, sempre és possible obtenir determinats estadístics mostrals: mitjana, mediana, desviació estàndard..., i considerar-los com a estimadors potencials dels paràmetres poblacionals.

Cal remarcar que l'obtenció d'estimadors pot fer-se mitjançant l'aplicació de dos criteris diferents:

1. **Estimació puntual**: s'anomena *estimació puntual* el procés d'estimació que assigna a cada paràmetre cert valor estimat. L'estimació puntual té l'avantatge de deixar especificat unívocament el paràmetre que es pretén estimar, però té l'inconvenient que, si agafem dues mostres diferents, es podrien produir diferències importants entre els valors estimats.
2. **Estimació per intervals**: consisteix a marcar un interval al qual, amb una certa probabilitat, pertanyi l'estimador escollit, de manera que l'estimació del paràmetre poblacional serà donada per un conjunt de valors, qualsevol dels quals podria ser agafat com a expressió del paràmetre poblacional que es vol estimar.

4.4 Noció d'interval de confiança. Coeficient de confiança

Els mètodes d'estimació puntual presenten un gran inconvenient: no proporcionen informació sobre com és de gran l'error comès en l'estimació. L'error només seria conegut amb precisió en el cas que el paràmetre fos conegut, però en aquest cas no seria necessari fer cap estimació.

L'estimació per interval sorgeix per solucionar aquest problema. Aquest procediment es basa en el fet que un estimador és una v. a. caracteritzada per:

- una distribució de probabilitat
- una esperança matemàtica
- una variància

Definició. Ara posarem les bases per donar la definició d'*interval de confiança*. Sigui $X_1, ..., X_n$ una mostra aleatòria simple d'una v. a. X amb una distribució que depèn d'un paràmetre θ (i possiblement d'altres paràmetres). Es diu que els estadístics:

$$U = g_1(X_1, ..., X_n) \qquad V = g_2(X_1, ..., X_n)$$

constitueixen un **interval de confiança** per a θ, amb coeficient (o nivell) de confiança $1 - \alpha$, o al $100(1 - \alpha)\%$, si es verifica:

1. $U < V$ per a tota mostra de grandària n.
2. $P(U < \theta < V) = 1 - \alpha$.

Definició. $1 - \alpha$ s'anomena **nivell de confiança**. α s'anomena **nivell d'error o de significació o de significança**.

Cal arribar a una mena d'equilibri entre els diversos aspectes (amplada, confiança, error, precisió, utilitat) que entren en joc quan es construeixen intervals de confiança.

Exemple

Suposem que volem fer una inferència sobre l'alçada mitjana de tots els habitants de Catalunya. Com que no es disposa de les alçades de tots els habitants de Catalunya (població), s'agafa una mostra de la població, es mira la seva alçada i es construeix un interval de confiança per a l'alçada mitjana de tota la població. Podem veure què passa en dues situacions extremes:

INTERVAL	*AMPLADA*	*CONFIANÇA*	*ERROR*	*PRECISIÓ*	*UTILITAT*
(1.50 , 2.00)	Gran	Gran	Petit	Poca	Poca
(1.7234 , 1.7236)	Petita	Petita	Gran	Molta	Molta

Observacions

1. U i V són els límits de confiança de θ. En general, són estimadors per defecte i per excés de θ.
2. L'interval depèn de la mostra.

3. El coeficient de confiança $1 - \alpha$ és un valor que escull l'experimentador. S'acostuma a agafar $\alpha = 0.10$, 0.05 o 0.01, és a dir, $1 - \alpha = 0.9$, 0.95 o 0.99. Com més gran és $1 - \alpha$, més gran és el grau de confiança, però també serà més gran l'interval de confiança. Interessa obtenir un interval reduït, però amb una probabilitat relativament alta de contenir el valor del paràmetre.
4. El valor de θ és constant, mentre que l'interval de confiança $I = (U,V)$ és d'extrems aleatoris, incloent-hi el vertader valor de θ amb probabilitat $1 - \alpha$.
5. L'interval de confiança s'ha d'interpretar segons una visió freqüencial en el sentit següent: si, per exemple, agafem $1 - \alpha = 0.95$, en una llarga sèrie de determinació d'intervals de confiança, en el 95% dels casos l'interval inclourà el vertader valor de θ.

4.5 Determinació d'intervals de confiança

A la taula d'intervals es donen les instruccions per calcular diferents intervals de confiança, segons les condicions amb les quals estem treballant i el paràmetre que volem estimar.

Els passos que hauríem de seguir per construir un interval de confiança serien:

1. Determinar de què es vol fer l'interval:
   a) una mitjana poblacional
   b) una diferència de mitjanes poblacionals
   c) una variància poblacional
   d) un quocient de variàncies poblacionals
   e) una proporció poblacional
   f) una diferència de proporcions poblacionals
2. Determinar les condicions generals de l'exercici per treballar amb l'interval adient.
3. Consultar les taules estadístiques i determinar el valor corresponent que cal per trobar el marge d'error de l'interval.
4. Calcular el marge d'error de l'interval.
5. Calcular els extrems inferior i superior de l'interval de confiança.

Notació. La notació emprada a la taula d'intervals és la següent:

- n és el nombre d'elements de la mostra. Si té un subíndex, significa que hi ha més d'una mostra i amb el subíndex indiquem amb quina mostra treballem.

- $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ és la mitjana de la mostra. El significat dels subíndexs és el mateix que en el cas anterior.
- $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$ és la variància de la mostra. El significat dels subíndexs és el mateix que en el primer cas. Cal observar que, per calcular la variància mostral, s'ha de dividir entre $n - 1$ i no entre n (com fèiem en el tema d'estadística descriptiva), ja que l'estimador S^2 té millors propietats que si dividíssim entre n.
- $\hat{p}$ és la proporció en què apareix certa característica en una mostra. El significat dels subíndexs és el mateix que en el primer cas.
- $z_{\alpha/2}$ és el valor trobat a la taula $N(0,1)$ que deixa a la seva dreta una àrea igual a $\alpha/2$.
- $t_{\alpha/2}$ és el valor trobat a la taula t de Student que deixa a la seva dreta una àrea igual a $\alpha/2$.
- $\chi^2_{\alpha/2}$ i $\chi^2_{1-\alpha/2}$ són els valors trobats a la taula χ^2 que deixa a la seva dreta una àrea igual a $\alpha/2$ i $1 - \alpha/2$, respectivament.
- $F_{\alpha/2}$ i $F_{1-\alpha/2}$ són els valors trobats a la taula F de Fisher que deixa a la seva dreta una àrea igual a $\alpha/2$ i $1 - \alpha/2$, respectivament.

Observacions

1. Dues mostres, tant si tenen la mateixa mida com si no, són independents quan s'observa una variable sobre individus diferents. Per exemple, es mesura el pes abans de fer un règim en 8 persones, i després de fer el règim es mesura el pes en altres 8 persones diferents.
2. Dues mostres són dependents quan s'observa una variable sobre els mateixos individus. En aquest cas, les mides de les mostres seran iguals i té sentit crear una nova mostra que sigui la diferència individu a individu de les mostres originals. En aquesta situació, per fer els càlculs es treballarà sobre aquesta nova mostra i no sobre les mostres originals; és a dir, es calcularà la mitjana i la desviació estàndard d'aquesta nova mostra. Per exemple, es mesura el pes abans de fer un règim en 8 persones i després de fer el règim es mesura el pes en les mateixes 8 persones anteriors. També es consideren mostres dependents quan, per exemple, es mesuren els temps assolits per dos atletes de 100 metres

llisos en 6 carreres diferents on han participat els dos atletes i on se suposa que les condicions (de vent, temperatura i altura sobre el nivell del mar) són les mateixes per als dos atletes; serien mostres independents si les condicions fossin diferents per als dos atletes.

Mostra 1	*Mostra 2*	*Mostra 1 – Mostra 2*
x_{11}	x_{21}	$x_{11} - x_{21}$
x_{12}	x_{22}	$x_{12} - x_{22}$
...	...	...
x_{1n}	x_{2n}	$x_{1n} - x_{2n}$

Exemple 1. Amb l'objectiu de verificar el pes mitjà de les caixes de cereals de certa marca, s'han agafat 16 d'aquestes caixes i s'han pesat. Els resultats són: 506, 508, 499, 503, 504, 510, 497, 512, 514, 505, 493, 496, 506, 502, 509 i 496. Si suposem que el pes de les caixes segueix una v. a. normal amb variància desconeguda, busquem un interval de confiança del 95% per al pes mitjà de les caixes de cereals.

Solució

1. Es vol construir un interval per al pes mitjà de les caixes de cereals, per tant, d'una mitjana poblacional.
2. Ens diuen que la distribució del pes de les caixes de cereals, la variable que es vol estudiar, segueix una normal i no es coneix la variància poblacional σ^2. A més, la mostra és de 16 elements. Per això, agafarem el tercer interval de la taula.
3. Cal consultar les taules de la t de Student amb 15 graus de llibertat i mirar quin punt deixa a la seva dreta una àrea de $\alpha/2 = 0.05/2 = 0.025$. Aquest valor és $t_{0.025} = 2.13$.
4. El marge d'error de l'interval es calcula a partir de $t_{\alpha/2}\frac{S}{\sqrt{n}}$. Sabent que $S = 6.2$, $n = 16$ i $t_{0,025} = 2.13$, tenim que el marge d'error de l'interval és 3.30.
5. Com que $\overline{X} = 503.75$ i l'interval es construeix com a $\overline{X}$ ± marge d'error, l'interval que obtenim és $500.45 < \mu < 507.05$. Aquest resultat vol dir que, amb un 95% de confiança, la mitjana real del pes de les caixes de cereals es troba en aquest interval.

Exemple 2. Per comparar dos mètodes pedagògics diferents, s'han fet uns tests a dos grups d'alumnes, cada grup dels quals ha après segons un mètode diferent. El primer grup és de 10 alumnes i el segon de 16 alumnes. Suposem que les puntuacions segueixen v. a. normals amb esperança i variància desconegudes i iguals. Els resultats van ser $\overline{X}_1 = 6.3$, $S_1^2 = 3.4$, $\overline{X}_2 = 5.8$ i $S_2^2 = 3.1$. Volem un interval de confiança del 90% per a la diferència d'esperances.

Solució

1. Es vol construir un interval per a la diferència de puntuacions mitjanes obtingudes segons cada mètode pedagògic, per tant, d'una diferència de mitjanes poblacionals.
2. Ens diuen que la distribució de les puntuacions de cada mètode segueix unes distribucions normals, amb variàncies desconegudes però iguals a les dues poblacions. Per això, agafarem el sisè interval de la taula.
3. Cal consultar les taules de la t de Student amb $10 + 16 - 2 = 24$ graus de llibertat i mirar quin punt deixa a la seva dreta una àrea de $\alpha / 2 = 0.1 / 2 = 0.05$. Aquest valor és $t_{0.05} = 1.71$.
4. El marge d'error de l'interval es calcula a partir de

$$t_{\alpha/2}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right)}.$$

   Amb les dades de l'enunciat tenim que el marge d'error de l'interval és 1.24.
5. Com que $\overline{X}_1 = 6.3$ i $\overline{X}_2 = 5.8$ i l'interval es construeix com a $\overline{X}_1 - \overline{X}_2 \pm$ marge d'error, l'interval que obtenim és $-0.74 < \mu_1 - \mu_2 < 1.74$. Aquest resultat vol dir que, amb un 90% de confiança, la diferència de les esperances de les puntuacions segons els mètodes proposats està en aquest interval.

4.5.1 Càlcul del nivell d'error associat a un marge d'error donat

Ens podríem demanar amb quin nivell de confiança (o d'error) caldria treballar per aconseguir un determinat marge d'error. En aquest cas, caldria igualar la fórmula del marge d'error corresponent al valor concret del marge d'error desitjat deixant com a incògnita el valor de taules estadístiques. A partir del valor de taules estadístiques es pot trobar el nivell de confiança o el d'error.

4.5.2 Càlcul de la mida d'una mostra

Donat el nivell de confiança (o el d'error), ens podríem demanar quina ha de ser la mida de la mostra per aconseguir un cert marge d'error. En aquest cas, caldria igualar la fór-

mula del marge d'error corresponent al valor concret del marge d'error desitjat deixant com a incògnita la mida de la mostra. Aquesta variant es podria aplicar als intervals 1 (interval sobre la mitjana poblacional amb variància poblacional coneguda), 2 (interval sobre la mitjana poblacional amb variància poblacional desconeguda i suposant que la mida de la mostra, tot i que no es coneix *a priori*, serà gran) i 10 (interval sobre la proporció poblacional).

Observacions

1. El cas de l'interval 1 realment serà difícilment aplicable, perquè, generalment, la variància poblacional serà desconeguda.
2. Per al cas de l'interval 2, si es demana la mida de la mostra, implica que encara no s'ha agafat realment cap mostra, però en el càlcul de la mida de la mostra intervé la variabilitat de les dades (a través de la desviació estàndard d'una mostra). En aquest cas, per tenir una estimació de la variabilitat de les dades, caldria agafar una mostra prèvia i usar la variabilitat d'aquesta mostra prèvia per determinar la mida de la mostra definitiva.
3. En el cas de l'interval 10 també caldria tenir una estimació prèvia de la proporció mostral que pot sortir de la mostra definitiva. De totes maneres, per curar-se en salut, es pot agafar com a proporció mostral $\hat{p} = 0.5$ i llavors tindríem que:

$$n = 0{,}25\left(\frac{z_{\alpha/2}}{marge\ \ error}\right)^2$$

CONDICIONS	*PARÀMETRE*	*MARGE D'ERROR*	*INTERVAL*	*GRAUS LLIBERTAT*
X normal, σ coneguda X no normal, σ coneguda $n \geq 30$	μ	$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\overline{X}$ ± marge error	
X qualsevol, σ desconeguda $n \geq 30$	μ	$z_{\alpha/2}\frac{S}{\sqrt{n}}$	$\overline{X}$ ± marge error	
X normal, σ desconeguda X_1, X_2 normals, mostres dependents	μ $\mu_1 - \mu_2$	$t_{\alpha/2}\frac{S}{\sqrt{n}}$	$\overline{X}$ ± marge error $\overline{X}_d$ ± marge error	$n-1$
X_1, X_2 normals, indep. σ_1, σ_2 conegudes X_1, X_2 no normals, indep. σ_1, σ_2 conegudes $n_1, n_2 \geq 30$	$\mu_1 - \mu_2$	$z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$\overline{X}_1 - \overline{X}_2$ ± marge error	
X_1, X_2 qualssevol, indep. σ_1, σ_2 desconegudes $n_1, n_2 \geq 30$	$\mu_1 - \mu_2$	$z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$	$\overline{X}_1 - \overline{X}_2$ ± marge error	
X_1, X_2 normals, indep. σ_1, σ_2 desconegudes $\sigma_1 = \sigma_2$	$\mu_1 - \mu_2$	$t_{\alpha/2}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right)}$	$\overline{X}_1 - \overline{X}_2$ ± marge error	n_1+n_2-2
X_1, X_2 normals, indep. σ_1, σ_2 desconegudes $\sigma_1 \neq \sigma_2$	$\mu_1 - \mu_2$	$t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$	$\overline{X}_1 - \overline{X}_2$ ± marge error	$n_1+n_2-2-\frac{(n_2(n_2-1)s_1^2-n_1(n_1-1)s_2^2)^2}{n_2^2(n_2-1)s_1^4+n_1^2(n_1-1)s_2^4}$

CONDICIONS	*PARÀMETRE*	*MARGE D'ERROR*	*INTERVAL*	*GRAUS LLIBERTAT*
X normal	σ^2		$\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$	$n-1$
X_1, X_2 normals, indep.	$\sigma_2^{\,2} / \sigma_1^{\,2}$		$F_{1-\alpha/2}\frac{S_2^2}{S_1^2}, F_{\alpha/2}\frac{S_2^2}{S_1^2}$	(n_1-1, n_2-1)
X binomial, $n \geq 30$	p	$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\hat{p}$ ± marge error	
X_1, X_2 binomials, indep. $n_1, n_2 \geq 30$	$p_1 - p_2$	$z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$\hat{p}_1 - \hat{p}_2$ ± marge error	

5. Contrastos d'hipòtesis

El contrast d'hipòtesis es relaciona amb el conjunt de tècniques i mètodes estadístics que tenen com a objectiu la verificació de determinades afirmacions o suposicions fetes sobre algun o alguns paràmetres desconeguts que caracteritzen una població estadística. En aquest tema només tractarem els contrastos d'hipòtesis quan estem mostrejant poblacions normals o quan el nombre d'elements de les mostres sigui prou gran.

Exemple. Suposem que tenim una moneda i fem l'afirmació que la moneda és correcta: la probabilitat d'obtenir cara és la mateixa que la d'obtenir creu; per tant, aquesta probabilitat és 0.5. Com es pot comprovar aquesta afirmació? Una manera seria fer un nombre elevat de llançaments i comptar quantes cares s'han obtingut. Segons el nombre de cares obtingudes decidirem si la moneda és correcta o, en canvi, està trucada.

Per exemple, si llancem la moneda 200 vegades i ens surten 15 cares, sospitarem que la moneda està trucada; en canvi, si en aquests llançaments s'obtenen 95 cares, podrem afirmar que la moneda és correcta.

El problema és determinar, amb un cert nivell d'error, quin és el nombre de cares que fa de frontera entre una decisió i l'altra. Per tant, hem de trobar un interval, una zona, on, si el nombre de cares que hem obtingut està dintre de l'interval, acceptarem que la moneda és correcta.

Suposem que, en llançar la moneda, el nombre de cares que surt no es troba dins de l'interval proposat. Aquest fet pot ser degut a dues causes:

1. La moneda no és correcta i és lògic el resultat que hem obtingut. Haurem de rebutjar la hipòtesi que $p = 0.5$.
2. La moneda és correcta, però el resultat obtingut és estrany.

Entre aquestes dues alternatives sembla més raonable justificar el resultat per la primera causa.

L'exemple comentat conté alguns **elements bàsics de la teoria del contrast d'hipòtesis:**

a) Especificació de les hipòtesis: p = 0.5 (la moneda és correcta) i p ≠ 0.5 (la moneda està trucada).

b) Definició d'un nivell d'error.

c) Construcció de zones d'acceptació (o de rebuig) de la hipòtesi proposada.

d) Determinació d'un estadístic de prova: el nombre de cares que han sortit.

e) Decisió final sobre la hipòtesi (acceptem o rebutgem la hipòtesi proposada).

5.1 Hipòtesis estadístiques. Tipus d'hipòtesis

Definició. Sigui una població estadística caracteritzada pel comportament d'una variable la distribució de la qual depèn d'un vector de paràmetres θ desconegut. Definim l'**espai paramètric** com el conjunt de valors compatibles amb θ, és a dir, els possibles valors que pot prendre θ. El notarem amb el símbol Ω.

Exemples

a) Si *p* és una proporció poblacional, llavors:

$$\Omega = \{p \mid 0 \leq p \leq 1\}$$

b) Si $X \sim N(\mu,\sigma)$, amb μ i σ desconeguts:

$$\Omega = \{(\mu,\sigma) \mid -\infty < \mu < \infty\ ,\ \sigma^2 > 0\}$$

c) Si $X \sim N(\mu,\sigma)$, amb $\sigma = \sigma_0$ conegut:

$$\Omega = \{(\mu,\sigma_0) \mid -\infty<\mu<\infty\}$$

d) Si $X \sim N(\mu,\sigma)$, amb $\mu = \mu_0$ conegut:

$$\Omega = \{(\mu_0,\sigma) \mid \sigma^2 > 0\}$$

Definició. Una **hipòtesi estadística** (o simplement *hipòtesi*) és una suposició que determina, parcialment o totalment, la distribució de probabilitat d'una v. a. Les hipòtesis es poden classificar en dos grups:

a) Les que especifiquen un valor concret o un interval per al vector de paràmetres del model. Per exemple, $p = 0.5$ o $\mu < 4$.

b) Les que determinen el tipus de distribució de probabilitat que ha generat les dades. Per exemple, la distribució de la variables que s'està estudiant és normal.

Encara que la metodologia per realitzar el contrast d'hipòtesis és semblant en els dos casos, distingir entre els dos tipus d'hipòtesis és important, perquè molts proble-

mes de contrast d'hipòtesis respecte d'un paràmetre són en realitat problemes d'estimació, que tenen una resposta més clara donant un interval de confiança per al paràmetre que es vol estimar. En canvi, les hipòtesis respecte a la forma de la distribució pertanyen a la fase de diagnòstic i validació del model i s'estudien a banda. En aquest tema ens centrarem en les hipòtesis del primer grup.

Definició. Anomenarem **hipòtesi simple** aquella que especifica un únic valor de l'espai paramètric. En cas contrari, estem davant d'una **hipòtesi composta**.

Definició. Anomenarem **hipòtesi nul·la**, H_0, la hipòtesi que es contrasta. H_0 representa la hipòtesi que mantindrem llevat que les dades n'indiquin la falsedat. Complementàriament a H_0 es defineix la **hipòtesi alternativa**, H_1. Quan rebutgem H_0 estem acceptant una hipòtesi alternativa: que H_0 és falsa. Un contrast implica l'elecció entre dues hipòtesis: la H_0 que contrastem i una hipòtesi alternativa, H_1, que està implícita en el rebuig de H_0.

La H_0 i H_1 no tenen un comportament simètric, és a dir, si tenim dues hipòtesis, no és indiferent quina s'agafa com a H_0 i quina s'agafa com a H_1. Al final s'acceptarà una hipòtesi o l'altra en funció de quina sigui més coherent amb les evidències de les proves (les dades de la mostra) que hi hagi. En general:

- Si les proves demostren que és certa la H_1, s'accepta la H_1.
- Si les proves demostren que és certa la H_0, s'accepta la H_0.
- Si les proves són dubtoses, s'accepta la H_0.

Per això, quan s'accepta la H_1 és que realment s'ha demostrat que és certa, mentre que quan s'accepta la H_0 realment el que pot haver passat és que no s'hagi demostrat que la H_1 sigui certa. Per aquest motiu, moltes vegades, en comptes de dir que s'accepta la H_0 es diu que no hi ha prou evidències que la H_1 sigui certa. Per tant, per regla general, la hipòtesi que es vol demostrar és la que triarem com a H_1.

Quan es fa un contrast d'hipòtesis, sovint existeix més d'una hipòtesi alternativa respecte d'una hipòtesi simple nul·la H_0, és a dir, tenim una hipòtesi simple davant d'una hipòtesi alternativa composta. Ens podem trobar amb tres casos diferents:

Cas 1 (H_1 bilateral)	Cas 2 (H_1 unilateral dreta)	Cas 3 (H_1 unilateral esquerra)
$H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$	$H_0: \theta \leq \theta_0$ $H_1: \theta > \theta_0$	$H_0: \theta \geq \theta_0$ $H_1: \theta < \theta_0$

Exemples

a) El cas de comprovar si una moneda és correcta o està trucada es correspondria amb el cas 1, ja que la hipòtesi alternativa és bilateral:

$$H_0: p = 0.5$$
$$H_1: p \neq 0.5$$

b) Si anem a un concessionari de cotxes i el venedor ens diu que el consum d'un determinat model de cotxe és de 3 litres cada 100 quilòmetres i no ens acabem de creure que aquest consum sigui el real, estarem en el cas 2, ja que el que nosaltres voldrem demostrar, H_1, és que el consum mitjà del cotxe és superior a 3 litres cada 100 quilòmetres (si és menor, no protestarem):

$$H_0: \mu \leq 3$$
$$H_1: \mu > 3$$

c) Si som d'una associació de consumidors i volem demostrar que l'empresa que envasa l'aigua en ampolles de 50 cl ens estafa, estarem en el cas 3, ja que el que nosaltres voldrem demostrar, H_1, és que la quantitat mitjana d'aigua per ampolla és inferior a 50 cl (si és superior, no ens estaran estafant):

$$H_0: \mu \geq 50$$
$$H_1: \mu < 50$$

5.2 Concepte de zona crítica i zona d'acceptació

Definició. Per fer més operativa la decisió que s'ha de prendre, es defineix un **estadístic de prova** com una funció dels elements mostrals que no depengui explícitament dels paràmetres poblacionals desconeguts. En l'exemple de la moneda, l'estadístic de prova era el nombre de cares que surten en 200 llançaments. Prenent com a referència aquest estadístic de prova i a partir d'un punt crític *c*, es determinaran les zones crítica i d'acceptació de cada hipòtesi.

Definició: si l'estadístic de prova pren un valor que està dintre d'un rang de valors "coherents" amb la hipòtesi nul·la, acceptarem aquesta hipòtesi. Aquest rang de valors l'anomenarem **zona d'acceptació A**. La **zona crítica**, o **zona de rebuig,** serà la zona complementària a la zona d'acceptació. Per tant, qualsevol mostra estarà en una zona o una altra.

La regla de decisió per decidir entre H_0 i H_1 serà la següent:

a) si es presenta el succés {estadístic de prova $\in A$} o {mostra $\in A$}, s'accepta H_0.

b) si es presenta el succés {estadístic de prova $\notin A$} o {mostra $\notin A$}, s'accepta H_1.

5.3 Tipus d'errors. Nivell de significació

Contrastar una hipòtesi suposa que hem de prendre una decisió en la qual hem d'acceptar o rebutjar H_0. Si s'accepta la hipòtesi nul·la s'està rebutjant la hipòtesi alternativa; si es rebutja la hipòtesi nul·la s'està acceptant la hipòtesi alternativa. Per tant, en el contrast d'hipòtesis es poden cometre dos tipus d'errors:

		Decisió	
		Zona acceptació, H_0	Zona crítica, H_1
Hipòtesi certa	H_0	No hi ha error	Error de tipus I
	H_1	Error de tipus II	No hi ha error

Definició. Definim:
Error de tipus I: rebutjar la hipòtesi nul·la quan és certa.
Error de tipus II: acceptar la hipòtesi nul·la quan és falsa.
També s'anomenen **errors de 1a i 2a espècie**, respectivament.

Definició. La probabilitat de cometre un error de tipus I es coneix amb el nom de **nivell de significació** α del contrast. α es fixa, normalment, en 0.1, 0.05 o 0.01, depenent de la importància de la hipòtesi en joc. Denotarem mitjançant β la probabilitat de fer un error de tipus II:

$$\alpha = P(\text{error de tipus I}) = P(\text{rebutjar } H_0 \text{ sent certa})$$

$$\beta = P(\text{error de tipus II}) = P(\text{acceptar } H_0 \text{ sent falsa})$$

Els valors α i β mantenen entre si una relació inversa: per a una determinada mida mostral, si α el fem més petit, β serà més gran i a l'inrevés. L'única manera de disminuir ambdós a la vegada és augmentant la mida de la mostra.

Definició. Es defineix la **potència del contrast, 1 – β**, com la probabilitat de rebutjar H_0 sent falsa. Per determinar les zones d'acceptació de cada hipòtesi, fixat un nivell de significació α, s'intentarà que tingui la màxima potència, ja que llavors l'error de tipus II serà el més petit possible.

El procediment de selecció d'una zona crítica mitjançant el nivell de significació té dues crítiques principals:

1. El resultat del test pot dependre molt del valor de α, que és arbitrari, sent possible rebutjar H_0 amb $\alpha = 0.05$ i acceptar-la amb $\alpha = 0.04$.
2. Donar només el resultat del test no permet diferenciar el grau d'evidència que la mostra indica a favor o en contra de H_0.

Definició. Per contrarestar aquestes crítiques, definirem de manera simple el **nivell crític** α_c o **p-valor** com, donada una mostra, la probabilitat que sigui certa la H_0.

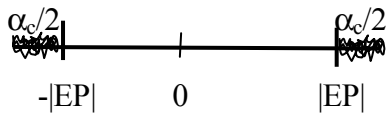
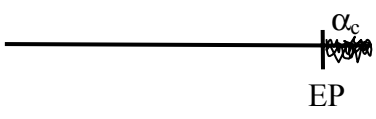
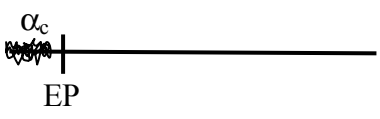
Anomenant EP el valor observat de l'estadístic de prova i suposant que la H_1 sigui unilateral dreta, tenim que:

$$\alpha_c = P(H_0 \geq EP)$$

Si H_1 és unilateral esquerra, obtenim que: $\alpha_c = P(H_0 \leq EP)$.

Si H_1 és bilateral i la distribució de l'estadístic de prova quan H_0 és certa és simètrica respecte del zero, aleshores $\alpha_c = P(|H_0| \geq EP)$.

Esquemàticament tindrem:

H_1 bilateral	*H_1 unilateral dreta*	*H_1 unilateral esquerra*
		

Per tant, el valor de α_c no es fixa *a priori*, com passava amb el nivell de significació α, sinó que es determina a partir de la mostra. Donada la interpretació del nivell crític α_c com, donada una mostra, la probabilitat que sigui certa la H_0, és lògic pensar que, com més petit sigui el valor de α_c, menys possibilitats hi ha que sigui certa la H_0 i, al contrari, més n'hi haurà que sigui certa la H_1. Basant-nos en α, la regla per decidir entre H_0 i H_1 serà la següent:

- Si $\alpha_c < \alpha$, acceptarem H_1.
- Si $\alpha_c > \alpha$, acceptarem H_0.

El valor de α amb què es vol treballar el decideix l'experimentador. Normalment, els programes informàtics d'estadística ens donen el p-valor α_c. Com a regla orientativa es pot dir que:

- Si $\alpha_c < 0.01$, la H_1 s'acceptarà amb una seguretat molt alta.
- Si $0.01 < \alpha_c < 0.1$, la hipòtesi que s'acceptarà dependrà del valor concret de α amb què es vol treballar, però hi ha força evidències que la H_1 és certa.
- Si $0.1 < \alpha_c < 0.25$, la hipòtesi que s'acceptarà generalment serà la H_0, però hi haurà molts dubtes que sigui realment certa. Tampoc no hi ha gaires evidències que la certa sigui la H_1.
- Si $0.25 < \alpha_c$, la H_0 s'acceptarà amb una seguretat alta i que serà més forta com més alt sigui α_c.

5.4 Aplicació dels contrastos d'hipòtesis a diferents paràmetres i condicions

A la taula de contrastos es donen les instruccions per fer diversos contrastos d'hipòtesis, segons les condicions amb les quals estem treballant i el paràmetre sobre el qual volem fer el contrast.

A la columna de l'esquerra hi ha les condicions en les quals es pot aplicar cada contrast. A la segona columna hi trobem la hipòtesi nul·la que volem contrastar. A la columna central tenim l'estadístic de prova que hem de calcular en cada cas i la distribució que segueix. En les dues últimes columnes s'hi troben la hipòtesi alternativa (sempre n'hi ha 3 per a cada H_0) i el criteri que ha de complir l'estadístic de prova per acceptar la H_1 (el criteri és diferent per a cada H_1).

Sempre hem de tenir present quin és el paràmetre sobre el qual volem fer contrastos i les condicions amb les quals estem treballant per usar l'estadístic adient.

Els passos que hauríem de seguir per fer un contrast d'hipòtesis serien:

1. Determinar sobre què es vol fer el contrast (depèn de l'enunciat):
   a) una mitjana poblacional
   b) una comparació de mitjanes poblacionals
   c) una variància poblacional
   d) una comparació de variàncies poblacionals
   e) una proporció poblacional
   f) una comparació de proporcions poblacionals
2. Determinar les condicions generals de l'exercici per treballar amb l'estadístic de prova adient.
3. Determinar la hipòtesi nul·la, H_0.
4. Determinar la hipòtesi alternativa, H_1.
5. Determinar la zona de les taules estadístiques on s'accepta la H_0.
6. Determinar la zona de les taules estadístiques on s'accepta la H_1. Serà la zona complementària a la trobada al pas 5).
7. Determinar la zona de l'estadístic mostral on s'accepta la H_0. Entenem per estadístic mostral l'estadístic que resulta de fer un càlcul directe sobre la mostra i que està associat al paràmetre sobre el qual es vol fer el contrast. Per determinar aquesta zona cal igualar, a la fórmula de l'estadístic de prova, l'*EP* als valors trobats en el pas 5) deixant com a incògnita l'estadístic mostral corresponent. Els estadístics mostrals són:
   a) la mitjana mostral si estem fent un contrast sobre una mitjana poblacional,
   b) la diferència de mitjanes mostral si estem fent un contrast sobre una diferència de mitjanes poblacionals,

c) la variància mostral si estem fent un contrast sobre una variància poblacional,

d) el quocient de variàncies mostrals si estem fent un contrast sobre una comparació de variàncies poblacionals,

e) la proporció mostral si estem fent un contrast sobre una proporció poblacional,

f) la diferència de proporcions mostrals si estem fent un contrast sobre una comparació de proporcions poblacionals.

8. Determinar la zona de l'estadístic mostral on s'accepta la H_1. Serà la zona complementària a la trobada al pas 7).
9. Calcular l'estadístic de prova, *EP*.
10. Segons els valors de l'EP i l'estadístic mostral i les zones d'acceptació de cada hipòtesi, decidir quina és la hipòtesi certa.
11. Calcular el valor del nivell de significació crític o p-valor α_c.
12. Segons els valors de α_c i de α, decidir quina és la hipòtesi certa.

Notació. La notació emprada a la taula dels contrastos és la següent:
n és el nombre d'elements de la mostra. Si té un subíndex, significa que hi ha més d'una mostra i amb el subíndex indiquem amb quina mostra treballem.

- $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ és la mitjana de la mostra. El significat dels subíndexs és el mateix que en el cas anterior.
- $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$ és la variància de la mostra. El significat dels subíndexs és el mateix que en el primer cas. Cal observar que per calcular la variància mostral s'ha de dividir entre $n - 1$ i no entre n (com fèiem al tema d'estadística descriptiva), ja que l'estimador S^2 té millors propietats que si dividíssim entre n.
- $\hat{p}$ és la proporció en què apareix certa característica en una mostra. El significat dels subíndexs és el mateix que en el primer cas.
- $z_{\alpha/2}$ i z_α són els valors trobats a la taula $N(0,1)$, que deixen a la seva dreta una àrea igual a $\alpha / 2$ i α, respectivament.
- $t_{\alpha/2}$ i t_α són els valors trobats a la taula t de Student que deixen a la seva dreta una àrea igual a $\alpha/2$ i α, respectivament.

- $\chi^2_{\alpha/2}$, $\chi^2_{1-\alpha/2}$, χ^2_{α} i $\chi^2_{1-\alpha}$ són els valors trobats a la taula χ^2 que deixen a la seva dreta una àrea igual a $\alpha / 2$, $1 - \alpha / 2$, α i $1 - \alpha$, respectivament.
- $F_{\alpha/2}$, $F_{1-\alpha/2}$, F_{α} i $F_{1-\alpha}$ són els valors trobats a la taula F de Fisher que deixen a la seva dreta una àrea igual a $\alpha/2$, $1 - \alpha / 2$, α i $1 - \alpha$, respectivament.

Exemple 1. Amb l'objectiu de verificar el consum d'un cotxe s'han agafat 16 mesures del consum d'aquest cotxe en trajectes de 100 km. Els resultats són: 5.06, 5.08, 4.99, 5.03, 5.04, 5.10, 4.97, 5.12, 5.14, 5.05, 4.93, 4.96, 5.06, 5.02, 5.09 i 4.96. Si suposem que el consum del cotxe segueix una v. a. normal amb variància desconeguda, ¿existeix alguna raó per creure, amb $\alpha = 0.05$, que el consum mitjà del cotxe és superior a 5 litres cada 100 quilòmetres?

Solució

1. Es vol fer un contrast sobre el consum mitjà, per tant, sobre una mitjana poblacional.
2. Ens diuen que la distribució del consum del cotxe, la variable que es vol estudiar, segueix una normal i no es coneix la variància poblacional σ^2. A més, la mostra és de 16 elements. Per això, agafarem el tercer estadístic de la taula.
3. H_0: $\mu \leq 5$.
4. H_1: $\mu > 5$.
5. S'ha de consultar la taula t de Student amb $16 - 1 = 15$ graus de llibertat. La H_1 és unilateral dreta i el valor de α és 0.05. Per tant, el punt $t_{0.05}$ és 1.75. En definitiva, la zona de les taules estadístiques on s'accepta la H_0 és $(-\infty , 1.75)$.
6. La zona de les taules estadístiques on s'accepta la H_1 és la complementària a la trobada al pas 5), per tant, $(1.75 , \infty)$.
7. Tenim l'$EP = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$ i l'igualem a 1.75 deixant $\bar{x}$ com a incògnita, és a dir, resolem $\dfrac{\overline{X} - 5}{0.062/\sqrt{16}} = 1.75$ i obtenim $\bar{x} = 5.03$. En definitiva, la zona de la mitjana mostral on s'accepta la H_0 és $(-\infty, 5.03)$.
8. La zona de la mitjana mostral on s'accepta la H_1 és la complementària a la trobada al pas 7), per tant, $(5.03 , \infty)$.
9. Els resultats mostrals són $\bar{x} = 5.0375$, $n = 16$, $S = 0.062$ i, per tant, $EP = 2.42$.
10. Com que la mitjana de la mostra és $\bar{x} = 5.0375$ i aquest valor es troba a la zona de la mitjana mostral on s'accepta la H_1, acceptem aquesta H_1, amb un $\alpha = 0.05$,

és a dir, el consum d'aquest cotxe és superior a 5 litres cada 100 km. Arribem a la mateixa conclusió si comprovem que $EP = 2.42$ es troba a la zona de les taules estadístiques on s'accepta la H_1.

11. Com que la H_1 és unilateral dreta i estem treballant amb el tercer estadístic de la taula i EP = 2.42, cal buscar l'àrea que hi ha a la dreta de l'EP a les taules de la t de Student amb 15 graus de llibertat. Interpolant, el resultat és $\alpha_c = 0.0143$.

12. Si volem treballar amb $\alpha = 0.05$ i com $\alpha_c = 0.0143 < \alpha = 0.05$, s'accepta H_1, és a dir, el consum d'aquest cotxe és superior a 5 litres cada 100 km.

Exemple 2. Per comparar dos mètodes pedagògics diferents, s'han fet uns tests a dos grups d'alumnes; cada grup ha après segons un mètode diferent. El primer grup és de 10 alumnes i el segon de 16 alumnes. Suposem que les puntuacions segueixen v. a. normals amb esperança i variància desconegudes i iguals. Els resultats van ser $\overline{x_1} = 6.3$, $S_1^2 = 3.4$, $\overline{x_2} = 5.8$ i $S_2^2 = 3.1$. Es tracta de verificar, amb $\alpha = 0.1$, la suposició que les variàncies poblacionals són iguals.

Solució

1. Es vol fer un contrast sobre una comparació de variàncies poblacionals.
2. Ens diuen que la distribució de les puntuacions amb un mètode i l'altre segueixen una normal i les mostres s'han agafat de manera independent. Usarem el novè estadístic de la taula.
3. H_0: $\sigma_1^2 = \sigma_2^2$.
4. H_1: $\sigma_1^2 \neq \sigma_2^2$.
5. S'ha de consultar la taula F de Fisher amb 9 i 15 graus de llibertat. La H_1 és bilateral i el valor de α és 0,1. Per tant, $\alpha / 2 = 0{,}05$ i el punt $F_{0.05}$ és 2.59 i el punt $F_{0.95}$ és $1 / 3.01 = 0.33$. En definitiva, la zona de les taules estadístiques on s'accepta la H_0 és (0.33, 2.59).
6. La zona de les taules estadístiques on s'accepta la H_1 és la complementària a la trobada al pas 5), per tant, serà $(0, 0.33) \cup (2.59 , \infty)$.
7. En aquest cas, l'estadístic mostral (quocient de variàncies mostrals) coincideix amb la definició de l'*EP* i, per tant, la zona del quocient de variàncies mostrals on s'accepta la H_0 també és (0.33, 2.59).
8. La zona del quocient de variàncies mostrals on s'accepta la H_1 és la complementària a la trobada al pas 7), per tant, $(0, 0.33) \cup (2.59 , \infty)$.
9. Els resultats mostrals són $S_1^2 = 3.4$ i $S_2^2 = 3.1$ i, per tant, $EP = 1.1$.

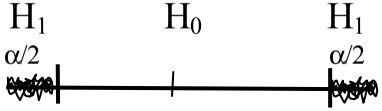
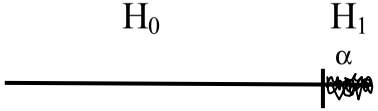
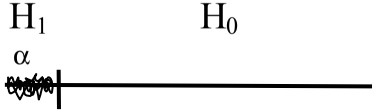
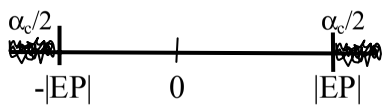
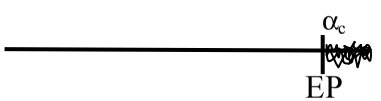
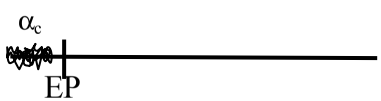
10. Com que $EP = 1.1$ es troba a la zona de les taules on s'accepta la H_0, acceptem aquesta H_0 amb un $\alpha = 0.1$, és a dir, les variàncies poblacionals podem assumir que són iguals (no hi ha evidències significatives que siguin diferents).

11. Com que la H_1 és bilateral i estem treballant amb el novè estadístic de la taula i $EP = 1.1$, cal buscar l'àrea que hi ha a la dreta de l'EP a les taules de la F de Fisher amb 9 i 15 graus de llibertat i multiplicar aquesta àrea per 2. El resultat és $\alpha_c = 0.8358$.

12. Si volem treballar amb $\alpha = 0.1$ i com que $\alpha_c = 0.8358 > \alpha = 0.1$, s'accepta H_0, és a dir, les variàncies poblacionals podem assumir que són iguals.

CONDICIONS	H_0	ESTADÍSTIC DE PROVA	H_1	ZONA H_1
X normal, σ coneguda	$\mu = \mu_0$	$EP = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$\mu \neq \mu_0$	$EP < -z_{\alpha/2}$ o $EP > z_{\alpha/2}$
X no normal, σ coneguda $n \geq 30$			$\mu > \mu_0$	$EP > z_\alpha$
			$\mu < \mu_0$	$EP < -z_\alpha$
X qualsevol, σ desconeguda $n \geq 30$	$\mu = \mu_0$	$EP = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim N(0,1)$	$\mu \neq \mu_0$	$EP < -z_{\alpha/2}$ o $EP > z_{\alpha/2}$
			$\mu > \mu_0$	$EP > z_\alpha$
			$\mu < \mu_0$	$EP < -z_\alpha$
X normal, σ desconeguda	$\mu = \mu_0$ $d = d_0$	$EP = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t\,(n-1)$ g. l.	$\mu \neq \mu_0$	$EP < -t_{\alpha/2}$ o $EP > t_{\alpha/2}$
X_1, X_2 normals, mostres dependents			$\mu > \mu_0$	$EP > t_\alpha$
			$\mu < \mu_0$	$EP < -t_\alpha$
X_1, X_2 normals, indep. σ_1, σ_2 conegudes	$d = d_0$ $(d = \mu_1 - \mu_2)$	$EP = \dfrac{(\overline{X_1} - \overline{X_2}) - d_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$	$d \neq d_0$	$EP < -z_{\alpha/2}$ o $EP > z_{\alpha/2}$
X_1, X_2 no normals, indep. σ_1, σ_2 conegudes $n_1, n_2 \geq 30$			$d > d_0$	$EP > z_\alpha$
			$d < d_0$	$EP < -z_\alpha$
X_1, X_2 qualssevol, indep. σ_1, σ_2 desconegudes $n_1, n_2 \geq 30$	$d = d_0$ $(d = \mu_1 - \mu_2)$	$EP = \dfrac{(\overline{X_1} - \overline{X_2}) - d_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim N(0,1)$	$d \neq d_0$	$EP < -z_{\alpha/2}$ o $EP > z_{\alpha/2}$
			$d > d_0$	$EP > z_\alpha$
			$d < d_0$	$EP < -z_\alpha$
X_1, X_2 normals, indep. σ_1, σ_2 desconegudes $\sigma_1 = \sigma_2$	$d = d_0$ $(d = \mu_1 - \mu_2)$	$EP = \dfrac{(\overline{X_1} - \overline{X_2}) - d_0}{\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)\left(\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right)}} \sim t\,(n_1+n_2-2)$ g. l.	$d \neq d_0$	$EP < -t_{\alpha/2}$ o $EP > t_{\alpha/2}$
			$d > d_0$	$EP > t_\alpha$
			$d < d_0$	$EP < -t_\alpha$
X_1, X_2 normals, indep. σ_1, σ_2 desconegudes $\sigma_1 \neq \sigma_2$	$d = d_0$ $(d = \mu_1 - \mu_2)$	$EP = \dfrac{(\overline{X_1} - \overline{X_2}) - d_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim t\,(n_1+n_2-2-\dfrac{(n_2(n_2-1)s_1^2 - n_1(n_1-1)s_2^2)^2}{n_2^2(n_2-1)s_1^4 + n_1^2(n_1-1)s_2^4})$ g. l.	$d \neq d_0$	$EP < -t_{\alpha/2}$ o $EP > t_{\alpha/2}$
			$d > d_0$	$EP > t_\alpha$
			$d < d_0$	$EP < -t_\alpha$

CONDICIONS	H_0	*ESTADÍSTIC DE PROVA*	H_1	*ZONA* H_1
X normal	$\sigma^2 = \sigma_0^2$	$EP = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2\,(n-1)$ g. l.	$\sigma^2 \neq \sigma_0^2$	$EP < \chi^2_{1-\alpha/2}$ o $EP > \chi^2_{\alpha/2}$
			$\sigma^2 > \sigma_0^2$	$EP > \chi^2_{\alpha}$
			$\sigma^2 < \sigma_0^2$	$EP < \chi^2_{1-\alpha}$
X_1, X_2 normals, indep.	$\sigma_1^2 = \sigma_2^2$	$EP = \dfrac{S_1^2}{S_2^2} \sim F\,(n_1-1, n_2-1)$ g. l.	$\sigma_1^2 \neq \sigma_2^2$	$EP < F_{1-\alpha/2}$ o $EP > F_{\alpha/2}$
			$\sigma_1^2 > \sigma_2^2$	$EP > F_{\alpha}$
			$\sigma_1^2 < \sigma_2^2$	$EP < F_{1-\alpha}$
X binomial, $n \geq 30$	$p = p_0$	$EP = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \sim N(0,1)$	$p \neq p_0$	$EP < -z_{\alpha/2}$ o $EP > z_{\alpha/2}$
			$p > p_0$	$EP > z_{\alpha}$
			$p < p_0$	$EP < -z_{\alpha}$
X_1, X_2 binomials, indep. $n_1, n_2 \geq 30$	$p_1 = p_2$	$EP = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1)$	$p_1 \neq p_2$	$EP < -z_{\alpha/2}$ o $EP > z_{\alpha/2}$
			$p_1 > p_2$	$EP > z_{\alpha}$
			$p_1 < p_2$	$EP < -z_{\alpha}$

	H_1 *bilateral*	H_1 *unilateral dreta*	H_1 *unilateral esquerra*
CÀLCUL ZONES H_0 I H_1			
CÀLCUL DEL P-VALOR (α_c)			

6. Anàlisi de la variància (ANOVA)

6.1 Generalitats sobre l'anàlisi de la variància

Davant un fenomen de naturalesa aleatòria, els possibles resultats poden estar influïts per una sèrie de condicionaments, externs i/o interns, els quals no sempre podem controlar. L'objecte de l'anàlisi de la variància se centra en la mesura de la influència d'aquests condicionaments en una variable resposta o observada. Anem a introduir una sèrie de conceptes per tenir el marc teòric on es desenvoluparà l'ANOVA:

a) *Variable resposta o observada*: variable que es vol estudiar si està influenciada per d'altres. La variable resposta serà una variable numèrica i fa el paper de variable dependent en la funció que relaciona el factor i la variable resposta. Per exemple, el nombre d'avellanes produïdes per avellaner pot ser una variable resposta.

b) *Factor*: condicionament que afecta el resultat d'un fenomen. Serà la qualitat o propietat a partir de la qual classifiquem les observacions. El factor serà una variable qualitativa o numèrica categoritzada en un nombre concret de valors i fa el paper de variable independent en la funció que relaciona el factor i la variable resposta. Per exemple, el nombre d'avellanes produïdes per avellaner pot dependre del sòl; en aquest cas el factor que observem és el sòl.

c) *Nivell*: cada una de les maneres en què es pot presentar un factor. Seguint amb l'exemple anterior, els diferents nivells del factor sòl podrien ser: calcari, sorrenc i argilós.

d) *Efecte assignable*: conseqüències dels factors considerats. En la producció d'avellanes seria la quantitat d'avellanes que correspondria al tipus de sòl on està situat cada avellaner.

e) *Efecte no assignable, residual o aleatori*: conseqüències que no provenen dels factors considerats. Per exemple, no s'han considerat els efectes del tipus d'adob, del tipus de reg o de la quantitat de pluja caiguda el mes de maig.

Per tant, cada observació la podem descompondre en una part deguda als efectes assignables i una altra deguda a l'atzar. Per fer l'anàlisi de la variància hem de suposar que les observacions de cada nivell del factor que volem analitzar provenen de variables aleatòries que es distribueixen normalment amb la mateixa variància, encara que s'ha vist que, si no es compleixen aquestes condicions, els resultats de l'anàlisi de la variància segueixen sent vàlids si les mides de les mostres de cada nivell són semblants.

6.2 Disseny ANOVA d'un factor

Primer estudiarem el cas en què només considerem un factor per explicar els resultats d'una sèrie d'observacions. L'objectiu serà trobar si existeixen diferències entre els diferents nivells considerats del factor o no existeixen. L'objectiu és semblant al que plantejàvem quan vam estudiar si hi havia diferències entre dues poblacions, però ara podem treballar amb més de dues poblacions o nivells, cosa que no podíem fer abans. Les dades les podem posar de la manera següent:

Població o nivell			
1	2	...	k
x_{11}	x_{21}	...	x_{k1}
.	.	...	.
.	.	...	.
.	.	...	.
x_{1n_1}	x_{2n_2}	...	x_{kn_k}

Amb aquestes dades podem calcular una sèrie d'estadístics:

$$\overline{x_{i\cdot}} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$ (mitjana mostral a la població o nivell i)

$$n = \sum_{i=1}^{k} n_i$$ (nombre total d'observacions)

$$\overline{x} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{n}$$ (mitjana mostral general)

Els passos que s'han de seguir per fer una anàlisi de la variància d'un factor serien:

1. Determinar quin és el factor que es vol considerar, quins nivells s'han agafat d'aquest factor i quina és la variable resposta o variable observada que es vol estudiar.
2. Comprovar que les observacions de cada nivell del factor es distribueixen normalment i que hi ha la mateixa variància en els diversos nivells.
3. Determinar la hipòtesi nul·la, H_0. La H_0 sempre serà la mateixa:

   $$H_0: \mu_1 = \ldots = \mu_k$$

   Això vol dir que, sota H_0, no existeixen diferències entre les diferents poblacions o entre els diferents nivells del factor considerat. També significa que el factor considerat no té influència sobre el resultat de la variable resposta.
4. Determinar la hipòtesi alternativa, H_1. La H_1 sempre serà la mateixa:

   $$H_1: \mu_i \neq \mu_j \text{ per a alguna parella } i \neq j$$

   H_1 significa que hi ha diferències entre, almenys, dues poblacions o nivells. També significa que el factor considerat té influència sobre el resultat de la variable resposta.
5. Determinar la zona de les taules on s'accepta la H_0. La taula que s'ha de consultar és la F de Fisher amb $k-1$ i $n-k$ graus de llibertat. Per a un nivell de significació α, si F_α és tal que $P(F > F_\alpha) = \alpha$ (és a dir, el punt de la taula F de Fisher que deixa a la seva dreta una àrea igual a α), la zona de les taules on s'accepta la H_0 és l'interval $(0, F_\alpha)$.
6. Determinar la zona de les taules on s'accepta la H_1. Serà la zona complementària a la trobada al pas 5), és a dir, serà l'interval (F_α, ∞).
7. Calcular l'estadístic de prova, que, en aquest cas, anomenarem F.

Els passos que s'han de seguir per calcular F els podem resumir a la taula següent:

Font de variació	*g. l.*	*Suma quadrats*	*Quadrats mitjans*	*F*
Entre grups	$k-1$	$Q_e = \sum_i n_i(\overline{x_{i\cdot}} - \overline{x})^2$	$\overline{Q_e} = \frac{Q_e}{k-1}$	$\frac{\overline{Q_e}}{\overline{Q_d}}$
Dintre grups	$n-k$	$Q_d = \sum_{i,j} (x_{ij} - \overline{x_{i\cdot}})^2$	$\overline{Q_d} = \frac{Q_d}{n-k}$	
Total	$n-1$	$Q_t = \sum_{i,j} (x_{ij} - \overline{x})^2$		

Per facilitar els càlculs, les sumes de quadrats es poden calcular amb fórmules algebraicament idèntiques:

$$Q_t = \sum_{i,j} (x_{ij} - \bar{x})^2 = \sum_{i,j} x^2_{ij} - n\bar{x}^2$$

$$Q_e = \sum_i n_i (\bar{x_{i\cdot}} - \bar{x})^2 = \sum_i n_i \bar{x_{i\cdot}}^2 - n\bar{x}^2$$

$$Q_d = Q_t - Q_e$$

8. Segons els valors de l'estadístic de prova F i les zones d'acceptació de cada hipòtesi, decidir quina és la hipòtesi certa.
9. Calcular el valor del nivell de significació crític o p-valor α_c. El p-valor es troba calculant la probabilitat que una distribució F de Fisher amb $k-1$ i $n-k$ graus de llibertat sigui més gran que l'estadístic de prova F, és a dir, $P(F_{k-1,n-k} > F) = \alpha_c$.
10. Segons els valors de α_c i de α, decidir quina és la hipòtesi certa.

Observació

En el cas que s'accepti que hi ha diferències entre almenys dos nivells o poblacions, ens pot interessar fer algun contrast parcial per comprovar entre quins nivells o poblacions hi ha diferències amb un nivell de significació α. Les hipòtesis per contrastar serien:

$$H_0^{(ij)}: \mu_i = \mu_j$$

$$H_1^{(ij)}: \mu_i \neq \mu_j$$

Aleshores, s'utilitza l'estadístic:

$$t = \frac{\bar{x_{i\cdot}} - \bar{x_{j\cdot}}}{\sqrt{\frac{Q_d}{n-k}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

Per contrastar la significativitat d'aquest estadístic s'ha de consultar la taula de la distribució t de Student amb $(n-k)$ graus de llibertat i mirar quin punt deixa a la seva dreta una àrea de $\alpha/2$.

Exemple

La gent que es preocupa per la seva salut prefereix hamburgueses que tenen poques calories. Les hamburgueses es poden classificar segons la seva composició: vedella, pollastre i carn (sobretot de carn de porc i vedella, però fins a un 15% de carn de pollastre). S'han agafat hamburgueses de 54 marques diferents, s'han classificat segons la seva composició i s'han mesurat les calories que contenen. Els resultats obtinguts són:

Vedella	186	181	176	149	184	190	158	139	175	148
	152	111	141	153	190	157	131	149	135	132
Pollastre	129	132	102	106	94	102	87	99	107	113
	135	142	86	143	152	146	144			
Carn	173	191	182	190	172	147	146	139	175	136
	179	153	107	195	135	140	138			

Suposant que les calories de cada grup d'hamburgueses es distribueixen normalment i que hi ha la mateixa variància en els diversos nivells, es vol analitzar si hi ha diferències significatives, amb $\alpha = 0.05$, entre les calories dels diversos grups d'hamburgueses.

Solució

1. El factor que es vol considerar és la composició de les hamburgueses. D'aquest factor es consideren 3 nivells: vedella, pollastre i carn. La variable observada que es vol estudiar són les calories.
2. Hem suposat que es compleixen les condicions per aplicar ANOVA d'un factor: les observacions de cada nivell del factor es distribueixen normalment i hi ha la mateixa variància en els diversos nivells.
3. La H_0 és:

$$H_0: \mu_{vedella} = \mu_{pollastre} = \mu_{carn}$$

   Això vol dir que, sota H_0, no existeixen diferències entre les calories mitjanes de les hamburgueses segons els diversos nivells considerats de composició.
4. La H_1 és:

$$H_1: \mu_i \neq \mu_j \text{ per a alguna parella } i \neq j$$

   H_1 significa que hi ha diferències entre les calories mitjanes de les hamburgueses d'almenys dos grups considerats de composició.
5. S'ha de consultar la taula F de Fisher amb (2,51) graus de llibertat. A la taula anterior hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $F_\alpha = 3.18$. Per tant, la zona de les taules on s'accepta la H_0 és l'interval (0 , 3.18).
6. La zona de les taules on s'accepta la H_1 és l'interval (3.18 , ∞).
7. Calculem l'estadístic de prova, F.

Font de variació	*g. l.*	*Suma quadrats*	*Quadrats mitjans*	*F*
Entre grups	2	17692.1951	8846.10	16.1
Dintre grups	51	28067.1382	550.33	
Total	53	45759.3333		

8. Com que l'estadístic de prova $F = 16.1$ es troba a l'interval on s'accepta H_1, direm que acceptem H_1 i que, per tant, les calories mitjanes de les hamburgueses no són les mateixes en els diversos grups de composició d'hamburgueses considerats.
9. El valor del nivell de significació crític o p-valor, α_c, és $3.86 \cdot 10^{-6}$.
10. Com que el p-valor, $\alpha_c = 3.86 \cdot 10^{-6}$, és més petit que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_1 i arribem a la mateixa conclusió que abans.
11. Com que s'ha acceptat que hi ha diferències entre les calories mitjanes entre almenys dos grups d'hamburgueses, anem a contrastar, amb $\alpha = 0.05$, si hi ha diferències entre les calories mitjanes de les hamburgueses de vedella i les de pollastre. El contrast que s'ha de realitzar és:

$$H_0^{(12)}: \mu_{vedella} = \mu_{pollastre}$$
$$H_1^{(12)}: \mu_{vedella} \neq \mu_{pollastre}$$

Com que $\alpha / 2 = 0.025$, busquem el punt de la taula t de Student, amb 51 graus de llibertat, que deixa a la seva dreta una àrea de 0.025. Aquest punt és 2.01 ja que $P(t_{51} > 2.01) = 0.025$. Per tant, les zones de la taula t de Student on s'accepta la hipòtesi nul·la i la hipòtesi alternativa són:

- Zona on s'accepta H_0: $(-2.01 , 2.01)$.
- Zona on s'accepta H_1: $(-\infty , -2.01) \cup (2.01 , \infty)$.

Tenim que:

$$\bar{x}_{vedella} = 156.85 \qquad \bar{x}_{pollastre} = 118.76$$

Aleshores, s'utilitza l'estadístic:

$$t = \frac{\bar{x}_{vedella} - \bar{x}_{pollastre}}{\sqrt{\frac{Q_d}{n-k}\left(\frac{1}{n_{vedella}} + \frac{1}{n_{pollastre}}\right)}} = \frac{156.85 - 118.76}{\sqrt{\frac{28067.1382}{54-3}\left(\frac{1}{20} + \frac{1}{17}\right)}} = 4.92$$

Com que el valor de l'estadístic *t* és 4.92, aquest valor es troba dins de la zona on s'accepta la H_1 i podem afirmar que hi ha diferències significatives entre les calories mitjanes de les hamburgueses de vedella i les hamburgueses de pollastre.

També s'arriba a aquesta conclusió si calculem el p-valor de l'estadístic *t*: $\alpha_c = 2 \cdot P(t_{51} > |4.92|) = 2 \cdot 4.6973 \cdot 10^{-6} = 9.3946 \cdot 10^{-6}$. Com que el p-valor, $\alpha_c = 9.3946 \cdot 10^{-6}$, és més petit que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_1.

12. També podem estar interessats a contrastar, amb $\alpha = 0.05$, si hi ha diferències entre les calories mitjanes de les hamburgueses de pollastre i les de carn. El contrast que s'ha de realitzar és:

$$H_0^{(23)}: \mu_{pollastre} = \mu_{carn}$$
$$H_1^{(23)}: \mu_{pollastre} \neq \mu_{carn}$$

Les zones de la taula t de Student on s'accepta la hipòtesi nul·la i la hipòtesi alternativa seran les mateixes d'abans:

- Zona on s'accepta H_0: (–2.01, 2.01).
- Zona on s'accepta H_1: $(-\infty, -2.01) \cup (2.01, \infty)$.

Tenim que:

$$\bar{x}_{pollastre} = 118.76 \qquad \bar{x}_{carn} = 158.71$$

Aleshores, s'utilitza l'estadístic:

$$t = \frac{\bar{x}_{pollastre} - \bar{x}_{carn}}{\sqrt{\frac{Q_d}{n-k}\left(\frac{1}{n_{pollastre}} + \frac{1}{n_{carn}}\right)}} = \frac{118.76 - 158.71}{\sqrt{\frac{28067.1382}{54-3}\left(\frac{1}{17} + \frac{1}{17}\right)}} = -4.96$$

Com que el valor de l'estadístic *t* és –4.96, aquest valor es troba dins de la zona on s'accepta la H_1 i podem afirmar que hi ha diferències significatives entre les calories mitjanes de les hamburgueses de pollastre i les hamburgueses de carn.

També s'arriba a aquesta conclusió si calculem el p-valor de l'estadístic *t*: $\alpha_c = 2 \cdot P(t_{51} > |-4.96|) = 2 \cdot 4.0541 \cdot 10^{-6} = 8.1082 \cdot 10^{-6}$. Com que el p-valor, $\alpha_c = 8.1082 \cdot 10^{-6}$, és més petit que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_1.

13. Per finalitzar, anem a contrastar, amb $\alpha = 0.05$, si hi ha diferències entre les calories mitjanes de les hamburgueses de vedella i les de carn. El contrast que s'ha de realitzar és:

$$H_0^{(13)}: \mu_{vedella} = \mu_{carn}$$
$$H_1^{(13)}: \mu_{vedella} \neq \mu_{carn}$$

Les zones de la taula t de Student on s'accepta la hipòtesi nul·la i la hipòtesi alternativa continuen sent les mateixes d'abans:

- Zona on s'accepta H_0: $(-2.01 , 2.01)$.
- Zona on s'accepta H_1: $(-\infty , -2.01) \cup (2.01 , \infty)$.

Tenim que:

$$\bar{x}_{vedella} = 156.85 \qquad \bar{x}_{carn} = 158.71$$

Aleshores, s'utilitza l'estadístic:

$$t = \frac{\bar{x}_{vedella} - \bar{x}_{carn}}{\sqrt{\frac{Q_d}{n-k}\left(\frac{1}{n_{vedella}} + \frac{1}{n_{carn}}\right)}} = \frac{156.85 - 158.71}{\sqrt{\frac{28067.1382}{54-3}\left(\frac{1}{20} + \frac{1}{17}\right)}} = -0.24$$

Com que el valor de l'estadístic t és -0.24, aquest valor es troba dins de la zona on s'accepta la H_0 i hem de concloure que no s'ha demostrat que hi hagi diferències significatives entre les calories mitjanes de les hamburgueses de vedella i les hamburgueses de carn.

També s'arriba a aquesta conclusió si calculem el p-valor de l'estadístic t: $\alpha_c = 2 \cdot P(t_{51} > |-0.24|) = 2 \cdot 0.4057 = 0.8114$. Com que el p–valor, $\alpha_c = 0.8114$, és més gran que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_0.

6.2 Excel: ANOVA d'un factor

El programa Excel permet fer de manera automàtica els càlculs que calen per realitzar un ANOVA d'un factor.

Prèviament hem de tenir instal·lat el mòdul de "Análisis de datos". Per instal·lar aquest mòdul cal seguir els passos següents:

1. Del menú "Herramientas", triem "Complementos".
2. Marquem la casella "Herramientas para análisis" i acceptem.
3. Comprovem que al final del menú "Herramientas" apareix "Análisis de datos".

Per fer una ANOVA d'un factor amb Excel cal seguir els passos següents:

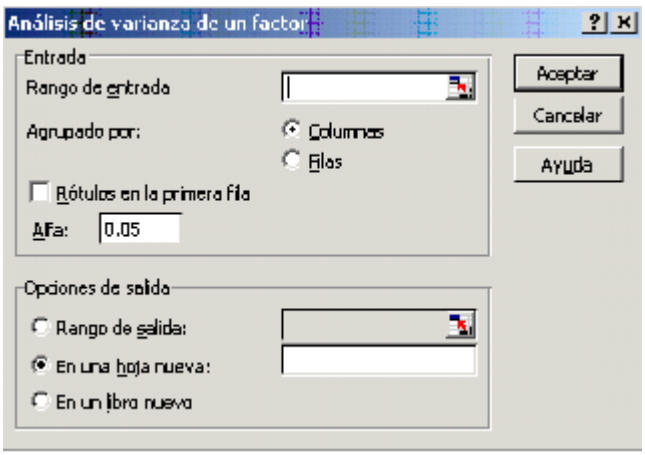
1. Per començar a fer qualsevol tipus d'anàlisi, primer hem de tenir les dades entrades en el full de càlcul. Si seguim amb l'exemple de les hamburgueses, hem de posar les dades de cada nivell (vedella, pollastre i carn) en columnes (o files) diferents; en començar la columna (o la fila) podem posar un rètol que indiqui a quin nivell corresponen les dades que hi ha en cada columna (o fila). La pantalla 1 ens mostra com han de quedar les dades una vegada introduïdes a Excel.

Pantalla 1. Dades d'un factor introduïdes a Excel

	A	B	C
1	Vedella	Pollastre	Carn
2	186	129	173
3	181	132	191
4	176	102	182
5	149	106	190
6	184	94	172
7	190	102	147
8	158	87	146
9	139	99	139
10	175	107	175
11	148	113	136
12	152	135	179
13	111	142	153
14	141	86	107
15	153	143	195
16	190	152	135
17	157	146	140
18	131	144	138
19	149		
20	135		
21	132		

2. Del menú "Herramientas", triem "Análisis de datos".
3. De les funcions que apareixen, triem "Análisis de varianza de un factor" i acceptem. Ha d'aparèixer el quadre que es veu a la pantalla 2.

Pantalla 2. Quadre de "Análisis de varianza de un factor"



4. Del quadre anterior, hem d'emplenar:
   a) **Rango de entrada**. Hem de seleccionar les caselles on es troben les dades que volem analitzar.
   b) **Agrupado por**. Hem de triar "Columnas" o "Filas" depenent de com hàgim entrat les dades de cada nivell (en el nostre exemple, hem de marcar "Columnas").
   c) **Rótulos**. Hem de marcar aquesta casella si a "Rango de entrada" hem seleccionat les cel·les on hi ha els rètols descriptius del nivell a què corresponen les dades de cada columna. Si no s'han posat aquests rètols descriptius o no s'han inclòs aquestes cel·les a "Rango de entrada", no hem de marcar la casella "Rótulos".
   d) **Alfa**. S'ha de posar el nivell d'error amb què volem treballar.
   e) **Opciones de salida**. Aquí triem on volem els resultats. Deixem marcada l'opció "En una hoja nueva".
5. El resultat apareix en una fulla nova. Hi ha dos quadres: quadre "Resumen" i quadre "Análisis de varianza".
6. Quadre **Resumen**. Obtenim, de cada nivell considerat, quantes dades hi ha, la seva suma, la seva mitjana i la seva variància.
7. Quadre **Análisis de varianza**. És el quadre amb el resultat dels càlculs que s'han de fer per obtenir l'estadístic de prova. El més interessant són les tres últimes columnes d'aquest quadre:
   a) **F**. És el valor de l'estadístic de prova F. En el nostre exemple, tenim que $F = 16.1$.
   b) **Probabilidad**. És el valor del nivell de significació crític o p-valor. En el nostre exemple, tenim que $\alpha_c = 3.86 \cdot 10^{-6}$.
   c) **Valor crítico para F**. És el valor de la taula F, amb els graus de llibertat corresponents, que fa de frontera entre acceptar la H_0 i la H_1. En el nostre exemple, tenim que $F_\alpha = 3.18$.

6.3 Comparació de variàncies: test de Levene

Per poder comprovar si podem assumir que les variàncies de diverses poblacions són iguals o no, hi ha diverses opcions, en forma de contrastos, que podem aplicar. En aquest cas comentarem el test de Levene, ja que és aplicable en condicions bastant generals i segueix el mateix procediment que l'ANOVA d'un factor, però aplicat a una transformació de les dades mostrals originals. A cada dada original cal restar-li la mitjana mostral del

seu nivell o població al qual pertany; el resultat d'aquesta resta s'ha d'agafar en valor absolut. Per tant, seguint amb la notació de l'apartat d'ANOVA amb un factor, tindrem:

Població o nivell			
1	2	...	k
$x'_{11} = \lvert x_{11} - \overline{x_{1\cdot}} \rvert$	$x'_{21} = \lvert x_{21} - \overline{x_{2\cdot}} \rvert$	...	$x'_{k1} = \lvert x_{k1} - \overline{x_{k\cdot}} \rvert$
.	.	...	.
.	.	...	.
.	.	...	.
$x'_{1n_1} = \lvert x_{1n_1} - \overline{x_{1\cdot}} \rvert$	$x'_{2n_2} = \lvert x_{2n_2} - \overline{x_{2\cdot}} \rvert$	...	$x'_{kn_k} = \lvert x_{kn_k} - \overline{x_{k\cdot}} \rvert$

Amb aquestes dades podem calcular una sèrie d'estadístics:

$$\overline{x'_{i\cdot}} = \frac{\sum_{j=1}^{n_i} x'_{ij}}{n_i} \quad \text{(mitjana mostral a la població o nivell } i\text{)}$$

$$n = \sum_{i=1}^{k} n_i \quad \text{(nombre total d'observacions)}$$

$$\overline{x'} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x'_{ij}}{n} \quad \text{(mitjana mostral general)}$$

Els passos que s'han de seguir per fer una comparació de variàncies segons el test de Levene són els mateixos que per fer una ANOVA d'un factor, excepte que ara es treballa amb les dades x'_{ij} i que a H_0 i H_1 es comparen variàncies en comptes de mitjanes. Aquestes hipòtesis ara són:

1. La hipòtesi nul·la, H_0, és:

   $$H_0: \sigma^2_1 = \ldots = \sigma^2_k$$

   Això vol dir que, sota H_0, podem assumir que les variàncies poblacionals són iguals entre les diferents poblacions o entre els diferents nivells del factor considerat.

2. La hipòtesi alternativa, H_1, és:

   $$H_1: \sigma^2_i \neq \sigma^2_j \text{ per a alguna parella } i \neq j$$

   H_1 significa que hi ha diferències entre almenys les variàncies poblacionals de dues poblacions o nivells.

Posant les dades adequadament, aquest contrast també es pot fer amb Excel usant ANOVA d'un factor.

Exemple

Seguint amb l'exemple de les hamburgueses i les seves calories, s'ha suposat que les variàncies poblacionals eren iguals en els 3 grups d'hamburgueses. Contrastem aquesta suposició, amb $\alpha = 0.05$.

Solució

1. Com que hem de fer el contrast per comparar variàncies, calcularem la mitjana de calories de cada grup d'hamburgueses. Obtenim:

$$\bar{x}_{vedella} = 156.85 \qquad \bar{x}_{pollastre} = 118.76 \qquad \bar{x}_{carn} = 158.71$$

2. A cada dada original li restem la mitjana del seu grup i agafem el resultat en valor absolut.

Vedella	*Pollastre*	*Carn*
29.15	10.24	14.29
24.15	13.24	32.29
19.15	16.76	23.29
7.85	12.76	31.29
27.15	24.76	13.29
33.15	16.76	11.71
1.15	31.76	12.71
17.85	19.76	19.71
18.15	11.76	16.29
8.85	5.76	22.71
4.85	16.24	20.29
45.85	23.24	5.71
15.85	32.76	51.71
3.85	24.24	36.29
33.15	33.24	23.71
0.15	27.24	18.71
25.85	25.24	20.71
7.85		
21.85		
24.85		

3. La H_0 és:

$$H_0: \sigma^2_{vedella} = \sigma^2_{pollastre} = \sigma^2_{carn}$$

Això vol dir que, sota H_0, no existeixen diferències entre les variàncies de les calories de les hamburgueses segons els diversos nivells considerats de composició.

4. La H_1 és:

$$H_1: \sigma^2_i \neq \sigma^2_j \text{ per a alguna parella } i \neq j$$

H_1 significa que hi ha diferències entre la variància de les calories d'almenys dos grups d'hamburgueses.

5. S'ha de consultar la taula F de Fisher amb (2,51) graus de llibertat. A la taula anterior hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $F_\alpha = 3.18$. Per tant, la zona de les taules on s'accepta la H_0 és l'interval (0, 3.18).
6. La zona de les taules on s'accepta la H_1 és l'interval (3.18, ∞).
7. Calculem l'estadístic de prova, F.

Font de variació	*g. l.*	*Suma quadrats*	*Quadrats mitjans*	*F*
Entre grups	2	113.42	56.71	0.49
Dintre grups	51	5904.58	115.78	
Total	53	6018		

8. Com que l'estadístic de prova $F = 0.49$ es troba a l'interval on s'accepta H_0, direm que acceptem H_0 i que, per tant, podem assumir que les variàncies de les calories de les diverses composicions d'hamburgueses són iguals.
9. El valor del nivell de significació crític o p-valor, α_c, és 0.6156.
10. Com que el p-valor, $\alpha_c = 0.6156$, és més gran que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_0 i arribem a la mateixa conclusió que abans.

6.4 Disseny ANOVA de dos factors sense interacció. Blocs aleatoritzats

Suposem que l'observació d'una v. a. X està influïda per dos factors. En l'exemple de la producció d'avellanes podem considerar el factor sòl i el factor que mesura la quantitat de pluja caiguda durant el mes de maig. En aquest cas podríem estudiar si hi ha diferències entre els diferents tipus de sòl o entre les diferents quantitats de pluja caigudes.

Suposem que el primer factor (factor fila) té a nivells i que el segon factor (factor columna) té b nivells. Suposem també que l'efecte dels dos factors és additiu i que disposem de $n = a \cdot b$ observacions, una per a cada combinació dels nivells del factor fila amb els nivells del factor columna. Amb aquestes condicions, les dades les podem posar de la manera següent:

		Factor columna			
		1	2	...	b
Factor fila	1	x_{11}	x_{12}	...	x_{1b}
	2	x_{21}	x_{22}	...	x_{2b}
	...	...	...	...	...
	a	x_{a1}	x_{a2}	...	x_{ab}

De la taula anterior podem calcular:

$\overline{x_{i\cdot}}$ (mitjana de la fila i, $1 \leq i \leq a$),

$\overline{x_{\cdot j}}$ (mitjana de la columna j, $1 \leq j \leq b$),

$\overline{x}$ (mitjana general).

Els passos que hem de seguir per fer una anàlisi de la variància de dos factors sense interacció serien:

1. Determinar quins són els factors que es volen considerar, quins nivells s'han agafat de cadascun d'aquests factors i quina és la variable resposta o variable observada que es vol estudiar.
2. Comprovar que les observacions de cada nivell de cada factor es distribueixen normalment i que hi ha la mateixa variància en els diversos nivells.
3. Determinar la hipòtesi nul·la, H_0. En aquest cas, hi haurà dues H_0, una per a cada factor, i seran:
   - H_0^f: $\alpha_1 = ... = \alpha_a$ (no hi ha efecte del factor fila)
   - H_0^c: $\beta_1 = ... = \beta_b$ (no hi ha efecte del factor columna)
4. Determinar la hipòtesi alternativa, H_1. En aquest cas, hi haurà dues H_1, una per a cada factor, i seran:
   - H_1^f: $\alpha_i \neq \alpha_j$ per a alguna parella $i \neq j$ (hi ha efecte fila)
   - H_1^c: $\beta_i \neq \beta_j$ per a alguna parella $i \neq j$ (hi ha efecte columna)

5. Determinar la zona de les taules on s'accepta la H_0. Hi haurà una zona per a cadascun dels dos contrastos que es volen realitzar. La taula que s'ha de consultar és la F de Fisher, però els graus de llibertat seran diferents segons el contrast que es vulgui realitzar:
   - Contrast del factor fila: $a-1$ i $(a-1)(b-1)$ graus de llibertat. Buscarem el valor F_α^f que deixa a la seva dreta una àrea igual a α i la zona de les taules on s'accepta la H_0^f és l'interval $(0, F_\alpha^f)$.
   - Contrast del factor columna: $b-1$ i $(a-1)(b-1)$ graus de llibertat. Buscarem el valor F_α^c que deixa a la seva dreta una àrea igual a α i la zona de les taules on s'accepta la H_0^c és l'interval $(0, F_\alpha^c)$.
6. Determinar la zona de les taules on s'accepta la H_1. Seran les zones complementàries a les trobades al pas 5), és a dir, seran:
   - Contrast del factor fila: (F_α^f, ∞).
   - Contrast del factor columna: (F_α^c, ∞).
7. Calcular els estadístics de prova. Per fer cadascun dels contrastos anteriors, un per al factor fila i un per al factor columna, necessitem uns estadístics diferents segons el contrast que vulguem fer. Els passos que s'han de seguir per calcular les F per a un disseny de dos factors (sense interacció) els podem resumir a la taula:

Font de variació	*g. l.*	*Suma quadrats*	*Quadrats mitjans*	*F*
Entre files	$a-1$	$Q_f = b\sum_i (\overline{x_{i\cdot}} - \overline{x})^2$	$\overline{Q_f} = \frac{Q_f}{a-1}$	$\frac{\overline{Q_f}}{\overline{Q_r}}$
Entre columnes	$b-1$	$Q_c = a\sum_j (\overline{x_{\cdot j}} - \overline{x})^2$	$\overline{Q_c} = \frac{Q_c}{b-1}$	$\frac{\overline{Q_c}}{\overline{Q_r}}$
Residu o error	$(a-1)(b-1)$	$Q_r = \sum_{i,j} (x_{ij} - \overline{x_{i\cdot}} - \overline{x_{\cdot j}} + \overline{x})^2$	$\overline{Q_r} = \frac{Q_r}{(a-1)\cdot(b-1)}$	
Total	$ab-1$	$Q_t = \sum_{i,j} (x_{ij} - \overline{x})^2$		

Per facilitar els càlculs, les sumes de quadrats es poden calcular amb fórmules algebraicament idèntiques:

$$Q_t = \sum_{i,j} x^2_{ij} - ab\,\bar{x}^{\,2}$$

$$Q_f = b \sum_{i} \bar{x}_{i\cdot}^{\,2} - ab\,\bar{x}^{\,2}$$

$$Q_c = a \sum_{j} \bar{x}_{\cdot j}^{\,2} - ab\,\bar{x}^{\,2}$$

$$Q_r = Q_t - Q_f - Q_c$$

8. A partir dels valors dels estadístics de prova *F* i les zones d'acceptació de cada hipòtesi, decidir quina és la hipòtesi certa per a cada contrast.
9. Calcular el valor del nivell de significació crític o p-valor de cada contrast: α_c^f i α_c^c. El p-valor de cada contrast es troba calculant la probabilitat que una distribució F de Fisher amb els graus de llibertat corresponents al contrast que es vol analitzar sigui més gran que l'estadístic de prova *F* del contrast corresponent, és a dir:
   - $P(F_{a-1,(a-1)(b-1)} > F^f) = \alpha_c^f$.
   - $P(F_{b-1,(a-1)(b-1)} > F^c) = \alpha_c^c$.
10. A partir dels valors de α_c^f, α_c^c i de α, decidir quina és la hipòtesi certa per a cada contrast.

L'anomenat disseny en **blocs aleatoritzats** és un disseny que s'usa especialment en l'experimentació agrícola, en el qual es volen comparar *a* tractaments (per exemple, *a* fertilitzants), assignant els tractaments en *b* blocs (exemple: *b* finques), de manera que es reparteixen els *a* tractaments aleatòriament a cada bloc (exemple: els fertilitzants s'apliquen aleatòriament en *a* parcel·les d'una mateixa finca). Interessarà saber si hi ha diferències entre els tractaments (α_i) i entre els blocs (β_j).

Blocs	A	2	1	3	4	5
	B	4	3	1	5	2
	C	5	1	2	4	3

Exemple

Es va fer un experiment per estudiar l'efecte de dos factors, altura d'escala i ritme de pujada d'escales, en el ritme cardíac de les persones. Es van considerar dues altures diferents d'escala, 14.6 cm i 29.2 cm, i tres ritmes diferents de pujada d'escales, 14 escales/minut, 21 escales/minut i 28 escales/minut. Per tant, hi ha 6 combinacions diferents dels diversos nivells considerats dels factors. Es van agafar 6 persones. Cada persona que va fer l'experiment va pujar les escales durant 3 minuts amb unes condicions particulars d'altura d'escala i ritme de pujada. La variable que es va mesurar és

la diferència entre el ritme cardíac abans de fer l'activitat i el que tenia després de fer l'activitat. Els resultats van ser:

		Ritme pujada		
		14	21	28
Altura escala	14.6	9	15	24
	29.2	16	26	50

Suposant que l'increment de ritme cardíac de cada nivell dels factors considerats es distribueix normalment i que hi ha la mateixa variància en els diversos nivells, es vol analitzar si hi ha diferències significatives, amb $\alpha = 0.05$, entre l'increment del ritme cardíac dels diversos nivells considerats dels dos factors que es volen estudiar.

Solució

1. Els factors que es volen considerar són l'altura d'escala i el ritme de pujada d'escales. Del factor altura d'escala es consideren dos nivells diferents: 14.6 cm i 29.2 cm. Del factor ritme de pujada d'escales es consideren tres nivells diferents: 14 escales/minut, 21 escales/minut i 28 escales/minut. La variable resposta que es vol estudiar és l'increment de ritme cardíac després de fer l'activitat.
2. Hem suposat que es compleixen les condicions per aplicar ANOVA de dos factors: les observacions de cada nivell del factor es distribueixen normalment i hi ha la mateixa variància en els diversos nivells.
3. Hi ha dues H_0, una per a cada factor, i són:
   - H_0^f: $\alpha_{14.6} = \alpha_{29.2}$ (no hi ha efecte del factor altura de l'escala)
   - H_0^c: $\beta_{14} = \beta_{21} = \beta_{28}$ (no hi ha efecte del factor ritme de pujada d'escales)
4. Hi ha dues H_1, una per a cada factor, i són:
   - H_1^f: $\alpha_{14.6} \neq \alpha_{29.2}$ (hi ha efecte de l'altura de l'escala en l'increment del ritme cardíac).
   - H_1^c: $\beta_i \neq \beta_j$ per a alguna parella $i \neq j$ (hi ha efecte del ritme de pujada d'escales en l'increment del ritme cardíac).
5. Determinem la zona de les taules on s'accepta la H_0 de cada contrast:
   - Contrast del factor altura d'escala: s'ha de consultar la taula F de Fisher amb (1,2) graus de llibertat. A la taula anterior hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $F_\alpha^f = 18.5$. Per tant, la zona de les taules on s'accepta la H_0^f és l'interval (0, 18.5).

- Contrast del factor ritme de pujada: s'ha de consultar la taula F de Fisher amb (2,2) graus de llibertat. A la taula anterior hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $F_{\alpha}^{c} = 19.0$. Per tant, la zona de les taules on s'accepta la H_0^c és l'interval (0, 19.0).

6. Determinem la zona de les taules on s'accepta la H_1 de cada contrast:
   - Contrast del factor altura d'escala: la zona de les taules on s'accepta la H_1^f és l'interval (18.5, ∞).
   - Contrast del factor ritme de pujada: la zona de les taules on s'accepta la H_1^c és l'interval (19.0, ∞).
7. Calculem els estadístics de prova, un per al factor altura d'escala i un per al factor ritme de pujada d'escales:

Font de variació	*g. l.*	*Suma quadrats*	*Quadrats mitjans*	*F*
Altura escala	1	322.7	322.7	6.43
Ritme pujada	2	624.3	312.2	6.22
Residu o error	2	100.3	50.2	
Total	5	1047.3		

8. Hi haurà una conclusió per a cada contrast:
   - Contrast del factor altura d'escala: com que l'estadístic de prova $F^f = 6.43$ es troba a l'interval on s'accepta H_0^f, direm que acceptem H_0^f i que, per tant, no s'ha demostrat que l'increment mitjà del ritme cardíac sigui diferent per a les dues altures d'escala considerades.
   - Contrast del factor ritme de pujada: com que l'estadístic de prova $F^c = 6.22$ es troba a l'interval on s'accepta H_0^c, direm que acceptem H_0^c i que, per tant, no s'ha demostrat que l'increment mitjà del ritme cardíac sigui diferent per als tres ritmes de pujada d'escales considerats.
9. Hi ha un p-valor per a cada contrast:
   - Contrast del factor altura d'escala: el valor del nivell de significació crític o p-valor, α_c^f, és 0.127.
   - Contrast del factor ritme de pujada: el valor del nivell de significació crític o p-valor, α_c^c, és 0.138.
10. Segons el p-valor, hi haurà una conclusió per a cada contrast:
    - Contrast del factor altura d'escala: com que el p-valor corresponent, $\alpha_c^f = 0.127$, és més gran que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_0^f i arribem a la mateixa conclusió que abans.

- Contrast del factor ritme de pujada: com que el p-valor corresponent, $\alpha_c^c = 0.138$, és més gran que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_0^c i arribem a la mateixa conclusió que abans.

6.4.1 Excel: ANOVA de dos factors amb una sola mostra per grup

Per fer una ANOVA de dos factors amb una sola mostra per grup amb Excel cal seguir els passos següents:

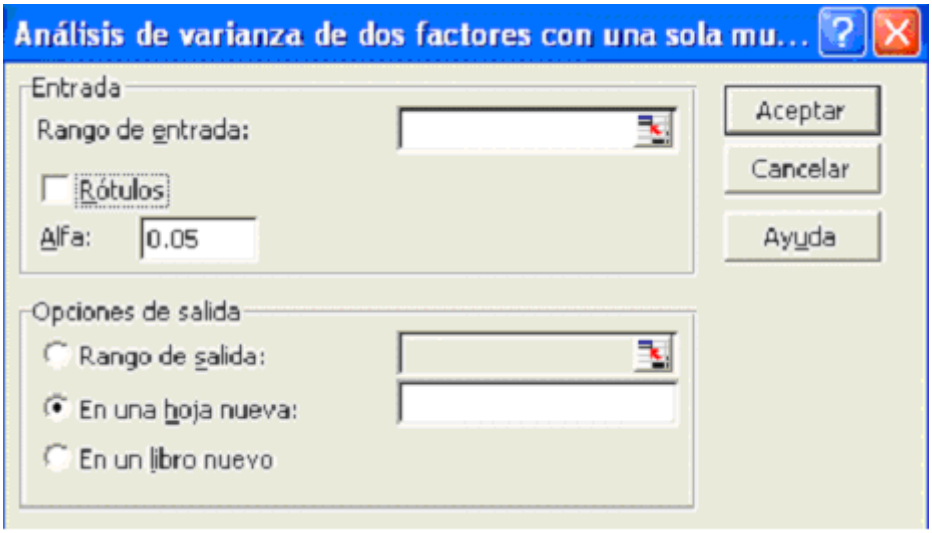
1. Hem d'introduir les dades en el full de càlcul. Si seguim amb l'exemple de la pujada d'escales, hem de posar les dades d'un factor en files i les de l'altre factor en columnes. També podem posar un rètol que indiqui a quin nivell corresponen les dades de cada fila i de cada columna. La pantalla 3 ens mostra com han de quedar les dades una vegada introduïdes a Excel.

Pantalla 3. Dades de dos factors amb una sola mostra per grup

	A	B	C	D
1		14	21	28
2	14.6	9	15	24
3	29.2	16	26	50

2. Del menú "Herramientas", triem "Análisis de datos".
3. De les funcions que apareixen, triem "Análisis de varianza de dos factores con una sola muestra por grupo" i acceptem. Ha d'aparèixer el quadre que es veu a la pantalla 4.

Pantalla 4. Quadre de "Análisis de varianza de dos factores con una sola muestra por grupo"



4. Del quadre anterior, hem d'emplenar:
   a) **Rango de entrada**. Hem de seleccionar les caselles on es troben les dades que volem analitzar.
   b) **Rótulos**. Hem de marcar aquesta casella si a "Rango de entrada" hem seleccionat les cel·les on hi ha els rètols descriptius dels nivells a què corresponen les dades de cada fila i de cada columna. Si no s'han posat aquests rètols descriptius o no s'han inclòs aquestes cel·les a "Rango de entrada", no hem de marcar la casella "Rótulos".
   c) **Alfa**. Hem de posar el nivell d'error amb què volem treballar.
   d) **Opciones de salida**. Aquí triem on volem els resultats. Deixem marcada l'opció "En una hoja nueva".
5. El resultat apareix en una fulla nova. Hi ha dos quadres: quadre "Resumen" i quadre "Análisis de varianza".
6. Quadre **Resumen**. Obtenim, de cada nivell considerat de cada factor, quantes dades hi ha, la seva suma, la seva mitjana i la seva variància.
7. Quadre **Análisis de varianza**. És el quadre amb el resultat dels càlculs que s'han de fer per obtenir els estadístics de prova. El més interessant són les tres últimes columnes d'aquest quadre:
   a) **F**. Són els valors dels estadístics de prova F^f i F^c. En el nostre exemple tenim que $F^f = 6.43$ i $F^c = 6.22$.
   b) **Probabilidad**. Són els valors del nivell de significació crític o p-valor de cada contrast. En el nostre exemple tenim $\alpha_c^f = 0.127$ i $\alpha_c^c = 0.138$.
   c) **Valor crítico para F**. Són els valors de la taula F, amb els graus de llibertat corresponents, que fan de frontera entre acceptar la H_0 i la H_1 de cada contrast. En el nostre exemple tenim que $F_\alpha^f = 18.5$ i $F_\alpha^c = 19.0$.

6.5 Disseny ANOVA de dos factors amb interacció

Suposem que l'observació d'una v. a. X està influïda per dos factors i que el nombre d'observacions per a cada combinació dels nivells del factor fila amb els nivells del factor columna és més gran que 1.

Per tant, suposem que el primer factor (factor fila) té *a* nivells i que el segon factor (factor columna) té *b* nivells. Suposem que disposem de $n = a \times b \times r$ observacions, *r* per a cada combinació dels nivells del factor fila amb els nivells del factor columna. Amb aquestes condicions, les dades les podem posar de la manera següent:

		Factor columna			
		1	2	...	b
Factor fila	1	$x_{111} \dots x_{11r}$	$x_{121} \dots x_{12r}$	...	$x_{1b1} \dots x_{1br}$
	2	$x_{211} \dots x_{21r}$	$x_{221} \dots x_{22r}$	...	$x_{2b1} \dots x_{2br}$
	...	...	...	...	...
	a	$x_{a11} \dots x_{a1r}$	$x_{a21} \dots x_{a2r}$	...	$x_{ab1} \dots x_{abr}$

i podem calcular:

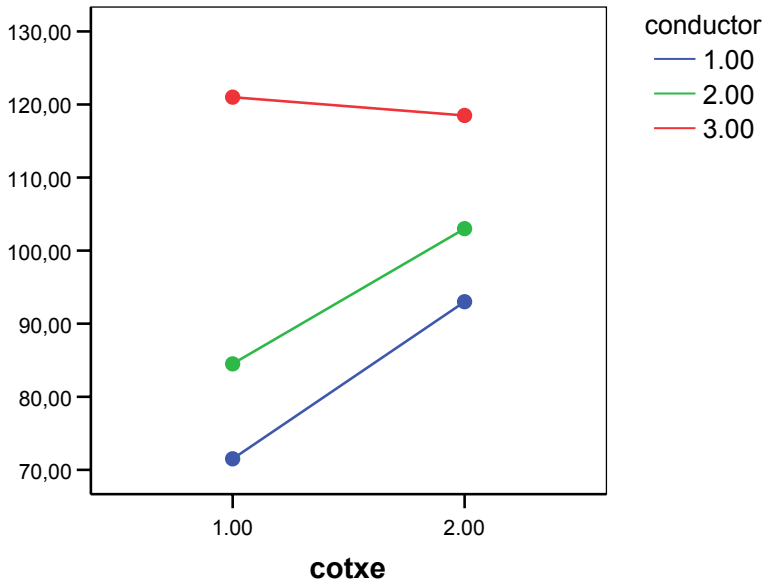
$\overline{x_{i\cdot}}$ (mitjana de la fila i, $1 \leq i \leq a$),

$\overline{x_{\cdot j}}$ (mitjana de la columna j, $1 \leq j \leq b$),

$\overline{x_{ij}}$ (mitjana de la combinació de la fila i amb la columna j, $1 \leq i \leq a$, $1 \leq j \leq b$),

$\overline{x}$ (mitjana general).

Amb aquestes condicions ens apareix un nou element, que és la interacció que hi ha entre els factors considerats. Per explicar el significat de la interacció ho farem a través d'un exemple: suposem que agafem el temps que es triga a fer una volta al circuit de Montmeló en dos cotxes diferents i amb tres conductors diferents. Si no hi hagués interacció entre cotxes i conductors, amb cotxes diferents el pilot més ràpid sempre tardaria menys temps a fer una volta que els altres pilots o, si més no, la diferència amb els altres pilots es mantindria en canviar de cotxe. Però pot passar que un pilot diferent estigui molt avesat a un tipus de cotxe en concret i amb aquest cotxe sigui el més ràpid o la diferència no sigui tan gran com en l'altre cotxe. D'aquesta manera es veu que hi pot haver una interacció entre cotxes i pilots. Gràficament ho podem representar com a:



Si no hi hagués interacció entre cotxes i pilots, les tres línies serien més o menys paral·leles, la qual cosa voldria dir que les diferències de temps es mantenen entre els conductors sigui quin sigui el cotxe que agafin. Però en el gràfic anterior podem veure que la diferència de temps entre el 3r conductor i els dos primers no és la mateixa amb un cotxe que amb l'altre; això indica que hi ha certa interacció entre cotxes i conductors.

En el cas que tinguem dos factors amb diverses mostres per a cada combinació de nivells dels factors, es poden fer 3 contrastos:

- Si hi ha diferències entre els nivells del factor fila.
- Si hi ha diferències entre els nivells del factor columna.
- Si hi ha interacció entre els dos factors considerats.

Els passos que hem de seguir per fer una anàlisi de la variància de dos factors amb interacció són:

1. Determinar quins són els factors que es volen considerar, quins nivells s'han agafat de cadascun d'aquests factors i quina és la variable resposta o variable observada que es vol estudiar.
2. Comprovar que les observacions de cada nivell de cada factor es distribueixen normalment i que hi ha la mateixa variància en els diversos nivells.
3. Determinar la hipòtesi nul·la, H_0. En aquest cas hi haurà tres H_0, una per a cada factor i una per contrastar si hi ha interacció entre els dos factors, i seran:

   H_0^f: $\alpha_1 = \ldots = \alpha_a$ (no hi ha efecte del factor fila)

   H_0^c: $\beta_1 = \ldots = \beta_b$ (no hi ha efecte del factor columna)

   H_0^I: no hi ha interacció entre el factor fila i el factor columna
4. Determinar la hipòtesi alternativa, H_1. En aquest cas, hi haurà tres H_1, una per a cada factor i una per contrastar si hi ha interacció entre els dos factors, i seran:

   H_1^f: $\alpha_i \neq \alpha_j$ per a alguna parella $i \neq j$ (hi ha efecte fila)

   H_1^c: $\beta_i \neq \beta_j$ per a alguna parella $i \neq j$ (hi ha efecte columna)

   H_1^I: hi ha interacció entre el factor fila i el factor columna
5. Determinar la zona de les taules on s'accepta la H_0. Hi haurà una zona per a cadascun dels tres contrastos que es volen realitzar. La taula que s'ha de consultar és la F de Fisher, però els graus de llibertat seran diferents segons el contrast que es vulgui realitzar:
   - Contrast del factor fila: $a - 1$ i $ab(r - 1)$ graus de llibertat. Buscarem el valor F_α^f que deixa a la seva dreta una àrea igual a α i la zona de les taules on s'accepta la H_0^f és l'interval $(0, F_\alpha^f)$.

- Contrast del factor columna: $b - 1$ i $ab\,(r - 1)$ graus de llibertat. Buscarem el valor $F_\alpha^{\,c}$ que deixa a la seva dreta una àrea igual a α i la zona de les taules on s'accepta la $H_0^{\,c}$ és l'interval $(0\,,\,F_\alpha^{\,c})$.
- Contrast de la interacció: $(a - 1)(b - 1)$ i $ab(r - 1)$ graus de llibertat. Buscarem el valor $F_\alpha^{\,I}$, que deixa a la seva dreta una àrea igual a α i la zona de les taules on s'accepta la $H_0^{\,I}$ és l'interval $(0\,,\,F_\alpha^{\,I})$.

6. Determinar la zona de les taules on s'accepta la H_1. Seran les zones complementàries a les trobades al pas 5), és a dir, seran:
   - Contrast del factor fila: $(F_\alpha^{\,f}\,,\,\infty)$.
   - Contrast del factor columna: $(F_\alpha^{\,c}\,,\,\infty)$.
   - Contrast de la interacció: $(F_\alpha^{\,I}\,,\,\infty)$.
7. Calcular els estadístics de prova. Per fer cadascun dels contrastos anteriors necessitem uns estadístics diferents segons el contrast que vulguem fer. Els passos que hem de seguir per calcular les F per a un disseny de dos factors (amb interacció) els podem resumir a la taula:

Font de variació	*g. l.*	*Suma quadrats*	*Quadrats mitjans*	*F*
Entre files	$a - 1$	$Q_f = br \sum_i (\overline{x_{i\cdot}} - \overline{x})^2$	$\overline{Q_f} = \dfrac{Q_f}{a\text{-}1}$	$\dfrac{\overline{Q_f}}{\overline{Q_r}}$
Entre columnes	$b - 1$	$Q_c = ar \sum_j (\overline{x_{\cdot j}} - \overline{x})^2$	$\overline{Q_c} = \dfrac{Q_c}{b\text{-}1}$	$\dfrac{\overline{Q_c}}{\overline{Q_r}}$
Interacció	$(a - 1)(b - 1)$	$Q_I = r \sum_{i,j} (\overline{x_{ij}} - \overline{x_{i\cdot}} - \overline{x_{\cdot j}} + \overline{x})^2$	$\overline{Q_I} = \dfrac{Q_I}{(a\text{-}1)\cdot(b\text{-}1)}$	$\dfrac{\overline{Q_I}}{\overline{Q_r}}$
Residu o error	$ab(r - 1)$	$Q_r = \sum_{i,j,k} (x_{ijk} - \overline{x_{ij}})^2$	$\overline{Q_r} = \dfrac{Q_r}{ab(r\text{-}1)}$	
Total	$abr - 1$	$Q_t = \sum_{i,j,k} (x_{ijk} - \overline{x})^2$		

Per facilitar els càlculs, les sumes de quadrats es poden calcular amb fórmules algebraicament idèntiques:

$$Q_t = \sum_{i,j,k} x^2_{ijk} - abr\,\bar{x}^{\,2}$$

$$Q_f = br \sum_{i} \overline{x_{i\cdot}}^{\,2} - abr\,\bar{x}^{\,2}$$

$$Q_c = ar \sum_{j} \overline{x_{\cdot j}}^{\,2} - abr\,\bar{x}^{\,2}$$

$$Q_I = r \sum_{i,j} \overline{x_{ij}}^{\,2} - br \sum_{i} \overline{x_{i\cdot}}^{\,2} - ar \sum_{j} \overline{x_{\cdot j}}^{\,2} + abr\,\bar{x}^{\,2}$$

$$Q_r = Q_t - Q_f - Q_c - Q_I$$

8. A partir dels valors dels estadístics de prova F i les zones d'acceptació de cada hipòtesi, decidir quina és la hipòtesi certa per a cada contrast.
9. Calcular el valor del nivell de significació crític o p-valor de cada contrast: α_c^f, α_c^c i α_c^I. El p-valor de cada contrast es troba calculant la probabilitat que una distribució F de Fisher amb els graus de llibertat corresponents al contrast que es vol analitzar sigui més gran que l'estadístic de prova F del contrast corresponent, és a dir:
   - $P(F_{a-1,ab(r-1)} > F^f) = \alpha_c^f$
   - $P(F_{b-1,ab(r-1)} > F^c) = \alpha_c^c$
   - $P(F_{(a-1)(b-1),ab(r-1)} > F^I) = \alpha_c^I$
10. A partir dels valors de α_c^f, α_c^c i α_c^I i de α, decidir quina és la hipòtesi certa per a cada contrast.

Exemple

Es va fer un experiment per estudiar l'efecte de dos factors, altura d'escala i ritme de pujada d'escales, en el ritme cardíac de les persones. Es van considerar dues altures diferents d'escala, 14.6 cm i 29.2 cm, i tres ritmes diferents de pujada d'escales, 14 escales/minut, 21 escales/minut i 28 escales/minut. Per tant, hi ha 6 combinacions diferents dels diversos nivells considerats dels factors. Es van agafar 5 persones per a cada combinació. La variable que es va mesurar és la diferència entre el ritme cardíac abans de fer l'activitat i el que tenia després de fer l'activitat. Els resultats van ser:

		Ritme pujada		
		14	21	28
Altura escala	14.6	10 15 14 6 0	10 22 20 14 9	15 24 22 39 20
	29.2	11 22 6 33 8	14 30 45 35 6	66 51 37 63 33

Suposant que l'increment de ritme cardíac de cada nivell dels factors considerats es distribueix normalment i que hi ha la mateixa variància en els diversos nivells, es vol analitzar si hi ha diferències significatives, amb $\alpha = 0.05$, entre l'increment del ritme cardíac dels diversos nivells considerats dels dos factors que es volen estudiar.

Solució

1. Els factors que es volen considerar són l'altura d'escala i el ritme de pujada d'escales. Del factor altura d'escala es consideren dos nivells diferents: 14.6 cm i 29.2 cm. Del factor ritme de pujada d'escales es consideren tres nivells diferents: 14 escales/minut, 21 escales/minut i 28 escales/minut. La variable resposta que es vol estudiar és l'increment de ritme cardíac després de fer l'activitat.
2. Hem suposat que es compleixen les condicions per aplicar ANOVA de dos factors: les observacions de cada nivell del factor es distribueixen normalment i hi ha la mateixa variància en els diversos nivells.
3. Hi ha tres H_0, una per a cada factor i una per contrastar si hi ha interacció entre els dos factors, i són:
   - H_0^f: $\alpha_{14.6} = \alpha_{29.2}$ (no hi ha efecte del factor altura de l'escala)
   - H_0^c: $\beta_{14} = \beta_{21} = \beta_{28}$ (no hi ha efecte del factor ritme de pujada d'escales)
   - H_0^I: no hi ha interacció entre l'altura de l'escala i el ritme de pujada
4. Hi ha tres H_1, una per a cada factor i una per contrastar si hi ha interacció entre els dos factors, i són:
   - H_1^f: $\alpha_{14.6} \neq \alpha_{29.2}$ (hi ha efecte de l'altura de l'escala en l'increment del ritme cardíac)
   - H_1^c: $\beta_i \neq \beta_j$ per a alguna parella $i \neq j$ (hi ha efecte del ritme de pujada d'escales en l'increment del ritme cardíac)
   - H_1^I: hi ha interacció entre l'altura de l'escala i el ritme de pujada
5. Determinem la zona de les taules on s'accepta la H_0 de cada contrast:
   - Contrast del factor altura d'escala: s'ha de consultar la taula F de Fisher amb (1,24) graus de llibertat. A la taula anterior, hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $F_\alpha^f = 4.26$. Per tant, la zona de les taules on s'accepta la H_0^f és l'interval (0, 4.26).
   - Contrast del factor ritme de pujada: s'ha de consultar la taula F de Fisher amb (2,24) graus de llibertat. A la taula anterior hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $F_\alpha^c = 3.40$. Per tant, la zona de les taules on s'accepta la H_0^c és l'interval (0, 3.40).

- Contrast de la interacció: s'ha de consultar la taula F de Fisher amb (2,24) graus de llibertat. A la taula anterior, hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $F_{\alpha}^{I} = 3.40$. Per tant, la zona de les taules on s'accepta la H_0^I és l'interval (0, 3.40).

6. Determinem la zona de les taules on s'accepta la H_1 de cada contrast:
   - Contrast del factor altura d'escala: la zona de les taules on s'accepta la H_1^f és l'interval (4.26, ∞).
   - Contrast del factor ritme de pujada: la zona de les taules on s'accepta la H_1^c és l'interval (3.40, ∞).
   - Contrast de la interacció: la zona de les taules on s'accepta la H_1^I és l'interval (3.40, ∞).
7. Calculem els estadístics de prova, un per al factor altura d'escala, un per al factor ritme de pujada d'escales i un per a la interacció entre l'altura de l'escala i el ritme de pujada:

Font de variació	*g. l.*	*Suma quadrats*	*Quadrats mitjans*	*F*
Altura escala	1	1613.33	1613.33	12.85
Ritme pujada	2	3121.67	1560.83	12.43
Interacció	2	501.67	250.83	2.00
Residu o error	24	3014	125.58	
Total	29	8250.67		

8. Hi haurà una conclusió per a cada contrast:
   - Contrast del factor altura d'escala: com que l'estadístic de prova $F^f = 12.85$ es troba a l'interval on s'accepta H_1^f, direm que acceptem H_1^f i que, per tant, l'increment mitjà del ritme cardíac no és el mateix per a les dues altures d'escala considerades.
   - Contrast del factor ritme de pujada: com que l'estadístic de prova $F^c = 12.43$ es troba a l'interval on s'accepta H_1^c, direm que acceptem H_1^c i que, per tant, l'increment mitjà del ritme cardíac no és el mateix per als tres ritmes de pujada d'escales considerats.
   - Contrast de la interacció: com que l'estadístic de prova $F^I = 2.00$ es troba a l'interval on s'accepta H_0^I, direm que acceptem H_0^I i que, per tant, no s'ha demostrat que la interacció entre l'altura de l'escala i el ritme de pujada sigui significativa.

9. Hi ha un p-valor per a cada contrast:
   - Contrast del factor altura d'escala: el valor del nivell de significació crític o p-valor, α_c^f, és 0.0015.
   - Contrast del factor ritme de pujada: el valor del nivell de significació crític o p-valor, α_c^c, és 0.0002.
   - Contrast de la interacció: el valor del nivell de significació crític o p-valor, α_c^I, és 0.1576.
10. Segons el p-valor, hi haurà una conclusió per a cada contrast:
   - Contrast del factor altura d'escala: com que el p-valor corresponent, $\alpha_c^f = 0.0015$, és més petit que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_1^f i arribem a la mateixa conclusió que abans.
   - Contrast del factor ritme de pujada: com que el p-valor corresponent, $\alpha_c^c = 0.0002$, és més petit que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_1^c i arribem a la mateixa conclusió que abans.
   - Contrast de la interacció: com que el p-valor corresponent, $\alpha_c^I = 0.1576$, és més gran que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_0^I i arribem a la mateixa conclusió que abans.

6.5.1 Excel: ANOVA de dos factors amb diverses mostres per grup

Per fer una ANOVA de dos factors amb diverses mostres per grup amb Excel cal seguir els passos següents:

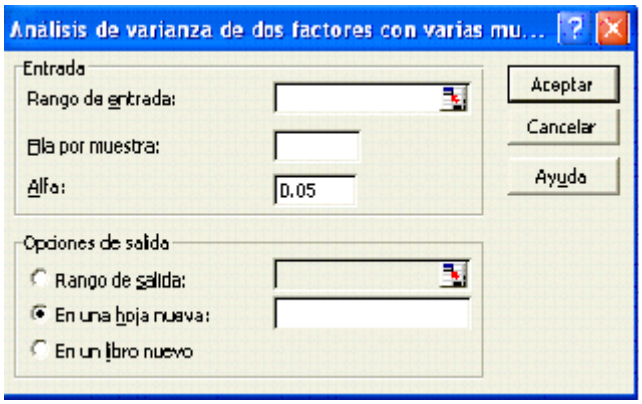
1. Hem d'introduir les dades en el full de càlcul. Si seguim amb l'exemple de la pujada d'escales, hem de posar les dades d'un factor en files i les de l'altre factor en columnes, però per a cada nivell del factor fila hem de reservar tantes files com observacions hi hagi a cada combinació de nivells dels dos factors considerats (en el nostre cas, 5). També podem posar un rètol que indiqui a quin nivell corresponen les dades de cada fila i de cada columna. La pantalla 5 ens mostra com han de quedar les dades una vegada introduïdes a Excel.

Pantalla 5. Dades Excel de dos factors amb diverses mostres per grup

	A	B	C	D
1		14	21	28
2	14.6	10	10	15
3		15	22	24
4		14	20	22
5		6	14	39
6		0	9	20
7	29.2	11	14	66
8		22	30	51
9		6	45	37
10		33	35	63
11		8	6	33

2. Del menú "Herramientas", triem "Análisis de datos".
3. De les funcions que apareixen, triem "Análisis de varianza de dos factores con varias muestras por grupo" i acceptem. Ha d'aparèixer el quadre que es veu a la pantalla 6.

Pantalla 6. Quadre de "Análisis de varianza de dos factores con varias muestras por grupo"

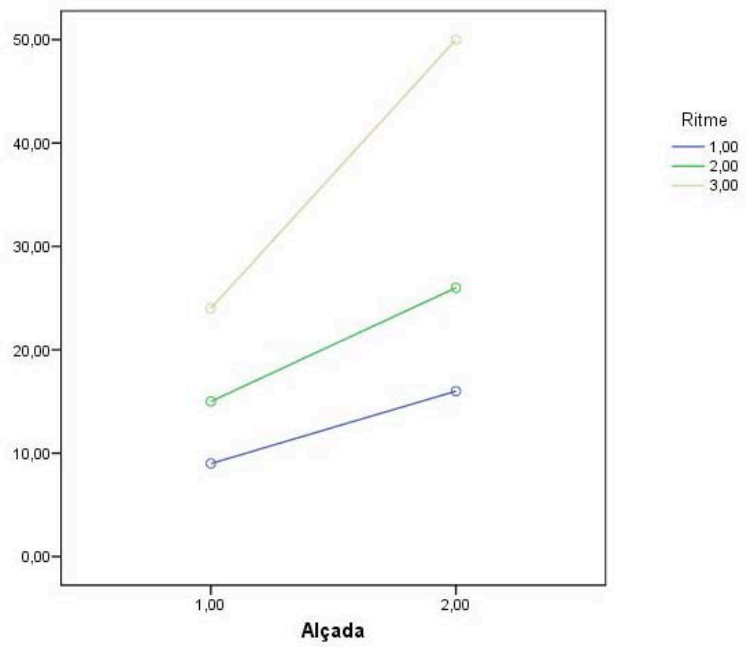


4. Del quadre anterior, hem d'emplenar:
   a) *Rango de entrada*. Hem de seleccionar les caselles on es troben les dades que volem analitzar. En aquest cas, és obligat posar i agafar els rètols dels nivells dels dos factors.
   b) *Fila por muestra*. Hem de posar quantes dades hi ha a cada combinació de nivells dels dos factors. Si seguim el nostre exemple, hem de posar 5.
   c) *Alfa*. S'ha de posar el nivell d'error amb què volem treballar.
   d) *Opciones de salida*. Aquí triem on volem els resultats. Deixem marcada l'opció "En una hoja nueva".

5. El resultat apareix en una fulla nova. Hi ha dos quadres: quadre "Resumen" i quadre "Análisis de varianza".
6. Quadre **Resumen**. Obtenim, de cada nivell considerat de cada factor, quantes dades hi ha, la seva suma, la seva mitjana i la seva variància.
7. Quadre **Análisis de varianza**. És el quadre amb el resultat dels càlculs que s'han de fer per obtenir els estadístics de prova. El més interessant són les tres últimes columnes d'aquest quadre:
   a) *F*. Són els valors dels estadístics de prova F^f, F^c i F^I. En el nostre exemple tenim que $F^f = 12.85$, $F^c = 12.43$ i $F^I = 2.00$.
   b) *Probabilidad*. Són els valors del nivell de significació crític o p-valor de cada contrast. En el nostre exemple, tenim $\alpha_c^f = 0.0015$, $\alpha_c^c = 0.0002$ i $\alpha_c^I = 0.1576$.
   c) *Valor crítico para F*. Són els valors de la taula *F*, amb els graus de llibertat corresponents, que fan de frontera entre acceptar la H_0 i la H_1 de cada contrast. En el nostre exemple tenim que $F_\alpha^f = 4.26$, $F_\alpha^c = 3.40$ i $F_\alpha^I = 3.40$.

Observacions

1. En aquest exemple de dos factors amb interacció, les mitjanes de cada combinació d'altura d'escala i ritme de pujada d'escales coincideix amb els valors de l'exemple de dos factors amb una sola mostra per a cada combinació de nivells dels dos factors considerats. Abans s'obtenia que no s'havia demostrat que l'altura de les escales o que el ritme de pujada de les escales provoqués diferències en l'increment del ritme cardíac i ara s'obté que sí que hi ha diferències significatives en l'increment del ritme cardíac, tant en canviar l'altura de les escales com en canviar el ritme de pujada de les escales. A què pot ser degut aquest canvi en les conclusions?
2. Per comprovar la força de la interacció, es pot fer un gràfic de les mitjanes de les diverses combinacions dels nivells dels dos factors. Realment es veu que hi ha una certa interacció, encara que numèricament s'obté que aquesta no està demostrada de manera clara.


50,00
40,00
30,00
20,00
10,00
0,00
1,00
2,00
Alçada
Ritme
1,00
2,00
3,00

7. Proves d'independència i de bondat d'ajustament

Les proves per contrastar si dues característiques observades d'una població són independents i per contrastar si les dades recollides en una mostra segueixen una certa distribució estadística tenen un punt en comú: es pot usar l'anomenada prova khi quadrat per fer cadascun d'aquests contrastos.

Aquesta prova es basa a comparar les freqüències que s'han observat en recollir les dades de les mostres amb les freqüències que s'esperarien si fos certa la hipòtesi nul·la que es vol contrastar. L'estadístic que s'ha de calcular per aplicar la prova khi quadrat és:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\textit{freqüència observada} - \textit{freqüència esperada})^2}{\textit{freqüència esperada}}$$

on k és el nombre de successos en què s'ha dividit cert espai mostral. Aquests successos han de formar una partició de l'espai mostral, és a dir, han de ser disjunts i la seva unió ha de coincidir amb l'espai mostral.

Si les freqüències observades s'assemblen a les freqüències esperades segons la hipòtesi nul·la, llavors és raonable pensar que la hipòtesi nul·la és certa. En canvi, si les freqüències observades són molt diferents de les freqüències esperades segons la hipòtesi nul·la, llavors sembla lògic pensar que la hipòtesi nul·la no és certa.

Exemple. Suposem que tenim un dau i volem comprovar si està trucat o no. Per comprovar-ho, llancem el dau 300 vegades i obtenim els resultats següents:

Resultat	1	2	3	4	5	6
Freq. observada	45	52	56	48	53	46

Si el dau no està trucat, esperaríem que la freqüència de cada puntuació del dau fos 50. Veiem que les freqüències observades en fer els 300 llançaments són semblants a les que s'esperarien si el dau no estigués trucat. Per tant, podem afirmar que el dau no està trucat.

Exemple. Suposem que agafem un altre dau i volem comprovar si està trucat o no. Per fer la comprovació, llancem aquest dau 300 vegades i obtenim els resultats següents:

Resultat	1	2	3	4	5	6
Freq. observada	20	90	10	80	15	85

En aquest cas, veiem que les freqüències observades en fer els 300 llançaments són bastant diferents de les que s'esperarien si el dau no estigués trucat. Per tant, podem afirmar que el dau està trucat.

7.1 Prova d'independència

Suposem que volem determinar si existeix alguna relació entre dues característiques diferents en les quals una població ha estat classificada i on cada característica es troba dividida en cert nombre de categories. Per exemple, existeix alguna relació entre l'edat de les persones i el color del seu cotxe? En aquest exemple, s'ha classificat la població en dues característiques, on suposem que cadascuna té almenys dues categories exhaustives i mútuament excloents. Aquestes dues característiques són la franja d'edat a la qual pertany una persona i el color del seu cotxe. Les categories per a aquestes dues característiques podrien ser:

- Per a l'edat de cada persona: té entre 18 i 30 anys, entre 30 i 45 anys o és major de 45 anys.
- Per al color del seu cotxe: vermell, blau, negre o gris.

Generalitzant, suposem que tenim una població Ω que admet dues descomposicions diferents en categories excloents:

$$\Omega = A_1 + \ldots + A_r = B_1 + \ldots + B_s$$

Suposem que tenim una mostra de n elements, de manera que:

$$n_{ij} \text{ és la freqüència absoluta del succés } A_i \cap B_j$$

Les **freqüències observades** de la mostra es posen en una **taula de contingència** $r \times s$. Una taula de contingència es forma per les freqüències que s'observen per a les dues classificacions i les seves categories corresponents.

	B_1	B_2	...	B_s	
A_1	n_{11}	n_{12}	...	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	...	n_{2s}	$n_{2\cdot}$
.					.
.	...	...	...	...	.
.					.
A_r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot s}$	n

$n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$ és la freqüència absoluta de A_i (total de la fila i).

$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$ és la freqüència absoluta de B_j (total de la columna j).

L'**objectiu** és contrastar si aquestes dues característiques en què s'ha dividit la població són independents entre elles o, en canvi, si hi ha relació.

Els passos que cal seguir per contrastar si dues característiques són independents o no són:

1. Determinar quines són les dues característiques, A i B, en què s'ha dividit la població i que són objecte d'estudi. També s'ha de precisar en quines categories, A_i i B_j, s'ha subdividit cada característica.
2. Determinar la hipòtesi nul·la, H_0. La H_0 sempre serà la mateixa:

   $$H_0: P(A_i \cap B_j) = P(A_i)P(B_j) \qquad \forall\ i,j$$

   Això vol dir que, si es compleix H_0, les característiques A i B són independents i que no hi ha relació entre A i B.
3. Determinar la hipòtesi alternativa, H_1. La H_1 sempre serà la mateixa:

   $$H_1: P(A_i \cap B_j) \neq P(A_i)P(B_j) \qquad \text{per a alguna parella } i,j$$

   H_1 significa que A i B són dependents i que hi ha relació entre les categories de les característiques A i B.
4. Construir la taula de **freqüències esperades** sota la suposició que H_0 és certa. En aquest cas, la freqüència esperada de la cel·la $A_i \cap B_j$, sota H_0, és:

   $$n_{i\cdot} n_{\cdot j} / n$$

La taula de freqüències esperades seria:

	B_1	B_2	...	B_s	
A_1	$n_{1\cdot}n_{\cdot 1}/n$	$n_{1\cdot}n_{\cdot 2}/n$	...	$n_{1\cdot}n_{\cdot s}/n$	$n_{1\cdot}$
A_2	$n_{2\cdot}n_{\cdot 1}/n$	$n_{2\cdot}n_{\cdot 2}/n$	...	$n_{2\cdot}n_{\cdot s}/n$	$n_{2\cdot}$
.					.
.	...	...	...	...	.
.					.
A_r	$n_{r\cdot}n_{\cdot 1}/n$	$n_{r\cdot}n_{\cdot 2}/n$	...	$n_{r\cdot}n_{\cdot s}/n$	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot s}$	n

5. Calcular l'estadístic de prova, que en aquest cas anomenarem χ^2. Aquest estadístic es calcula segons la fórmula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(obs_{ij} - esp_{ij})^2}{esp_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}$$

6. Determinar la zona de les taules on s'accepta la H_0. La taula que s'ha de consultar és la χ^2 amb $(r-1)(s-1)$ graus de llibertat. Per a un nivell de significació α, si χ^2_α és tal que $P(\chi^2 > \chi^2_\alpha) = \alpha$ (és a dir, el punt de la taula χ^2 que deixa a la seva dreta una àrea igual a α), la zona de les taules on s'accepta la H_0 és l'interval $(0, \chi^2_\alpha)$.
7. Determinar la zona de les taules on s'accepta la H_1. Serà la zona complementària a la que hem trobat al pas 6), és a dir, serà l'interval (χ^2_α, ∞).
8. Segons els valors de l'estadístic de prova χ^2 i les zones d'acceptació de cada hipòtesi, decidir quina és la hipòtesi certa.
9. Calcular el valor del nivell de significació crític o p-valor α_c. El p-valor es troba calculant la probabilitat que una distribució χ^2 amb $(r-1)(s-1)$ graus de llibertat sigui més gran que l'estadístic de prova χ^2, és a dir, $P(\chi^2_{(r-1)(s-1)} > \chi^2) = \alpha_c$.
10. Segons els valors de α_c i de α, decidir quina és la hipòtesi certa.

Observacions

1. Quan s'aplica la prova khi quadrat s'ha de procurar que les **freqüències esperades siguin superiors a 5**. Si aquesta condició no es verifica, és recomanable agrupar successos perquè les freqüències augmentin; en aquest cas disminuiran els graus de llibertat.
2. La fórmula per calcular l'estadístic de prova χ^2 també es pot expressar com a:

$$\chi^2 = n\left[\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}\right)-1\right]$$

3. En una taula 2 × 2, aquesta fórmula és:

$$\chi^2 = n\frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

Si en una taula 2 × 2 les freqüències esperades són petites, haurem d'usar la **correcció de Yates**, que consisteix a utilitzar la fórmula:

$$\chi^2 = n\frac{\left(n_{11}n_{22}-n_{12}n_{21}-\frac{n}{2}\right)^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

4. Quan en una taula de contingència s'accepta que hi ha dependència entre les dues característiques A i B estudiades, convé calcular el coeficient:

$$C = \frac{\chi^2/n}{q}$$

sent $q = mín\{r,s\} - 1$. C és el **coeficient de contingència de Cramer**, que proporciona una mesura del grau d'associació entre els successos A_1, ..., A_r i els successos B_1, ..., B_s. Aquest coeficient verifica que:

a) $0 \leq C \leq 1$.

b) Si $C = 0$ existeix independència completa.

c) Com més gran sigui el valor de C, més dependència hi ha entre les categories de les característiques A i B.

Exemple. Es vol estudiar si hi ha alguna relació entre l'edat de les persones i el color del seu cotxe. Per fer aquest estudi s'agafen 300 persones i es classifiquen segons la seva franja d'edat i el color del seu cotxe. Les franges d'edat que s'han considerat són: de 18 a 30 anys, de 30 a 45 anys i més de 45 anys. El color del cotxe pot ser: vermell, blau, negre i gris. Els resultats obtinguts es mostren a la taula de contingència següent:

	vermell	*blau*	*negre*	*gris*	
18 - 30 anys	60	40	3	7	110
30 - 45 anys	15	8	22	75	120
+ 45 anys	5	2	35	28	70
	80	50	60	110	300

Volem veure si hi ha independència o no ($\alpha = 0.05$) entre l'edat de les persones i el color del seu cotxe.

Solució

1. Les dues característiques que es consideren són: edat i color del cotxe. Les categories en què s'ha dividit l'edat són: de 18 a 30 anys, de 30 a 45 anys i més de 45 anys. Les categories en què s'ha tingut en compte el color del cotxe són: vermell, blau, negre i gris.
2. La H_0 és:

   H_0: l'edat i el color del cotxe són característiques independents (no estan relacionades).

   H_0 significa que, per exemple, dintre d'una franja d'edat, la fracció de persones que tenen un color de cotxe o un altre és semblant a la fracció del total de persones que tenen aquell color de cotxe, és a dir, trobar-se en una franja d'edat o una altra no té influència en el color del cotxe.
3. La H_1 és:

   H_1: l'edat i el color del cotxe són característiques dependents (estan relacionades).

   Això vol dir que, sota H_1, la fracció del nombre de cotxes d'un color o un altre depèn de la franja d'edat de l'usuari.
4. La taula de freqüències esperades sota H_0 és:

	vermell	*blau*	*negre*	*gris*	
18 - 30 anys	110·80/300	110·50/300	110·60/300	110·110/300	110
30 - 45 anys	120·80/300	120·50/300	120·60/300	120·110/300	120
+ 45 anys	70·80/300	70·50/300	70·60/300	70·110/300	70
	80	50	60	110	300

és a dir,

	vermell	*blau*	*negre*	*gris*	
18 - 30 anys	29.3	18.3	22	40.3	110
30 - 45 anys	32	20	24	44	120
+ 45 anys	18.7	11,7	14	25,7	70
	80	50	60	110	300

5. Calculem l'estadístic de prova, χ^2.

$$\chi^2 = \frac{(60-29.3)^2}{29.3} + \frac{(40-18.3)^2}{18.3} + \ldots + \frac{(28-25.7)^2}{25.7} = 189.59$$

6. S'ha de consultar la taula χ^2 amb $(3-1)(4-1) = 6$ graus de llibertat. A la taula anterior hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $\chi^2_\alpha = 12.6$. Per tant, la zona de les taules on s'accepta la H_0 és l'interval $(0, 12.6)$.
7. La zona de les taules on s'accepta la H_1 és l'interval $(12.6 , \infty)$.
8. Com que l'estadístic de prova $\chi^2 = 189.59$ es troba a l'interval on s'accepta H_1, direm que acceptem H_1 i que, per tant, hi ha dependència entre l'edat i el color de cotxe.
9. El valor del nivell de significació crític o p-valor, α_c, és $3.1 \cdot 10^{-38}$.
10. Com que el p-valor, $\alpha_c = 3.1 \cdot 10^{-38}$, és més petit que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_1 i arribem a la mateixa conclusió que abans.
11. En aquest cas té sentit calcular el coeficient de contingència de Cramer:

$$C = \frac{\chi^2 / n}{q} = \frac{189.59/300}{2} = 0.316$$

7.2 Proves de bondat d'ajustament a una distribució

L'objectiu d'aquest apartat és donar les bases per, donat un conjunt de dades, contrastar si aquestes dades podem suposar que segueixen una certa distribució determinada. Per realitzar aquest contrast, ho podem fer a través de dues aproximacions diferents:

Avaluant les diferències que hi ha entre les freqüències esperades sota la distribució posada com a hipòtesi nul·la i les freqüències observades; en aquest cas s'usarà la prova khi quadrat.

Comparant la funció de distribució de la distribució posada com a hipòtesi nul·la i la funció de distribució mostral; sota aquesta aproximació s'usarà el test de Kolmogorov-Smirnov.

7.2.1 La prova khi quadrat

Partim de la base que tenim unes dades mostrals, $x_1, x_2, ..., x_n$, que provenen d'una població. El que volem contrastar és si aquestes dades podem suposar que segueixen certa variable aleatòria X o no. X pot ser qualsevol de les distribucions conegudes: binomial, Poisson, normal, exponencial, uniforme...

Per fer aquest contrast de bondat d'ajustament segons la prova khi quadrat, cal seguir els passos següents:

1. Determinar la hipòtesi nul·la, H_0. S'ha d'especificar quina és la distribució de referència X a partir de la qual volem comprovar si les dades s'hi ajusten o no. La H_0:

$$H_0\text{: les dades segueixen la distribució } X$$

2. Determinar la hipòtesi nul·la, H_1. La H_1 serà la contrària que la H_0:

$$H_1\text{: les dades no segueixen la distribució } X$$

3. Calcular les freqüències observades. Per fer això, dividirem el camp on pot prendre valors la variable aleatòria X en intervals de classe disjunts $I_1, I_2, ..., I_k$ i calcularem les freqüències absolutes $n_1, n_2, ..., n_k$, sent n_i el nombre de valors mostrals que pertanyen a l'interval I_i.
4. Calcular les freqüències esperades sota la H_0. Les freqüències esperades seran $np_1, np_2, ..., np_k$, on p_i és la probabilitat de l'interval I_i sota H_0.
5. Calcular l'estadístic de prova, que en aquest cas anomenarem χ^2. Aquest estadístic es calcula segons la fórmula:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(obs_i - esp_i\right)^2}{esp_i} = \sum_{i=1}^{k} \frac{\left(n_i - np_i\right)^2}{np_i}$$

6. Determinar la zona de les taules on s'accepta la H_0. La taula que s'ha de consultar és la χ^2 amb $k - s - 1$ graus de llibertat, on s és el nombre de paràmetres que s'han hagut d'estimar, a partir de les dades mostrals, de la distribució X. Per a un nivell de significació α, si χ^2_α és tal que $P(\chi^2 > \chi^2_\alpha) = \alpha$ (és a dir, el punt de la taula χ^2 que deixa a la seva dreta una àrea igual a α), la zona de les taules on s'accepta la H_0 és l'interval $(0, \chi^2_\alpha)$.
7. Determinar la zona de les taules on s'accepta la H_1. Serà la zona complementària a la trobada al pas 6), és a dir, serà l'interval (χ^2_α, ∞).
8. Segons els valors de l'estadístic de prova χ^2 i les zones d'acceptació de cada hipòtesi, decidir quina és la hipòtesi certa.

9. Calcular el valor del nivell de significació crític o p-valor α_c. El p-valor es troba calculant la probabilitat que una distribució χ^2 amb $k - s - 1$ graus de llibertat sigui més gran que l'estadístic de prova χ^2, és a dir, $P(\chi^2_{k-s-1} > \chi^2) = \alpha_c$.
10. Segons els valors de α_c i de α, decidir quina és la hipòtesi certa.

Observacions

1. La prova khi quadrat és de naturalesa discreta, és a dir, es comparen les freqüències observades i les freqüències esperades per a un nombre finit de categories. Per tant, si a la hipòtesi nul×la tenim una variable aleatòria X contínua, caldrà dividir el rang de possibles valors de la variable X en un nombre finit d'intervals de classe.
2. La prova khi quadrat ens afirma que l'estadístic χ^2 tendeix a una khi quadrat a mesura que n augmenta. S'ha vist que, a partir que n sigui 5 vegades el nombre d'intervals de classe, els resultats són acceptables. Per tant, seria bo seleccionar els intervals de classe de manera que tota freqüència esperada sigui major que 5. Això es pot aconseguir ajuntant intervals de classe veïns, però cal tenir en compte que el nombre de graus de llibertat es redueix en 1 cada vegada que ajuntem dos intervals.

Exemple 1. A la taula següent tenim el nombre de faltes comeses per alumne (hi ha 50 alumnes) en un dictat i volem veure si aquestes dades s'ajusten a una distribució de Poisson ($\alpha = 0.05$).

Faltes	0	1	2	3	4	5	6	7
Nombre d'alumnes	2	6	11	11	12	4	3	1

Solució

1. La H_0 és:

   H_0: les faltes per alumne segueixen una distribució Poisson

2. La H_1 és:

   H_1: les faltes per alumne no segueixen una distribució Poisson

3. Les freqüències observades es donen a la taula de l'enunciat.

Faltes	0	1	2	3	4	5	6	7 o més
Nombre d'alumnes observats	2	6	11	11	12	4	3	1

4. Per calcular les freqüències esperades, hem de suposar que les dades segueixen una Poisson. Com que no ens donen el paràmetre λ que té la distribució de Poisson, l'hem d'estimar a partir de la mitjana mostral i obtenim $\hat{\lambda} = 3.08$.

   A partir d'aquí apliquem la fórmula de la funció de probabilitat de la Poisson, amb $\lambda = 3.08$, per saber quina és la probabilitat de cada nombre de faltes.

Faltes	0	1	2	3	4	5	6	7 o més
Probabilitat (p_i)	0.0460	0.1416	0.2180	0.2238	0.1723	0.1062	0.0545	0.0377

   Per exemple, la probabilitat esperada del nombre d'alumnes que fan zero faltes és:

$$P(\text{Poiss}(3.08)=0) = e^{-3.08}\frac{3.08^0}{0!} = 0.0460$$

   Les freqüències esperades seran:

Faltes	0	1	2	3	4	5	6	7 o més
Nombre d'alumnes esperats	2.30	7.08	10.90	11.19	8.62	5.31	2.72	1.89

   Com que hi ha freqüències esperades que són més petites que 5, ajuntem intervals de classe:

Faltes	1 o menys	2	3	4	5 o més
Nombre d'alumnes esperats	9.38	10.90	11.19	8.62	9.92

   Fem el mateix amb les observades:

Faltes	1 o menys	2	3	4	5 o més
Nombre d'alumnes observats	8	11	11	12	8

5. Calculem l'estadístic de prova:

$$\chi^2 = \frac{(8-9.38)^2}{9.38} + \frac{(11-10.90)^2}{10.90} + \frac{(11-11.19)^2}{11.19} + \frac{(12-8.62)^2}{8.62} + \frac{(8-9.92)^2}{9.92} = 1.91$$

6. S'ha de consultar la taula χ^2 amb $k - s - 1 = 5 - 1 - 1 = 3$ graus de llibertat (k és 5 perquè, al final, s'han agafat 5 intervals de classe i s és 1 perquè s'ha hagut d'estimar el paràmetre λ de la Poisson). A la taula anterior hem de trobar el punt

que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $\chi^2_{\alpha} = 7.8$. Per tant, la zona de les taules on s'accepta la H_0 és l'interval $(0, 7.8)$.

7. La zona de les taules on s'accepta la H_1 és l'interval $(7.8 , \infty)$.
8. Com que l'estadístic de prova $\chi^2 = 1.91$ es troba a l'interval on s'accepta H_0, direm que acceptem H_0 i que, per tant, podem suposar que el nombre de faltes per alumne s'ajusta a una distribució Poisson.
9. El valor del nivell de significació crític o p-valor, α_c, és 0.58.
10. Com que el p-valor, $\alpha_c = 0.58$, és més gran que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_0 i arribem a la mateixa conclusió que abans.

Exemple 2. S'han recollit dades corresponents als pesos de 150 persones i s'ha obtingut una mitjana de 72 quilos i una desviació estàndard de 6 quilos. Volem saber si aquestes dades s'ajusten a una distribució normal (amb $\alpha = 0.05$) i, per això, s'han agrupat les dades en 6 intervals de classe. La taula següent mostra els resultats una vegada s'han agrupat els pesos.

Pes	50-61	61-65	65-69	69-74	74-80	80-110
Nombre de persones	8	12	20	54	34	22

Resolució.

1. La H_0 és:

   H_0: el pes segueix una distribució normal

2. La H_1 és:

   H_1: el pes no segueix una distribució normal

3. Les freqüències observades es donen a la taula de l'enunciat.

Pes	61 o menys	61-65	65-69	69-74	74-80	80 o més
Nombre de persones	8	12	20	54	34	22

4. Per calcular les freqüències esperades, hem de suposar que les dades segueixen una normal on s'han estimat els valors de la mitjana i la desviació estàndard; aquests valors són 72 i 6, respectivament.

   A partir d'aquí busquem a les taules de la normal per saber quina és la probabilitat de cada interval de pesos fet.

Pes	61 o menys	61-65	65-69	69-74	74-80	80 o més
Probabilitat (p_i)	0.0334	0.0883	0.1869	0.3220	0.2782	0.0912

Per exemple, la probabilitat esperada del nombre d'alumnes que pesen entre 61 i 65 quilos és:

$$P(61 < N(72,6) < 650) = 0.0883$$

Les freqüències esperades seran:

Pes	61 o menys	61-65	65-69	69-74	74-80	80 o més
Nombre de persones esperades	5.01	13.24	28.03	48.30	41.73	13.68

Com que no hi ha freqüències esperades que siguin més petites que 5, no ajuntarem intervals de classe i treballarem amb els que tenim actualment.

5. Calculem l'estadístic de prova:

$$\chi^2 = \frac{(8-5.01)^2}{5.01} + \frac{(20-13.24)^2}{13.24} + \ldots + \frac{(22-13.68)^2}{13.68} = 11.37$$

6. S'ha de consultar la taula χ^2 amb $k - s - 1 = 6 - 2 - 1 = 3$ graus de llibertat (k és 6 perquè s'han agafat 6 intervals de classe i s és 2 perquè s'han hagut d'estimar els paràmetres mitjana i desviació estàndard de la distribució normal). A la taula anterior hem de trobar el punt que deixa a la seva dreta una àrea de $\alpha = 0.05$. Aquest punt és $\chi^2_\alpha = 7.8$. Per tant, la zona de les taules on s'accepta la H_0 és l'interval $(0, 7.8)$.
7. La zona de les taules on s'accepta la H_1 és l'interval $(7.8, \infty)$.
8. Com que l'estadístic de prova $\chi^2 = 11.37$ es troba a l'interval on s'accepta H_1, direm que acceptem H_1 i que, per tant, els pesos de les persones no s'ajusten a una distribució normal.
9. El valor del nivell de significació crític o p-valor, α_c, és 0.001.
10. Com que el p-valor, $\alpha_c = 0.001$, és més petit que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_1 i arribem a la mateixa conclusió que abans.

7.2.2 El test de Kolmogorov-Smirnov

Per aplicar la prova de la khi quadrat cal un nombre relativament elevat de dades perquè es compleixi que les freqüències esperades, sota la H_0, de cada interval de classe sigui major que 5 i que hi hagi un nombre d'intervals de classe suficient per no quedar-nos sense graus de llibertat quan consultem la taula χ^2. Per exemple, si disposem de menys de 20 dades i volem comprovar si podem suposar que les dades provenen d'una distribució normal, hauríem de tenir un màxim de 3 intervals de classe perquè a cadascun hi hagués un mínim de 5 dades esperades. A més, cal estimar 2 paràmetres: la mitjana i la desviació estàndard de la normal; en aquest cas, els graus de llibertat amb què caldria consultar la taula χ^2 serien $3 - 2 - 1 = 0$ graus de llibertat!!!

La prova de Kolmogorov-Smirnov no necessita que les dades es trobin agrupades i es pot aplicar quan les mostres tenen pocs elements. Com s'ha comentat abans, en aquest test es compararà la funció de distribució de la distribució proposada a la hipòtesi nul·la i la funció de distribució de la mostra una vegada aquesta ha estat ordenada. Si aquesta comparació ens mostra una diferència prou gran entre les funcions de distribució mostral i la proposada sota H_0, llavors la hipòtesi nul·la es rebutja; si la diferència és petita, llavors s'accepta la H_0.

Per aplicar el test de Kolmogorov-Smirnov els passos que cal seguir són:

1. Determinar la hipòtesi nul·la, H_0. S'ha d'especificar quina és la distribució de referència X a partir de la qual volem comprovar si les dades s'hi ajusten o no. La H_0:

   H_0: les dades segueixen la distribució X.

2. Determinar la hipòtesi nul·la, H_1. La H_1 serà la contrària que la H_0:

   H_1: les dades no segueixen la distribució X.

3. Calcular la funció de distribució mostral de les dades ordenades. Per això, suposem que tenim una mostra $x_1, ..., x_n$ i l'ordenem. Notem mitjançant $x_{(1)}, ..., x_{(n)}$ la mostra ordenada:

$$x_{(1)} < ... < x_{(n)}$$

La funció de distribució mostral és:

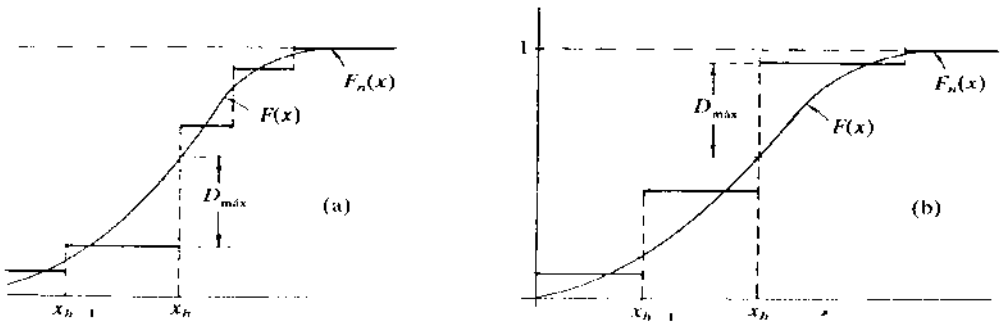
$$S_n(x) = \begin{cases} 0 & si\ x < x_{(1)} \\ \dfrac{i}{n} & si\ x_{(i)} \le x < x_{(i+1)} \quad i = 1, \ldots, n-1 \\ 1 & si\ x \ge x_{(n)} \end{cases}$$

És a dir, per a qualsevol valor ordenat x de la mostra aleatòria, $S_n(x)$ és la proporció del nombre de valors a la mostra que són iguals o menors que x.

4. Calcular la funció de distribució, suposant certa la H_0, en cadascun dels valors mostrals ordenats. És a dir, per a cada valor $x_{(i)}$, cal trobar $F(x_{(i)}) = P(X \leq x_{(i)})$.
5. Calcular l'estadístic de prova, que en aquest cas notarem amb D_n i s'anomena *estadístic de Kolmogorov-Smirnov*. Aquest estadístic és la diferència màxima entre la funció de distribució mostral ordenada i la funció de distribució sota H_0 i es defineix com a:

$$D_n = màx\ |S_n(x) - F_0(x)|$$

A efectes pràctics, cal tenir present que es poden donar dues situacions diferents:



a) La distància màxima entre $F(x)$ i $S_n(x)$ l'obtenim just abans d'arribar fins a x_h i val $|S_n(x_{h-1}) - F(x_h)|$.

b) La distància màxima és $|S_n(x_h) - F(x_h)|$.

Per tant, quan apliquem el test, s'ha de calcular per a cada punt x_h:

$$D_n(x_h) = màx\ \{|S_n(x_{h-1}) - F_0(x_h)|, |S_n(x_h) - F_0(x_h)|\}$$

i després agafem el màxim d'aquests $D_n(x_h)$. Aquest últim valor serà D_n.

6. Determinar la zona de les taules on s'accepta la H_0. Tenim una taula especial, taula de Kolmogorov-Smirnov, que conté, per a cada $n > 1$, els punts crítics a_α tals que:

$$P(D_n > a_\alpha) = \alpha$$

per a diferents valors de α. Llavors, per a un nivell de significació α, la zona de les taules on s'accepta la H_0 és l'interval $(0, a_\alpha)$.
7. Determinar la zona de les taules on s'accepta la H_1. Serà la zona complementària a la que hem trobat al pas 6), és a dir, serà l'interval $(a_\alpha, 1)$.

8. Segons els valors de l'estadístic de prova D_n i les zones d'acceptació de cada hipòtesi, decidir quina és la hipòtesi certa.

Exemple. Volem contrastar, amb $\alpha = 0.5$, si la mostra següent de durades de vida de cert dispositiu es pot suposar exponencial:

16, 8, 10, 12, 6, 10, 20, 7, 2, 24.

Solució

1. La H_0 és:

   H_0: la durada segueix una distribució exponencial.

2. La H_1 és:

   H_1: la durada no segueix una distribució exponencial.

3, 4 i 5.

Com que no s'especifica quin és el paràmetre λ de la funció exponencial que hem d'usar, el primer que farem és estimar-lo a partir de les dades. Com que la mitjana mostral és 11.5, l'estimació del paràmetre λ serà 1/11.5.

Construïm la taula següent:

x_h	$F(x_h)$	$S_n(x_h)$	$S_n(x_{h-1})$	$\lvert S_n(x_h) - F(x_h)\rvert$	$\lvert S_n(x_{h-1}) - F(x_h)\rvert$	$D_n(x_h)$
2	0.16	0.1	0	0.06	0.16	0.16
6	0.41	0.2	0.1	0.21	0.31	0.31
7	0.46	0.3	0.2	0.16	0.26	0.26
8	0.5	0.4	0.3	0.10	0.20	0.20
10	0.58	0.5	0.4	0.08	0.18	0.18
10	0.58	0.6	0.5	0.02	0.08	0.08
12	0.65	0.7	0.6	0.05	0.05	0.05
16	0.75	0.8	0.7	0.05	0.05	0.05
20	0.82	0.9	0.8	0.08	0.02	0.08
24	0.88	1	0.9	0.12	0.02	0.12

L'estadístic D_n pren un valor de 0.31.

6. Agafant un nivell de significació $\alpha = 0.05$, obtenim a la taula de Kolmogorov-Smirnov ($n = 10$) un valor $a_{0.05} = 0.409$. Per tant, la zona de les taules on s'accepta la H_0 és l'interval (0, 0.409).
7. La zona de les taules on s'accepta la H_1 és l'interval (0.409, 1).

8. Com que l'estadístic de prova $D_n = 0.31$ es troba a l'interval on s'accepta H_0, direm que acceptem H_0 i que, per tant, podem suposar que la durada del dispositiu segueix una distribució exponencial.

7.2.2.1 El contrast de Kolmogorov-Smirnov-Lilliefors

Suposem que estem sota les mateixes condicions amb les quals aplicàvem el contrast de Kolmogorov-Smirnov i que volem contrastar que la distribució de X és $N(\mu,\sigma)$, on μ i σ són desconeguts o no. Aleshores, estimarem μ i σ mitjançant $\overline{X}$ i S, respectivament. Això ens permetrà especificar de manera completa el model sota la hipòtesi nul·la i, per tant, poder operar com en el cas del contrast de Kolmogorov-Smirnov utilitzant, en aquest cas, la taula de Lilliefors.

Exemple. Volem contrastar, amb $\alpha = 0.05$, si la mostra següent de durades de vida de cert dispositiu es pot suposar que segueix una distribució normal:

16, 8, 10, 12, 6, 10, 20, 7, 2, 24.

Solució

1. La H_0 és:

   H_0: la durada segueix una distribució normal.

2. La H_1 és:

   H_1: la durada no segueix una distribució normal.

3, 4 i 5.

Com que no s'especifica quins són els paràmetres μ i σ de la distribució normal, el primer que farem és estimar-los a partir de les dades. Aquestes estimacions són 11.5 i 6.72, respectivament.

Construïm la taula següent:

x_h	$F(x_h)$	$S_n(x_h)$	$S_n(x_{h-1})$	$\lvert S_n(x_h) - F(x_h)\rvert$	$\lvert S_n(x_{h-1}) - F(x_h)\rvert$	$D_n(x_h)$
2	0.08	0.1	0	0.02	0.08	0.08
6	0.21	0.2	0.1	0.01	0.11	0.11
7	0.25	0.3	0.2	0.05	0.05	0.05
8	0.30	0.4	0.3	0.10	0.00	0.10
10	0.41	0.5	0.4	0.09	0.01	0.09
10	0.41	0.6	0.5	0.19	0.09	0.19
12	0.53	0.7	0.6	0.17	0.07	0.17
16	0.75	0.8	0.7	0.05	0.05	0.05
20	0.90	0.9	0.8	0.00	0.10	0.10
24	0.97	1	0.9	0.03	0.07	0.07

L'estadístic D_n pren un valor de 0.19.

6. Agafant un nivell de significació $\alpha = 0.05$, obtenim a la taula de Lilliefors ($n = 10$) un valor $a_{0.05} = 0.262$. Per tant, la zona de les taules on s'accepta la H_0 és l'interval $(0, 0.262)$.
7. La zona de les taules on s'accepta la H_1 és l'interval $(0.262, 1)$.
8. Com que l'estadístic de prova $D_n = 0.19$ es troba a l'interval on s'accepta H_0, direm que acceptem H_0 i que, per tant, podem suposar que la durada del dispositiu segueix una distribució normal.

8. Regressió lineal

8.1 Relació entre variables

Per relacionar dues o més variables a través d'alguna funció tenim diverses tècniques estadístiques. L'aplicació d'una tècnica o d'una altra dependrà de la quantitat de variables, el tipus de variables i la funció que relaciona les variables.

El cas més general de relació entre variables seria:

$$G(y_1, y_2, ..., y_m) = F(x_1, x_2, ..., x_n)$$

on les variables y_i s'anomenen *variables dependents* i les variables x_i s'anomenen *variables independents*. L'objectiu general és predir valors de les variables dependents a partir dels valors de les variables independents.

En el cas particular que només hi hagi una sola variable dependent, la relació la notarem segons:

$$y = F(x_1, x_2, ..., x_n)$$

Les tècniques estadístiques que podem usar per relacionar variables són:

			Variable dependent		
			1		>1
			Numèrica	No numèrica	
Variable independent	1	Numèrica	Regressió simple	Regressió logística	
		No numèrica	ANOVA 1 factor		
	>1	Numèrica	Regressió múltiple	Anàlisi discriminant	Anàlisi canònica
		No numèrica	ANOVA 2 o més factors	Anàlisi conjunta	MANOVA 2 o més factors
		Barreja	ANCOVA	Regressió logística i ordinal	MANCOVA

En el cas que la funció *F* sigui lineal, es diu que fem regressió lineal simple i regressió lineal múltiple. En aquests casos tindrem que les relacions entre variables són donades per:

- Regressió lineal simple: $Y = \alpha + \beta X$, equació d'una recta on α és l'ordenada a l'origen i β és el pendent de la recta. El pendent d'una recta indica en quina quantitat augmenta (o disminueix, si el signe del pendent és negatiu) la variable *Y* per cada unitat que augmenta la variable *X*. L'ordenada a l'origen indica quin és el valor de la *Y* quan la *X* és 0 (punt de tall de la recta amb l'eix d'ordenades).
- Regressió lineal múltiple: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$, on β_0 és el terme independent de l'equació i els restants β_i són els coeficients que acompanyen cadascuna de les variables independents. La interpretació de cada coeficient β_i seria anàloga a la del pendent però referint-se a la variable X_i que acompanya el coeficient β_i, és a dir, és la quantitat que augmenta (o disminueix, si el signe del coeficient β_i és negatiu) la variable Y per cada unitat que augmenta la variable X_i (suposant que la resta de variables independents es mantenen fixes).

En el cas que la funció no sigui lineal, direm que fem regressió no lineal. Aquesta pot adoptar diverses formes: potencial, logarítmica, exponencial, polinòmica...

8.2 Model de regressió mostral simple

La formulació d'un model de regressió requereix delimitar el fenomen que es vol estudiar, localitzar les variables i establir les relacions que hi ha entre aquestes.

Exemples

- Podem establir una relació lineal entre el pes (P) i les altures (A):

$$P = \alpha + \beta A$$

- Existeix una relació no lineal entre quantitat produïda (Q) i capital (K) i treball (L) del tipus:

$$Q = AL^{\alpha}K^{\beta}$$

Per altra banda, l'observació de la realitat ens permetrà obtenir les dades necessàries sobre les variables que componen el model per tenir una base sobre la qual treballarem. A partir del model que proposem i de les dades observades, s'usaran tècniques d'inferència estadística per estimar i verificar el model, la finalitat del qual serà fer prediccions.

8.2.1 Components d'un model

Distingim quatre elements en un model: equacions, variables, paràmetres i terme de pertorbació.

- **Equacions**: són les relacions que hi ha o hi pot haver entre les variables que estem estudiant. Per trobar el tipus d'equació o funció que millor s'ajusta a les nostres dades, podem fer un gràfic amb les dades i veure quina relació existeix entre elles.
- **Variables**: la classificació més generalitzada ens divideix les variables que formen part del model en:
  a) **Variables explicades o endògenes**: són les variables dependents que es volen explicar a través del model.
  b) **Variables explicatives o exògenes**: són les variables independents i que intenten explicar el comportament de les variables explicades.
- **Paràmetres**: són els coeficients que afecten les variables explicatives i mesuren l'efecte de les fluctuacions d'aquestes variables sobre la variable explicada.
- **Terme de pertorbació**: introduirem aquest component del model amb un exemple. Suposem que ens trobem davant d'un model lineal on es vol explicar

la quantitat de vi en cert lloc (V) a partir de la pluja caiguda el mes de maig (P): $V = \alpha + \beta P$. Per cada valor de la pluja (P) existeix un valor de la quantitat de vi (V) que ens és donat per l'equació anterior. Es tracta d'una relació on no hi ha aleatorietat. Per tant, segons aquest model, tots els anys que plogui el mateix (P), la producció de vi serà la mateixa, i sabem que això no és així. Per tant, en aquest model hi falta algun terme que ens pugui explicar les diferències que hi ha entre els individus. Si introduïm aquest terme, el model ens queda: $V = \alpha + \beta P + u$, on u serà una v. a. que anomenem **terme de pertorbació**.

Les principals funcions del terme de pertorbació són:

a) recollir les variables explicatives que no són al model.

b) recollir especificacions incorrectes de l'equació del model.

c) recollir els errors en la mesura de les variables.

d) recollir el comportament aleatori dels resultats.

8.2.2 Hipòtesis bàsiques del model de regressió lineal simple

Suposarem que tenim una variable explicada i una variable explicativa. El model de regressió lineal simple es basa en una sèrie d'hipòtesis sobre els diferents components del model.

Respecte a l'equació. Existeix una relació lineal entre la variable explicada i la variable explicativa. Formalment escriurem:

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, 2, ..., n$$

on:

- Y_i: variable explicada,
- X_i: variable explicativa,
- α, β: paràmetres,
- u_i: terme de pertorbació.

El subíndex i indica les observacions mostrals que tenim.

Respecte a la variable explicativa (X_i). El nombre d'observacions ha de ser més gran que el nombre de paràmetres. En el cas de la regressió lineal simple, el nombre d'observacions ha de ser més gran que 2.

Respecte als paràmetres. α i β són constants al llarg del mostreig. Aquestes són les constants que tractarem d'aproximar mitjançant la inferència estadística.

8.3 Regressió lineal simple: estimació de la recta de regressió

A la pràctica la recta de regressió poblacional serà desconeguda i l'haurem d'estimar per obtenir $\hat{\alpha}$ i $\hat{\beta}$. L'estimació d'aquesta recta rep el nom de **recta de regressió mostral** i es representa mitjançant:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

Una vegada tinguem la recta de regressió mostral, ens apareixeran els **errors d'estimació o residus**, e_i, els quals es defineixen com la diferència entre el valor real Y_i i el valor estimat $\hat{y}_i$:

$$e_i = Y_i - \hat{Y}_i \quad i = 1, 2, ..., n$$

El problema que se'ns planteja ara és determinar un criteri per estimar els paràmetres α i β de la recta. El **criteri dels mínims quadrats ordinaris** (**MQO**) es basa en la minimització de la suma dels quadrats dels residus. Altres criteris serien el que es basa en la minimització de la suma dels residus (aquest criteri no dóna una solució única, ja que els residus poden ser positius o negatius i la manera de compensar-se entre ells pot ser diversa) i el que es basa en la minimització de la suma dels valors absoluts dels residus (aquest criteri té l'inconvenient, respecte al criteri de mínims quadrats ordinaris, que és més complicat de treballar amb valors absoluts de certes quantitats que amb els seus quadrats). Anem a deduir, doncs, els valors de $\hat{\alpha}$ i $\hat{\beta}$ segons el criteri MQO.

El que intenta el criteri MQO és minimitzar la funció següent:

$$\Phi = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

Això ho podem resoldre com un problema de màxims i mínims amb més d'una variable. El resultat que obtenim és:

$$\hat{\beta} = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2} \quad \text{i} \quad \hat{\alpha} = \overline{Y} - \hat{\beta} \overline{X}$$

Manipulant la primera expressió anterior obtenim altres expressions equivalents per a $\hat{\beta}$. En aquests casos obtenim:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n \overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n \overline{X}^2} \quad \text{i} \quad \hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Propietats

- La suma dels residus és zero: $\sum e_i = 0$.
- La suma dels valors reals de la variable Y és igual a la suma dels valors ajustats de la variable Y segons la recta de regressió: $\sum Y_i = \sum \hat{Y}_i$.

Exemple. Suposem que entre l'alçada (Y) i el pes (X) de les persones hi ha una relació lineal. A partir de les dades següents hem de trobar els paràmetres de la recta de regressió:

Y	174	168	181	170	158	177	159	164
X	77	70	79	68	56	80	56	64

$$\overline{X} = 68.75 \qquad \overline{Y} = 168.875$$

Y_i	X_i	X^2_i	X_iY_i
174	77	5929	13398
168	70	4900	11760
181	79	6241	14299
170	68	4624	11560
158	56	3136	8848
177	80	6400	14160
159	56	3136	8904
164	64	4096	10496
1351	550	38462	93425

Aleshores:

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{8 \cdot 93425 - 550 \cdot 1351}{8 \cdot 38462 - 550^2} = 0.837182$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 168.875 - 0.837182 \cdot 68.75 = 111.3187$$

Per tant, la recta de regressió que obtenim és:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i = 111.3187 + 0.837182 X_i$$

8.4 Regressió lineal simple: mesures de bondat d'ajustament

La bondat de l'ajustament d'una funció, en el nostre cas d'una recta, al núvol de punts es pot mesurar amb diferents coeficients: el coeficients de correlació *r*, el coeficient de determinació r^2 i l'error estàndard.

Si en un model obtenim que les mesures de bondat d'ajustament lineal no són bones, això pot ser degut a dues causes:

1. No existeix una relació lineal entre les dues variables, però existeix algun altre tipus de relació, per exemple, logarítmica, exponencial o quadràtica.
2. No existeix cap mena de relació entre les dues variables. Les dues variables són totalment independents.

Si es fa el gràfic de dispersió de les dues variables podrem tenir una idea aproximada de si el tipus de relació entre les dues variables és lineal, no lineal o inexistent.

8.4.1 Coeficient de correlació

El coeficient de correlació de Pearson, *r*, mesura la relació lineal que hi ha entre dues variables i es calcula mitjançant alguna de les fórmules següents:

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \cdot \sqrt{\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2}} =$$

$$= \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \cdot \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$$

Propietats

a) *r* és un estimador del coeficient de correlació poblacional ρ.

b) $-1 \leq r \leq 1$.

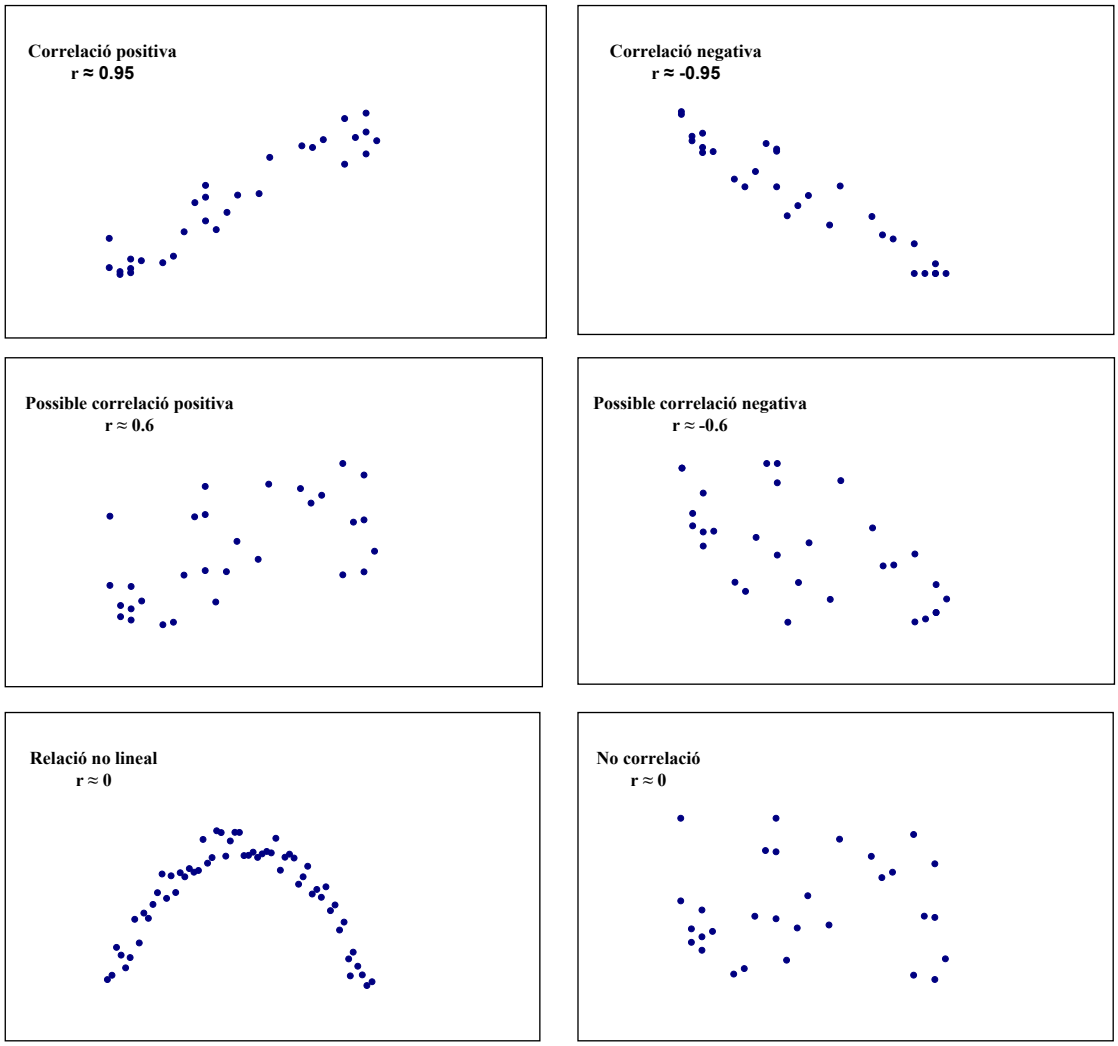
c) Si $0 < r < 1$, es diu que la correlació és positiva i indica que, si incrementem el valor de la variable *X*, també s'incrementarà el valor de la variable *Y*.

d) Si $-1 < r < 0$, es diu que la correlació és negativa i indica que, si incrementem el valor de la variable *X*, disminuirà el valor de la variable *Y*.

e) Si $r = \pm 1$, vol dir que una variable és exactament combinació lineal de l'altra i es diu que existeix **correlació total.**

f) Si $r = 0$, vol dir que no existeix cap mena de relació lineal entre les dues variables estudiades i diem que les variables estan **incorrelacionades.**

A les figures següents s'han representat diversos tipus de relacions que poden aparèixer.



8.4.2 Coeficient de determinació

Definició. El **coeficient de determinació** és el percentatge de la variació total de la variable explicada que queda explicada per la variable explicativa. Per calcular r^2 només cal elevar el coeficient de correlació r al quadrat.

Propietats

a) $0 \leq r^2 \leq 1$, fet que resulta evident de la primera propietat del coeficient de correlació.

b) Com més a prop d'1 valgui r^2, més significativa és la relació lineal entre X i Y i, en canvi, com més a prop de 0 valgui r^2, menys significativa és la relació lineal entre X i Y.

c) Si el model lineal és perfecte, el coeficient de determinació serà $r^2 = 1$.

d) Si el model lineal no explica res de la variació total de Y, el coeficient de determinació serà 0.

8.4.3 Error estàndard

Quan aproximem els valors Y_i mitjançant $\hat{y}_i = \hat{\alpha} + \hat{\beta} X_i$, ens trobem amb els errors d'estimació o residus e_i, on:

$$e_i = Y_i - \hat{Y}_i \quad i = 1, 2, ..., n$$

Com més grans siguin aquests residus, pitjor serà la relació lineal que existeix entre les variables X i Y. A l'inrevés, com més petits siguin aquests residus, millor serà la relació lineal que existeix entre les variables X i Y.

També s'ha de tenir en compte el nombre d'observacions que tenim. D'aquesta manera podem calcular l'error estàndard mitjançant l'expressió següent:

$$S_u = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - 2}}$$

8.4.4 Contrast de significativitat de la regressió

Quan tenim un coeficient de correlació donat ens podem preguntar:

- Aquest coeficient de correlació és significatiu?
- A partir de quin valor es pot considerar que un coeficient de correlació és significatiu?

Cal esmentar que, per a un nivell de significació donat, no hi ha un valor concret a partir del qual un coeficient de correlació es pugui considerar significatiu, ja que aquest valor depèn del nombre de dades de què disposem. Per determinar si el coeficient de correlació realment és significatiu o no, és a dir, per determinar si realment té sentit ajustar una recta de regressió a unes dades, cal fer un contrast d'hipòtesis.

Els passos que s'han de seguir per fer un contrast d'hipòtesis sobre el coeficient de correlació poblacional ρ serien:

1. Determinar la hipòtesi nul·la, H_0. En aquest cas, la H_0 és:

$$H_0: \rho = 0$$

Això vol dir que, sota H_0, el coeficient de correlació no és significatiu i no té sentit ajustar una recta de regressió.

2. Determinar la hipòtesi alternativa, H_1. En aquest cas, la H_1 és:

$$H_1: \rho \neq 0$$

Això vol dir que, sota H_1, el coeficient de correlació és significatiu i té sentit ajustar una recta de regressió.

3. Determinar la zona de les taules estadístiques on s'accepta la H_0. La taula que s'ha de consultar és la t de Student amb $n - 2$ graus de llibertat. Per a un nivell de significació α, si $t_{\alpha/2}$ és tal que $P(t > t_{\alpha/2}) = \alpha / 2$ (és a dir, el punt de la taula t de Student que deixa a la seva dreta una àrea igual a $\alpha / 2$), la zona de la taula t de Student on s'accepta la H_0 és l'interval $(-t_{\alpha/2} , t_{\alpha/2})$.
4. Determinar la zona de les taules estadístiques on s'accepta la H_1. Serà la zona complementària a la trobada al pas 3), és a dir, serà $(-\infty , -t_{\alpha/2}) \cup (t_{\alpha/2} , \infty)$.
5. Determinar la zona de l'estadístic mostral on s'accepta la H_0. En aquest cas, l'estadístic mostral és el coeficient de correlació r i s'acceptarà la H_0 si:

$$r \in \left(-\frac{t_{\alpha/2}}{\sqrt{t_{\alpha/2}^2 + n - 2}} , \frac{t_{\alpha/2}}{\sqrt{t_{\alpha/2}^2 + n - 2}} \right)$$

6. Determinar la zona de l'estadístic mostral on s'accepta la H_1. Serà la zona complementària a la trobada al pas 5), és a dir, s'acceptarà H_1 si:

$$r \in \left(-1, -\frac{t_{\alpha/2}}{\sqrt{t_{\alpha/2}^2 + n - 2}} \right) \cup \left(\frac{t_{\alpha/2}}{\sqrt{t_{\alpha/2}^2 + n - 2}} , 1 \right)$$

7. Calcular l'estadístic de prova, t, de la manera següent:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}}$$

8. Segons els valors de l'estadístic de prova t i l'estadístic mostral r i les zones d'acceptació de cada hipòtesi, decidir quina és la hipòtesi certa.
9. Calcular el valor del nivell de significació crític o p-valor α_c. Per portar a terme aquest càlcul cal trobar la probabilitat que una distribució t de Student,

amb $n - 2$ graus de llibertat, sigui major que l'estadístic $|t|$ i després multiplicar aquesta probabilitat per 2. Matemàticament, tindrem:

$$\alpha_c = 2 \cdot P(t_{n-2} > |t|).$$

10. Segons els valors de α_c i de α, decidir quina és la hipòtesi certa.

Exemple. Seguint amb l'exemple dels pesos i les alçades, calcularem les diverses mesures de bondat d'ajustament estudiades. Recordem que la recta de regressió obtinguda és:

$$\hat{Y}_i = 111.3187 + 0.837182X_i$$

El quadre següent mostra els resultats de les operacions prèvies que cal realitzar per calcular les diverses mesures de bondat d'ajustament:

Y_i	X_i	X^2_i	X_iY_i	Y^2_i	$\hat{Y}_i$	e_i	e^2_i
174	77	5929	13398	30276	175.782	−1.782	3.17
168	70	4900	11760	28224	169.921	−1.921	3.69
181	79	6241	14299	32761	177.456	3.544	12.56
170	68	4624	11560	28900	168.247	1.753	3.07
158	56	3136	8848	24964	158.201	−0.201	0.04
177	80	6400	14160	31329	178.293	−1.293	1.67
159	56	3136	8904	25281	158.201	0.799	0.64
164	64	4096	10496	26896	164.898	−0.898	0.81
1351	550	38462	93425	228631			25.66

Aleshores:

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}} =$$

$$= \frac{8 \cdot 93425 - 550 \cdot 1351}{\sqrt{8 \cdot 38462 - 550^2} \cdot \sqrt{8 \cdot 228631 - 1351^2}} = 0.972957$$

$$r^2 = 0.972957^2 = 0.946645$$

$$S_u = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{25.66}{6}} = 2.068$$

Finalment, per determinar si el coeficient de correlació és significatiu o no, amb $\alpha = 0.05$, seguirem els passos indicats per fer el contrast:

1. H_0: $\rho = 0$.
2. H_1: $\rho \neq 0$.
3. Com que $n = 8$, s'ha de consultar la taula t de Student amb 6 graus de llibertat i mirar quin punt deixa a la seva dreta una àrea de $\alpha / 2 = 0.025$. Aquest punt és el 2.45. Per tant, la zona de la taula t de Student on s'accepta la H_0 és l'interval $(-2.45 , 2.45)$.
4. La zona de la taula t de Student on s'accepta la H_1 és $(-\infty, -2.45) \cup (2.45, \infty)$.
5. La H_0 s'acceptarà si r compleix que:

$$r \in (-0.7072, 0.7072)$$

6. La H_1 s'acceptarà si r compleix que:

$$r \in (-1, -0.7072) \cup (0.7072,1)$$

7. L'estadístic de prova, t, és:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 10.32$$

8. Com que r i t pertanyen a la zona on s'accepta la H_1, s'accepta que el coeficient de correlació és significatiu i que té sentit fer la regressió lineal.
9. Calculem el p-valor α_c:

$$\alpha_c = 2 \cdot P(t_{n-2} > t) = 2 \cdot 2.42 \cdot 10^{-5} = 4.84 \cdot 10^{-5}.$$

10. Com que el p-valor, $\alpha_c = 4.84 \cdot 10^{-5}$, és més petit que el nivell d'error amb el qual volem treballar, $\alpha = 0.05$, acceptem H_1 i arribem a la mateixa conclusió que abans.

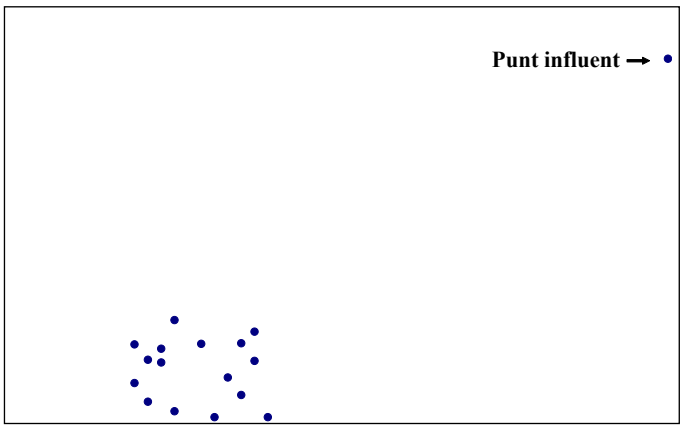
8.5 Regressió lineal simple: punts influents i punts atípics

Definició. Un **punt o observació influent** és un punt que, si es fan els càlculs de la regressió amb aquest punt o sense aquest punt, provoca resultats diferents ja sigui de l'estimació de la recta de regressió o del coeficient de correlació o d'ambdós.

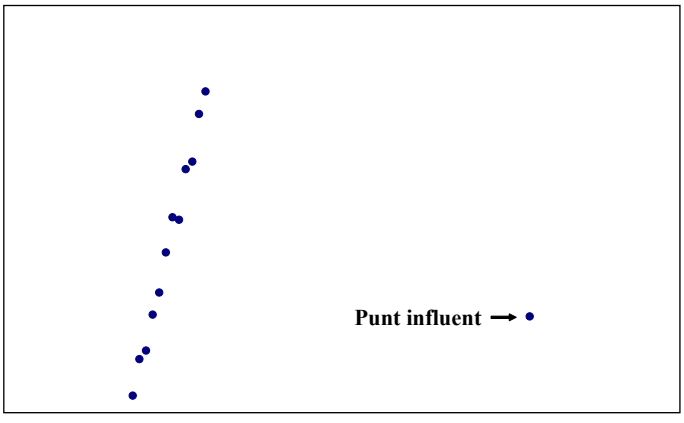
Definició: un **punt o observació atípic** és un punt que s'aparta del comportament de la resta de punts.

Per detectar punts influents i punts atípics hi ha diverses eines, la més senzilla de les quals és fer el gràfic de dispersió de les dades. Posteriorment es poden ajustar rectes de regressió amb i sense els punts candidats a ser influents i/o atípics per valorar-ne la influència i/o comportament.

Exemple 1. En el gràfic següent es mostra un punt influent, ja que el coeficient de correlació sense tenir en compte el punt influent és $r = -0.11$ (no significatiu), i tenint en compte el punt influent és $r = 0.85$ (significatiu).

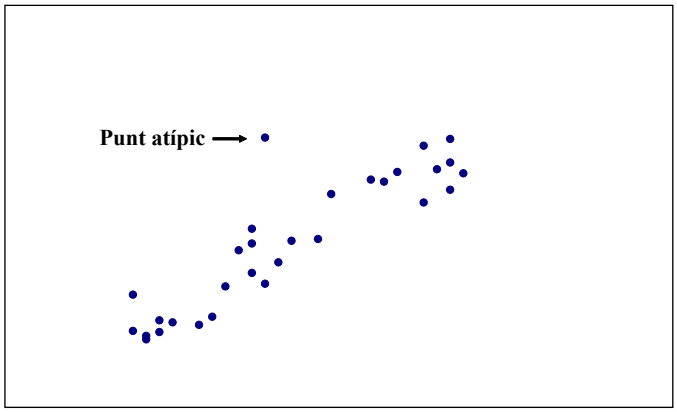


Exemple 2. En el gràfic següent també es mostra un punt influent, ja que el coeficient de correlació sense tenir en compte el punt influent és $r = 0.9955$ (molt significatiu) i tenint en compte el punt influent és $r = 0.02$ (no significatiu).



Exemple 3. En el gràfic següent es mostra un punt atípic, ja que té un comportament diferent de la resta de punts. La valoració de si és un punt influent no és tan clara,

ja que el coeficient de correlació sense tenir en compte el punt atípic és $r = 0.95$ i tenint en compte el punt atípic és $r = 0.89$.



8.6 Regressió lineal simple: construcció d'intervals de predicció

Una aplicació important de la regressió és la d'usar el model estimat per fer prediccions, és a dir, determinar el valor Y_0 que correspon a un valor determinat X_0 de la variable explicativa. Així, tenint en compte que:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

per a $X_i = X_0$ tenim que:

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0$$

on $\hat{Y}_0$ serà una **predicció puntual** de Y_0.

És clar que un valor concret de la variable explicativa X_0 no sempre produirà un mateix valor de la variable explicada Y_0. Seria més correcte pensar que els possibles valors de Y_0 estarien dintre d'un rang de nombres que posarem en forma d'interval. Cal distingir si aquest interval és per a valors particulars de Y_0 o per a l'esperança de Y_0.

8.6.1 Interval per a valors particulars de Y_0

Si volem construir un interval amb un nivell de confiança igual a $1 - \alpha$, farem servir les taules de la t de Student amb $n - 2$ graus de llibertat per tal de trobar el valor $t_{\alpha/2}$ que deixa a la seva dreta una àrea igual a $\alpha / 2$. Amb aquestes eines puc construir l'interval de predicció per a valors particulars de Y_0, amb un nivell de significació α:

$$\hat{Y}_0 - t_{\alpha/2} S_u \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} + 1} \leq Y_0 \leq \hat{Y}_0 + t_{\alpha/2} S_u \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} + 1}$$

8.6.2 Interval per a l'esperança de Y_0

És natural pensar que l'interval per a l'esperança d'un valor Y_0 serà més petit que l'interval per a valors particulars (suposant que treballem amb el mateix nivell de significació). L'interval el trobem fent els càlculs següents:

$$\hat{Y}_0 - t_{\alpha/2} S_u \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}} \leq Y_0 \leq \hat{Y}_0 + t_{\alpha/2} S_u \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}}$$

Exemple: seguint amb l'exemple dels pesos i les alçades anterior, suposem que volem estimar l'alçada d'una persona que sabem que pesa 75 kg. La recta de regressió obtinguda és:

$$\hat{Y}_i = 111.3187 + 0.837182 X_i$$

Per tant, una predicció puntual de l'alçada quan $X_0 = 75$ és:

$$\hat{Y}_0 = 111.3187 + 0.837182 \cdot 75 = 174.1074$$

Si volem els intervals de predicció per a $\alpha = 0{,}05$, tenim les dades següents:

$$n = 8 \qquad X_0 = 75 \qquad \hat{Y}_0 = 174.1074 \qquad t_{0.025,\,6} = 2.45$$
$$S_u = 2.068 \qquad \overline{X} = 68.75 \qquad \sum X^2_i = 38462$$

L'interval de predicció per a valors particulars de Y_0 és:

$$174.1074 - 2.45 \cdot 2.068 \cdot 1.0886 \leq Y_0 \leq 174.1074 + 2.45 \cdot 2.068 \cdot 1.0886$$
$$168.5920 \leq Y_0 \leq 179.6228$$

L'interval de predicció per a l'esperança de Y_0 és:

$$174.1074 - 2.45 \cdot 2.068 \cdot 0.4303 \leq Y_0 \leq 174.1074 + 2.45 \cdot 2.068 \cdot 0.4303$$
$$171.9274 \leq Y_0 \leq 176.2873$$

8.7 Regressió no lineal simple

En aquest apartat suposarem que la relació existent entre la variable explicada i la variable explicativa no és de tipus lineal. En alguns casos haurem de fer algun tipus de transformació per aconseguir una relació lineal, $Y' = \alpha' + \beta' X'$, entre les variables transformades. Posem 3 exemples de regressió no lineal entre una variable explicada i una d'explicativa:

Exemple 1. La relació que es compleix és del tipus:

$$e^Y = a \cdot X^b$$

on a i b són els paràmetres desconeguts.

Si prenem logaritmes, obtenim:

$$Y = ln\ a + b \cdot ln\ X$$

Aquesta equació és lineal i podem treballar normalment agafant com a:

- Variable explicada: $Y' = Y$
- Variable explicativa: $X' = ln\ X$
- Paràmetres: $\alpha' = ln\ a$ i $\beta' = b$.

Una vegada hem estimat els paràmetres α' i β' per MQO podrem recuperar l'equació inicial.

Exemple 2. La relació que es compleix és del tipus:

$$Y = a \cdot X^b$$

on a i b són els paràmetres desconeguts.

Si prenem logaritmes, obtenim:

$$ln\ Y = ln\ a + b \cdot ln\ X$$

Aquesta equació és lineal i podem treballar normalment agafant com a:

- Variable explicada: $Y' = ln\ Y$
- Variable explicativa: $X' = ln\ X$
- Paràmetres: $\alpha' = ln\ a$ i $\beta' = b$.

Una vegada hem estimat els paràmetres α' i β' per MQO podrem recuperar l'equació inicial.

Exemple 3. La relació que es compleix és del tipus:

$$Y = a + \frac{b}{X}$$

on a i b són els paràmetres desconeguts.

Aquesta equació és lineal i podem treballar normalment agafant com a:

- Variable explicada: $Y' = Y$
- Variable explicativa: $X' = 1/X$
- Paràmetres: $\alpha' = a$ i $\beta' = b$

Una vegada hem estimat els paràmetres α' i β' per MQO podrem recuperar l'equació inicial.

8.8 Regressió lineal múltiple

El model estudiat en el tema de regressió lineal simple es pot generalitzar i s'hi pot incloure el cas en què tinguem una variable dependent Y i K variables independents X_1, X_2, ..., X_K. En el model de regressió lineal múltiple també s'han d'especificar una sèrie d'hipòtesis bàsiques sobre les components del model, que, en tot cas, són semblants a les especificades en el model de regressió lineal simple.

8.8.1 Hipòtesis bàsiques del model de regressió lineal múltiple

Respecte a l'equació. Existeix una relació lineal entre la variable dependent i les variables independents. Formalment escriurem:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_K X_{iK} + u_i \quad i = 1, 2, ..., n$$

on:

- Y_i: variable dependent
- X_{ik}: variable independent, $k = 1, ..., K$
- β_k: paràmetres, $k = 0, 1, 2, ..., K$
- u_i: terme de pertorbació

El subíndex i indica les observacions mostrals que tenim, i el subíndex k, la variable independent amb la qual estem treballant o el seu paràmetre corresponent.

Respecte a les variables independents. El nombre d'observacions ha de ser més gran que el nombre de paràmetres. En el cas de la regressió lineal múltiple, el nombre d'observacions ha de ser més gran que $K + 1$.

Respecte als paràmetres. β_0, β_1, ..., β_K són constants al llarg del mostreig. Aquestes són les constants que s'aproximaran mitjançant la inferència estadística.

8.8.2 Mesures de bondat d'ajustament en regressió lineal múltiple

L'inconvenient que presenta el coeficient de determinació r^2 al model de regressió lineal múltiple és que, a mesura que augmentem el nombre de variables independents al model, r^2 també augmenta. Ens podem trobar que en un model afegim variables que no tinguin res a veure amb la variable que es vol explicar i r^2 sigui més alt que el model que no contingui aquesta variable no significativa. Per tant, el coeficient que es proposa per mesurar la bondat de l'ajustament és el que s'anomena **coeficient de determinació corregit** $\bar{r}^2$, que es defineix com a:

$$\bar{r}^2 = 1 - \frac{n - 1}{n - (K + 1)} (1 - r^2)$$

Aquest coeficient també estarà entre 0 i 1 i només augmenta en el cas que la variable introduïda serveixi per explicar la variable dependent.

Entre dos models diferents triarem el que tingui un $\bar{r}^2$ més gran.

Per altra banda, l'**error estàndard** en el cas de regressió lineal múltiple es calcula mitjançant l'expressió següent:

$$S_u = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - (K + 1)}}$$

8.9 Contrastos de significació en regressió lineal múltiple

El model de regressió lineal ens ofereix la possibilitat de contrastar diverses hipòtesis sobre els coeficients. Podem fer:

- Contrastos coeficient a coeficient per estudiar la rellevància de cada variable independent per separat. Les hipòtesis que es contrasten són:
  - H_0: $\beta_i = 0$, el coeficient β_i no és significatiu, la variable X_i no influeix de manera lineal sobre la variable Y i no té sentit que estigui en el model.
  - H_1: $\beta_i \neq 0$, el coeficient β_i és significatiu, la variable X_i influeix de manera lineal sobre la variable Y i té sentit que estigui en el model.
- En el cas de la regressió lineal múltiple, un contrast conjunt per a tots els coeficients que ens servirà per estudiar la significació del model en conjunt. Les hipòtesis que es contrasten són:

- H_0: el model lineal no és significatiu; les variables X_i, en conjunt, no influeixen de manera lineal sobre la variable Y.
- H_1: el model lineal és significatiu; les variables X_i, en conjunt, influeixen de manera lineal sobre la variable Y.

Per fer aquests contrastos, partirem d'un exemple concret on es fabriquen tapes d'alumini a partir de motlles on es posa alumini líquid a certa pressió. Es mesura la temperatura de l'alumini líquid, la pressió amb què aquest s'injecta al motlle i l'índex de porositat trobat a les tapes finals d'alumini. Les dades són:

Temperatura (ºC)	*Pressió (kg/cm²)*	*Índex de porositat*
640	950	6.09
660	954	5.53
638	1005	6.78
662	997	6.16
651	976	5.93
653	972	6.12
647	977	5.92

La temperatura i la pressió són variables independents i l'índex de porositat és la variable dependent. És a dir, hem de trobar els coeficients β de l'expressió:

$$\text{Porositat} = \beta_0 + \beta_1 \cdot \text{Temperatura} + \beta_2 \cdot \text{Pressió}$$

8.9.1 Contrastos per a coeficients particulars

En aquest subapartat volem fer contrastos del tipus:

H_0: el coeficient β_i no és significatiu.

H_1: el coeficient β_i és significatiu.

L'estimació mínim quadràtica dels coeficients del model anterior es pot fer usant les tècniques de regressió lineal. El programa Excel dóna els resultats d'aquesta estimació.

$$\hat{\beta} = \begin{bmatrix} 8.3953 \\ -0.023 \\ 0.013 \end{bmatrix} \qquad \alpha_c = \begin{bmatrix} 0.2576 \\ 0.0429 \\ 0.0221 \end{bmatrix}$$

Això vol dir que els valors de $\hat{\beta}_0$, $\hat{\beta}_1$ i $\hat{\beta}_2$ són 8.3953, –0.023 i 0.013 i que els valors de α_c per fer els contrastos individuals de cada coeficient són 0.2576, 0.0429 i 0.0221, respectivament. Si nosaltres volem agafar un valor de $\alpha = 0.05$, la conclusió que s'obté és que:

- El coeficient β_0 no és significatiu perquè $\alpha_c = 0.2576 > \alpha = 0.05$.
- El coeficient β_1 és significatiu perquè $\alpha_c = 0.0429 < \alpha = 0.05$. Conclusió: la temperatura influeix de manera lineal en l'índex de porositat.
- El coeficient β_2 és significatiu perquè $\alpha_c = 0.0221 < \alpha = 0.05$. Conclusió: la pressió influeix de manera lineal en l'índex de porositat.

8.9.2 Contrast global

En els resultats obtinguts a l'Excel també es pot veure si el model de regressió lineal en conjunt és significatiu, és a dir, si les variables independents usades en el model serveixen per explicar el comportament de la variable resposta. En aquest cas, el contrast que fem és:

- H_0: el model lineal no és significatiu.
- H_1: el model lineal és significatiu.

L'estadístic que s'usa per fer aquest contrast és l'estadístic F de la taula de l'anàlisi de la variància. En el nostre exemple, aquest estadístic té un valor d'11.6 amb un $\alpha_c = 0.0216$. Per tant, podem afirmar que el model de regressió lineal és significatiu en el conjunt de les variables independents, ja que $\alpha_c = 0.0216 < \alpha = 0.05$.

8.10 Resultats amb el programa Excel

Avui dia tenim programes informàtics que ens estalvien la feina de fer els càlculs per estimar els diferents paràmetres i mesures que ens apareixen quan volem fer estudis de regressió.

Presentarem les pantalles que apareixen al programa Excel quan fem regressió lineal simple i regressió lineal múltiple. Per fer regressió a Excel es pot anar a "Herramientas" → "Análisis de datos" → "Regresión".

Apareix un quadre on:

- Hem de triar les dades corresponents a la variable dependent *Y*.
- Hem de triar les dades corresponents a les variables independents (poden ser una o més d'una).
- Hem de marcar "Rótulos" si hem incorporat els rètols a les dades d'entrada.

Altres opcions ens permeten:

- Obligar que la recta ajustada passi pel punt (0,0) ("Constante igual a 0").
- Determinar un nivell de confiança per quan es fan intervals de confiança dels paràmetres de la recta de regressió ("Nivel de confianza").
- Mostrar els residus (ja sigui numèricament o a través d'un gràfic).
- Posar en un gràfic els valors pronosticats pel model i els valors reals de la variable dependent en funció de cada variable independent.
- Fer un gràfic de probabilitat normal amb els valors de la variable dependent.

8.10.1 Regressió lineal simple amb Excel

La pantalla que apareix quan treballem amb les dades de l'exemple dels pesos i les alçades és:

Estadísticas de la regresión					
Coeficiente de correlación	0.97295688				
Coeficiente de determinación	0.94664509				
R^2 ajustado	0.9377526				
Error típico	2.06789119				
Observaciones	8				
Análisis de varianza					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	455.217956	455.217956	106.454498	4.8446E-05
Residuos	6	25.6570439	4.27617398		
Total	7	480.875			
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	
Intercepción	111.318707	5.62612068	19.7860503	1.081E-06	
Pes	0.83718245	0.08114058	10.3176789	4.8446E-05	

Al començament apareix el valor del coeficient de correlació r (0.97295688), el valor del coeficient de determinació r^2 (0.94664509), el coeficient de determinació corregit $\bar{r}^2$ (0.9377526) i el valor de l'error estàndard S_u (2.0678119).

A la columna "Coeficientes" trobem el resultat de $\hat{\alpha}$ (correspon a "Intercepción") i el de $\hat{\beta}$ (correspon a "Pes").

A la columna "Probabilidad" apareixen els p-valors per contrastar si l'ordenada a l'origen i el pendent són significativament diferents de zero o no, respectivament. En aquest cas, com que tant el p-valor per fer el contrast sobre l'ordenada a l'origen, que és 1.081E-06, com el p-valor per fer el contrast sobre el pendent, que és 4.8446E-05, són més petits que 0.05, podem afirmar que tant l'ordenada a l'origen com el pendent són significativament diferents de zero.

8.10.2 Regressió lineal múltiple amb Excel

Per fer els càlculs quan treballem amb regressió lineal múltiple és quasi imprescindible treballar amb els ordinadors, ja que els càlculs es compliquen molt. Aquí presentem els resultats obtinguts quan agafem l'exemple de l'índex de porositat.

El model que ajustem és:

$$\text{Porositat} = \beta_0 + \beta_1 \cdot \text{Temperatura} + \beta_2 \cdot \text{Pressió}$$

La pantalla que ens apareix és la següent:

Estadísticas de la regresión					
Coeficiente de correlación	0.92355277				
Coeficiente de determinación	0.85294971				
R^2 ajustado	0.77942457				
Error típico	0.17662295				
Observaciones	7				
Análisis de varianza					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	0.72378876	0.36189438	11.6007894	0.02162379
Residuos	4	0.12478267	0.03119567		
Total	6	0.84857143			
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	
Intercepción	8.39525622	6.36515374	1.31894005	0.25762905	
Temperatura	–0.02301924	0.00786106	–2.92826328	0.04288807	
Pressió	0.01295912	0.00356909	3.63093621	0.02214152	

Al començament trobem el valor del coeficient de correlació r (0.92355277), el coeficient de determinació r^2 (0.85294971), el coeficient de determinació corregit $\bar{r}^2$ (0.77942457) i l'error estàndard S_u (0.17662295).

A la columna "Coeficientes" trobem l'estimació dels paràmetres: $\hat{\beta}_0 = 8.39525622$, $\hat{\beta}_1 = -0.02301924$, $\hat{\beta}_2 = 0.01295912$.

A la columna "Probabilidad" trobem el nivell de significació crític que resulta de contrastar si la variable independent corresponent és significativa, és a dir, si la variable independent aporta alguna cosa al model lineal per explicar el comportament de la variable Y. Com més petit sigui el valor del nivell de significació crític, més significativa és la variable independent. En aquest cas, com que tots els p-valors són més petits que 0.05, a banda del corresponent al terme independent, podem dir que les dues variables, per separat, del model són significatives.

A la columna "Valor crítico de F" de l'anàlisi de la variància trobem el nivell de significació crític 0.02162379, que resulta de contrastar si les variables independents en conjunt són significatives; en aquest cas, podem afirmar que, agafant un nivell de significació de 0.05, les variables explicatives en conjunt són significatives per explicar de manera lineal el comportament de la variable Y.

8.10.3 Regressió no lineal simple amb Excel

La regressió no lineal simple amb Excel la farem a partir del gràfic de les dades que volem treballar. Per tant, primer hem de fer un gràfic de dispersió. Agafant l'exemple dels pesos i les alçades, els passos per fer el gràfic són:

1. Posem les dades en dues columnes de manera que les dades de la variable independent, Pes, ocupin una columna i les de la variable dependent, Alçada, ocupin la columna del costat dret.
2. Seleccionem les dues columnes de dades.
3. Anem a "Insertar" → "Gráfico".
4. Triem "XY (Dispersión)". Com a "Subtipo", triem el gràfic de dalt. Cliquem sobre "Siguiente".
5. Si hem seleccionat bé les dades, ens surt una previsualització de com quedarà el gràfic i no hem de tocar cap opció. Cliquem sobre "Siguiente".
6. En aquesta pantalla surt un quadre de diàleg amb diverses pestanyes on hi ha diverses opcions del gràfic, com ara posar títol al gràfic i als eixos o mostrar una llegenda per al gràfic. Per ara, podem deixar les coses com estan i cliquem sobre "Siguiente".

7. Per no tenir problemes amb la mida del gràfic, triem "En una hoja nueva" i cliquem sobre "Finalizar".
8. Apareix una fulla nova amb el gràfic de dispersió de les dades on tenim, a l'eix d'abscisses, la variable Pes i, a l'eix d'ordenades, la variable Alçada.

A partir d'aquí podem ajustar diversos models a les dades. Nosaltres ajustarem els següents:

a) Lineal: $y = ax + b$.

b) Quadràtic: $y = ax^2 + bx + c$.

c) Logarítmic: $y = a\ ln\ x + b$.

d) Potencial: $y = ax^b$.

e) Exponencial: $y = ae^{bx}$.

Els passos que hem de seguir per fer els diversos ajustos a partir del gràfic obtingut són:

1. Cliquem, amb el botó dret, sobre una de les dades representades. Del menú que apareix, triem "Agregar línea de tendencia".
2. A la pestanya "Tipo" triem el model que volem ajustar. Si triem el model polinomial, també hem d'indicar el grau del polinomi que hi volem ajustar. Per seguir amb el nostre exemple, triem lineal.
3. A la pestanya "Opciones", marquem les caselles "Presentar ecuación en el gráfico" i "Presentar el valor R cuadrado en el gráfico". D'aquesta manera ens apareixerà l'equació ajustada al gràfic i el seu coeficient de determinació r^2. Cliquem sobre "Aceptar".
4. Al gràfic ens ha d'aparèixer la funció ajustada $y = 0.8372x + 111.32$ i un valor del coeficient de determinació de $r^2 = 0.9466$.

Els passos anteriors els repetiríem per a les altres funcions que volem ajustar i podem posar els diversos resultats obtinguts en una taula com la següent:

Model	*Funció*	r^2
Lineal	$y = 0.8372x + 111.32$	0.9466
Quadràtic	$y = 0.0054x^2 + 0.1067x + 135.65$	0.9485
Logarítmic	$y = 55.988 \ln x - 67.484$	0.9401
Potencial	$y = 41.371x^{0.3329}$	0.9464
Exponencial	$y = 119.84e^{0.005x}$	0.9513

Com que els models no tenen el mateix nombre de paràmetres per estimar, per decidir quin és el millor model que ajusta les dades, hem de calcular el coeficient de determinació ajustat i triar el model que tingui un valor més gran. Per fer-ho, hem de tenir en compte que en tots els models el valor de $K + 1$ és 2, perquè hi ha dos paràmetres per estimar, menys en el model quadràtic, on el valor de $K + 1$ és 3, perquè hi ha tres paràmetres per estimar. Per tant, la taula anterior es pot completar amb la columna dels valors calculats del coeficient de determinació ajustat:

Model	*Funció*	r^2	$\bar{r}^2$
Lineal	$y = 0.8372x + 111.32$	0.9466	0.9377
Quadràtic	$y = 0.0054x^2 + 0.1067x + 135.65$	0.9485	0.9279
Logarítmic	$y = 55.988 \ln x - 67.484$	0.9401	0.9301
Potencial	$y = 41.371x^{0.3329}$	0.9464	0.9375
Exponencial	$y = 119.84e^{0.005x}$	0.9513	0.9432

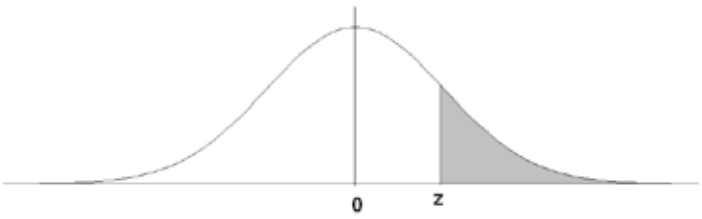
El model que té un valor $\bar{r}^2$ més gran és el model exponencial i considerem que aquest és el model que millor ajusta les dades.

Observació. Si volem ajustar un model no lineal diferent dels que apareixen en les opcions gràfiques d'Excel i aquest model és linealitzable mitjançant alguna transformació de les variables originals, caldrà fer aquesta transformació i, amb les variables transformades, seguir els passos que s'indiquen a l'apartat "Regressió lineal simple amb Excel". Posteriorment, i fent les operacions adequades a partir dels resultats obtinguts amb Excel, podrem recuperar l'equació no lineal inicial.

Taules estadístiques

TAULA NORMAL ESTÀNDARD (àrees a la dreta del punt)

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000



TAULA KHI QUADRAT

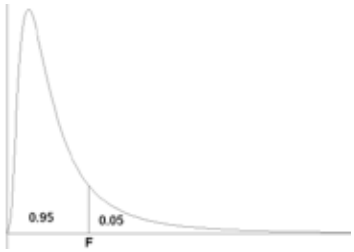
		àrees a la dreta del punt												
		0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
g. l.	1	0.0	0.0	0.0	0.0	0.0	0.1	0.5	1.3	2.7	3.8	5.0	6.6	7.9
	2	0.0	0.0	0.1	0.1	0.2	0.6	1.4	2.8	4.6	6.0	7.4	9.2	10.6
	3	0.1	0.1	0.2	0.4	0.6	1.2	2.4	4.1	6.3	7.8	9.3	11.3	12.8
	4	0.2	0.3	0.5	0.7	1.1	1.9	3.4	5.4	7.8	9.5	11.1	13.3	14.9
	5	0.4	0.6	0.8	1.1	1.6	2.7	4.4	6.6	9.2	11.1	12.8	15.1	16.7
	6	0.7	0.9	1.2	1.6	2.2	3.5	5.3	7.8	10.6	12.6	14.4	16.8	18.5
	7	1.0	1.2	1.7	2.2	2.8	4.3	6.3	9.0	12.0	14.1	16.0	18.5	20.3
	8	1.3	1.6	2.2	2.7	3.5	5.1	7.3	10.2	13.4	15.5	17.5	20.1	22.0
	9	1.7	2.1	2.7	3.3	4.2	5.9	8.3	11.4	14.7	16.9	19.0	21.7	23.6
	10	2.2	2.6	3.2	3.9	4.9	6.7	9.3	12.5	16.0	18.3	20.5	23.2	25.2
	11	2.6	3.1	3.8	4.6	5.6	7.6	10.3	13.7	17.3	19.7	21.9	24.7	26.8
	12	3.1	3.6	4.4	5.2	6.3	8.4	11.3	14.8	18.5	21.0	23.3	26.2	28.3
	13	3.6	4.1	5.0	5.9	7.0	9.3	12.3	16.0	19.8	22.4	24.7	27.7	29.8
	14	4.1	4.7	5.6	6.6	7.8	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
	15	4.6	5.2	6.3	7.3	8.5	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
	16	5.1	5.8	6.9	8.0	9.3	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
	17	5.7	6.4	7.6	8.7	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
	18	6.3	7.0	8.2	9.4	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
	19	6.8	7.6	8.9	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
	20	7.4	8.3	9.6	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
	21	8.0	8.9	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
	22	8.6	9.5	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
	23	9.3	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
	24	9.9	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
	25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9
	26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3
	27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6
	28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0
	29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3
	30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7
	40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8
	50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5
	60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0
	70	43.3	45.4	48.8	51.7	55.3	61.7	69.3	77.6	85.5	90.5	95.0	100.4	104.2
	80	51.2	53.5	57.2	60.4	64.3	71.1	79.3	88.1	96.6	101.9	106.6	112.3	116.3
	90	59.2	61.8	65.6	69.1	73.3	80.6	89.3	98.6	107.6	113.1	118.1	124.1	128.3
	100	67.3	70.1	74.2	77.9	82.4	90.1	99.3	109.1	118.5	124.3	129.6	135.8	140.2

TAULA t DE STUDENT

		àrees a la dreta del punt									
		0.45	0.4	0.3	0.25	0.2	0.1	0.05	0.025	0.01	0.005
g. l.	1	0.16	0.32	0.73	1.00	1.38	3.08	6.31	12.71	31.82	63.66
	2	0.14	0.29	0.62	0.82	1.06	1.89	2.92	4.30	6.96	9.92
	3	0.14	0.28	0.58	0.76	0.98	1.64	2.35	3.18	4.54	5.84
	4	0.13	0.27	0.57	0.74	0.94	1.53	2.13	2.78	3.75	4.60
	5	0.13	0.27	0.56	0.73	0.92	1.48	2.02	2.57	3.36	4.03
	6	0.13	0.26	0.55	0.72	0.91	1.44	1.94	2.45	3.14	3.71
	7	0.13	0.26	0.55	0.71	0.90	1.41	1.89	2.36	3.00	3.50
	8	0.13	0.26	0.55	0.71	0.89	1.40	1.86	2.31	2.90	3.36
	9	0.13	0.26	0.54	0.70	0.88	1.38	1.83	2.26	2.82	3.25
	10	0.13	0.26	0.54	0.70	0.88	1.37	1.81	2.23	2.76	3.17
	11	0.13	0.26	0.54	0.70	0.88	1.36	1.80	2.20	2.72	3.11
	12	0.13	0.26	0.54	0.70	0.87	1.36	1.78	2.18	2.68	3.05
	13	0.13	0.26	0.54	0.69	0.87	1.35	1.77	2.16	2.65	3.01
	14	0.13	0.26	0.54	0.69	0.87	1.35	1.76	2.14	2.62	2.98
	15	0.13	0.26	0.54	0.69	0.87	1.34	1.75	2.13	2.60	2.95
	16	0.13	0.26	0.54	0.69	0.86	1.34	1.75	2.12	2.58	2.92
	17	0.13	0.26	0.53	0.69	0.86	1.33	1.74	2.11	2.57	2.90
	18	0.13	0.26	0.53	0.69	0.86	1.33	1.73	2.10	2.55	2.88
	19	0.13	0.26	0.53	0.69	0.86	1.33	1.73	2.09	2.54	2.86
	20	0.13	0.26	0.53	0.69	0.86	1.33	1.72	2.09	2.53	2.85
	21	0.13	0.26	0.53	0.69	0.86	1.32	1.72	2.08	2.52	2.83
	22	0.13	0.26	0.53	0.69	0.86	1.32	1.72	2.07	2.51	2.82
	23	0.13	0.26	0.53	0.69	0.86	1.32	1.71	2.07	2.50	2.81
	24	0.13	0.26	0.53	0.68	0.86	1.32	1.71	2.06	2.49	2.80
	25	0.13	0.26	0.53	0.68	0.86	1.32	1.71	2.06	2.49	2.79
	26	0.13	0.26	0.53	0.68	0.86	1.31	1.71	2.06	2.48	2.78
	27	0.13	0.26	0.53	0.68	0.86	1.31	1.70	2.05	2.47	2.77
	28	0.13	0.26	0.53	0.68	0.85	1.31	1.70	2.05	2.47	2.76
	29	0.13	0.26	0.53	0.68	0.85	1.31	1.70	2.05	2.46	2.76
	30	0.13	0.26	0.53	0.68	0.85	1.31	1.70	2.04	2.46	2.75
	40	0.13	0.26	0.53	0.68	0.85	1.30	1.68	2.02	2.42	2.70
	60	0.13	0.25	0.53	0.68	0.85	1.30	1.67	2.00	2.39	2.66
	120	0.13	0.25	0.53	0.68	0.84	1.29	1.66	1.98	2.36	2.62
	∞	0.13	0.25	0.52	0.67	0.84	1.28	1.64	1.96	2.33	2.58

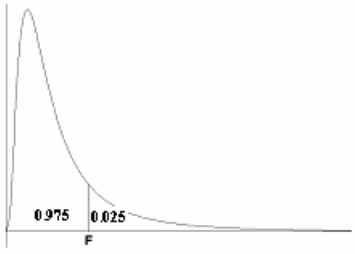
TAULA F (àrea de 0.05 a la dreta)

g. l. 2 \ g. l. 1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	199	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.10



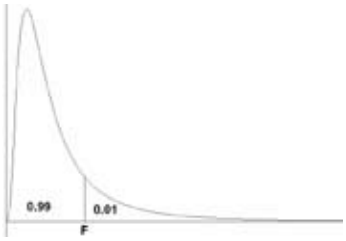
TAULA F (àrea de 0.025 a la dreta)

		g.l. 1																			
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
g.l. 2	1	648	799	864	900	922	937	948	957	963	969	977	985	993	997	1001	1006	1010	1014	1018
	2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5	39.5
	3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4	14.3	14.3	14.2	14.1	14.1	14.0	14.0	13.9	13.9
	4	12.2	10.6	10.0	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.86
	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	4.15
	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.68
	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.34
	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.09
	11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.89
	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.73
	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.50
	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
	16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
	17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.26
	18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.20
	19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.14
	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
	21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.05
	22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.01
	23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.98
	24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
	25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
	26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.89
	27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.86
	28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.84
	29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.82
	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.80
	40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.65
	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.49
	120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.33
	∞	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06	1.96	1.85	1.72	1.65	1.58	1.50	1.41	1.29	1.13



TAULA F (àrea de 0.01 a la dreta)

		g.l. 1																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
g.l. 2	1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6157	6209	6234	6260	6286	6313	6340	6363
	2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
	3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
	4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
	5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.03
	6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.89
	7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.66
	8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.87
	9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.32
	10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.92
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.61
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.37
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.18
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.02
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.88
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.76
	17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.66
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.58
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.50
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.43
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.37
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.32
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.27
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.22
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.18
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.14
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.11
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.08
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.05
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.02
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.82
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.62
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.40
	∞	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34	2.20	2.06	1.90	1.81	1.72	1.61	1.50	1.35	1.16



TAULA DE KOLMOGOROV-SMIRNOV

n	α = 0.1	α = 0.05	α = 0.01
1	0.95	0.975	0.995
2	0.776	0.842	0.929
3	0.636	0.708	0.829
4	0.565	0.624	0.734
5	0.509	0.563	0.669
6	0.468	0.519	0.617
7	0.436	0.483	0.576
8	0.41	0.454	0.542
9	0.387	0.43	0.513
10	0.369	0.409	0.489
11	0.352	0.0391	0.468
12	0.338	0.375	0.449
13	0.325	0.361	0.432
14	0.314	0.349	0.418
15	0.304	0.338	0.404
16	0.295	0.327	0.392
17	0.286	0.318	0.381
18	0.279	0.309	0.371
19	0.271	0.301	0.361
20	0.265	0.294	0.352
21	0.259	0.287	0.344
22	0.253	0.281	0.337
23	0.247	0.275	0.33
24	0.242	0.269	0.323
25	0.238	0.264	0.317
26	0.233	0.259	0.311
27	0.229	0.254	0.305
28	0.225	0.25	0.3
29	0.221	0.246	0.295
30	0.218	0.242	0.29
31	0.214	0.238	0.285
32	0.211	0.234	0.281
33	0.208	0.231	0.277
34	0.205	0.227	0.273
35	0.202	0.224	0.269
36	0.199	0.221	0.265
37	0.196	0.218	0.262
38	0.194	0.215	0.258
39	0.191	0.213	0.255
40	0.189	0.21	0.252
>40	$\frac{1.224}{\sqrt{n}}$	$\frac{1.358}{\sqrt{n}}$	$\frac{1.628}{\sqrt{n}}$

TAULA DE KOLMOGOROV-SMIRNOV-LILLIEFORS

n	α = 0.1	α = 0.05	α = 0.01
1	0.445	0.486	0.562
2	0.408	0.446	0.516
3	0.37	0.404	0.468
4	0.339	0.371	0.429
5	0.314	0.343	0.397
6	0.294	0.321	0.371
7	0.277	0.303	0.35
8	0.263	0.287	0.332
9	0.25	0.273	0.316
10	0.239	0.262	0.303
11	0.23	0.251	0.29
12	0.221	0.242	0.28
13	0.214	0.234	0.27
14	0.207	0.226	0.261
15	0.201	0.219	0.254
16	0.195	0.213	0.245
17	0.19	0.207	0.24
18	0.185	0.202	0.233
19	0.18	0.197	0.228
20	0.176	0.192	0.222
21	0.172	0.188	0.218
22	0.168	0.184	0.213
23	0.165	0.18	0.209
24	0.162	0.177	0.204
25	0.159	0.173	0.201
26	0.156	0.17	0.197
27	0.153	0.167	0.193
28	0.15	0.164	0.19
29	0.148	0.162	0.187
30	0.146	0.159	0.184
31	0.143	0.157	0.181
32	0.141	0.154	0.179
33	0.139	0.152	0.176
34	0.137	0.15	0.173
35	0.135	0.148	0.171
36	0.134	0.146	0.169
37	0.132	0.144	0.167
38	0.13	0.142	0.164
39	0.129	0.14	0.162
40	0.127	0.139	0.16
>40	$\frac{0.805}{\sqrt{n}}$	$\frac{0.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

L'objectiu principal d'aquest material és proporcionar als estudiants universitaris una eina que permeti consultar clarament i ràpidament com es pot aplicar de manera pràctica la teoria relacionada amb qualsevol dels continguts que formen part d'un curs bàsic d'estadística de nivell universitari. Per tal d'afavorir l'aprenentatge, s'explica pas a pas, i usant exemples, quan i com es poden aplicar aquestes tècniques estadístiques. També es comenta quines són les funcions i les eines estadístiques d'Excel.