Estadística aplicada a les ciències humanes i socials

Michael Greenacre

Amb la col·laboració de:
Anna Espinal
Jan Graffelman

PID_00147628

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índex

1. Què és l'estadística?

Estadística

El seu nom deriva de la paraula *estat*. Durant el segle XIX l'estadística era considerada com la **ciència de l'estat**. Després va depassar aquest àmbit i va adquirir una aplicació més universal.

De fet, és cert que l'estadística penetra en gairebé tots els aspectes de la nostra vida i es pot usar per a aconseguir una interpretació millor de tots aquells fenòmens que observem. En aquesta unitat introductòria veureu nou aplicacions diferents de l'estadística a problemes de meteorologia, medicina, ciències ambientals, estudis socials, recerca espacial, dret i benestar social.

En aquest apartat introductori aprendreu:

- què són les dades numèriques i les dades categòriques;
- què és una variable estadística;
- cómo se identifican el objetivo, las variables y los datos en un proyecto de investigación.

L'estadística mira les dades

En les ciències humanes i socials,...

.. les dades categòriques tenen un paper essencial perquè els fenòmens sociològics són difícils de mesurar quantitativament.

Les dades normalment són **numèriques**; per exemple, l'alçada d'una criatura o el preu d'una acció a la borsa. Però les dades també poden ser **categòriques**; per exemple, l'observació que una persona hagi tingut un atac de cor o no, o la província (comarca) on una persona hagi nascut.

L'objectiu de mirar dades

Per exemple:

- En la recollida de dades sobre el primer llamp que cau, el meteoròleg o la meteoròloga vol entendre a quina hora del dia és més probable que caigui un llamp, i l'estudi proposa de millorar la preparació per als perills d'un llamp.
- En recollir dades sobre l'alçada d'una criatura, el metge o la metgessa vol determinar el ritme de creixement d'un infant i comprovar que és normal.

Dades observades o dades creades mitjançant l'experimentació

1) Per una banda, simplement s'observen les dades tal com s'esdevenen naturalment; per exemple, cau un llamp i nosaltres observem l'hora en què cau el primer, o observem el nombre de morts d'aquella estranya criatura anomenada *manatí* alhora que el nombre de matrícules d'embarcacions.

2) Una manera alternativa de recollir dades és mitjançant un procediment més significatiu anomenat **experimentació**. Per exemple, en l'estudi de l'aspirina nosaltres no estudiem 20.000 persones i observem simplement quines tenen atacs de cor i quines han pres aspirina per veure si hi ha una connexió, com en l'estudi del manatí. En aquest cas s'ha dividit la gent, que en concret són tots metges i metgesses, en dos grups per un procés d'atzar (com ara a cara o creu) i després s'ha determinat que un grup prengui aspirina i l'altre, no. Podem fer experiments d'aquesta mena en comptades ocasions, però són més convincents a l'hora de poder demostrar resultats de debò.

Experimentació

No va ser fins al cap de molts anys d'observació i recerca que es va demostrar una connexió entre el càncer de pulmó i l'hàbit de fumar, però s'hauria pogut demostrar molt abans si haguéssim pogut fer experiments amb persones en els quals s'hagués demanat a algunes que fumessin durant un llarg període de temps i a altres que no ho fessin.

Les dades són les observacions sobre variables

Quan observem dades, mirem les diferents manifestacions d'una o més variables. Per exemple, l'alçada d'una criatura és una variable, mentre que les quantitats 95 cm, 83 cm i 88 cm són dades sobre aquesta variable. Les dades sobre la variable "comarca de Catalunya" podrien ser Barcelonès, Alt Empordà, Bages, etc. Aquesta variable l'anomenem **variable categòrica**. Sovint es representa una **variable** algebraicament amb una lletra majúscula, per exemple X, mentre que les **dades sobre una variable** es representen amb lletres minúscules, per exemple x_1, x_2, x_3. Per tant, podríem dir:

$$X = \text{alçada d'una criatura}$$

amb algunes observacions que poden ser $x_1 = 95$, $x_2 = 83$ i $x_3 = 88$ o, per a una variable categòrica:

$$Y = \text{comarca de Catalunya}$$

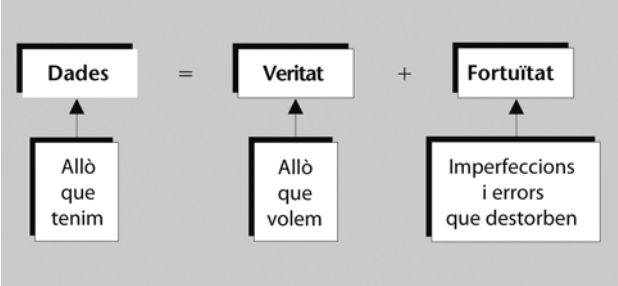
amb algunes observacions com, per exemple, y_1 = Barcelonès, y_2 = Alt Empordà i y_3 = Bages.

L'estadística descriu i analitza les dades

L'estadística s'usa per a descriure i analitzar les dades. Per exemple, en l'estudi del creixement d'una criatura, s'observen dues variables: l'alçada de la nena i l'edat. Es representen les dades de l'alçada contra les de l'edat en allò que anomenem un **diagrama de dispersió**. Això és una descripció de les dades, una descripció visual, de fet. Però amb uns estudis previs, els metges i les metgesses han establert un ritme de creixement normal per a una criatura, i això se superposa en el gràfic. Per mitjà del gràfic el metge o la metgessa dedueix ara que hi ha una alta probabilitat que la nena no creixi prou ràpid. És una anàlisi de les dades, i una anàlisi porta a una conclusió.

L'anàlisi estadística prova de separar la veritat de la fortuïtat

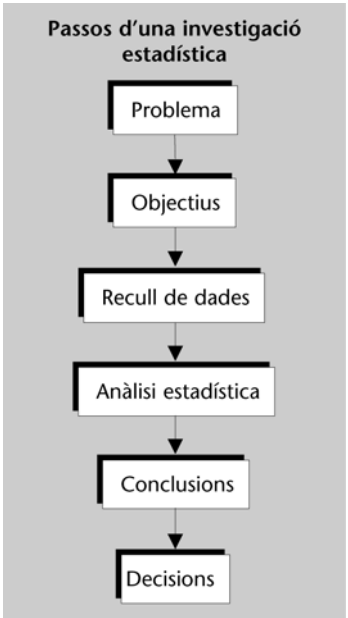
Les dades que observem no són perfectes –hi pot haver tota mena d'errors–. Si podíem preguntar una a una a totes les persones de Catalunya si treballen o no, aleshores tindríem una mesura perfecta del nivell d'ocupació. Però hem de recórrer a preguntar-ho a una mostra de la població, la qual cosa vol dir que les nostres dades no seran perfectes. Totes les dades consten d'un element de **veritat** i un element d'error que nosaltres anomenem **fortuïtat**, és a dir, un element que és imprevisible i fora del nostre control:



L'**anàlisi estadística** té el propòsit de separar la veritat de la fortuïtat de manera que puguem treure conclusions en ferm d'allò que observem. Aquest és un tema recurrent en aquesta assignatura i del qual parlarem sovint.

Els passos en una investigació estadística

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



1) En primer lloc, hi ha la definició d'un problema i els seus objectius.

2) En segon lloc, es recullen les dades de les variables rellevants.

3) En tercer lloc, es descriuen i possiblement s'analitzen les dades, cosa que porta a una conclusió pel que fa a l'objectiu de l'estudi.

Aquesta assignatura tracta principalment de la tercera part: la descripció i l'anàlisi de les dades dirigides a prendre decisions.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Variable: característica o atribut que pren valors potencials molt diversos, per exemple: el nivell socioeconòmic, l'alçada, els ingressos, un partit polític. D'acord amb els valors que una variable pot tenir podem parlar d'una variable numèrica (o contínua o quantitativa) i una variable categòrica (o discreta o qualitativa).

Dades numèriques: dades expressades en una escala contínua, per exemple: 3,5 anys; 210.000 pessetes; 21,44 grams.

Dades categòriques: dades que indiquen una categoria, grup o classe, per exemple: cabells negres, dona, amb feina, grup de tractament.

2. La descripció d'una variable numèrica: gràfics de tiges i fulles, i histogrames

El primer pas per a comprendre les dades numèriques és organitzar-les d'una manera que faci el dibuix global més clar. Les dues maneres d'organitzar i resumir dades: el gràfic de tiges i fulles, i l'histograma. Totes dues són mètodes per a visualitzar les observacions d'una variable única.

En aquest apartat aprendreu:

- què és la distribució d'una variable;
- com es visualitza la distribució d'una variable numèrica fent servir un gràfic de tiges i fulles, o un histograma;
- la diferència entre les distribucions simètriques i les asimètriques;
- com s'identifiquen valors insòlits, o allunyats, en les dades.

La distribució d'una variable

Imaginem-nos que tenim diverses observacions d'una variable. La **distribució** d'aquestes dades és el perfil dels valors observats, per exemple: quin és el valor més petit, quin és el més alt i on són més freqüents els valors. La idea d'una distribució suggereix que hauríem de provar de resumir el perfil dels valors en un quadre.

La freqüència és...

... el nombre de vegades que es repeteix una observació o valor determinat en el conjunt de les dades.

El gràfic de tiges i fulles

Una manera senzilla de visualitzar la distribució d'una variable numèrica és dibuixar un gràfic de tiges i fulles.

1) El primer pas és ordenar les dades de més petita a més gran. Això ho anomenem **classificar dades en ordre ascendent**.

2) Després, segons l'escala de valors, hem de triar quina part dels valors serà la **tija** i quina la **fulla**.

Com es dibuixa un gràfic de tiges i fulles

Per a dibuixar un gràfic de tiges i fulles cal seguir els quatre passos següents:

El primer pas...

... és opcional; treballar amb les dades classificades és una mica més fàcil, però també podem treballar amb les dades en l'ordre original.

1) Classificar les dades en ordre ascendent.

2) Decidir quina part dels valors és la tija i quina la fulla, arrodonint els valors si cal.

3) Escriure les tiges una sota l'altra en ordre ascendent i dibuixar una ratlla vertical al costat per a separar les tiges de les fulles.

4) Escriure totes les fulles al costat de cada tija, una per a cada un dels valors de les dades.

Per a cada tija, les fulles s'haurien de trobar també en un ordre ascendent. Si hem ordenat les dades en el primer pas, les fulles estaran ordenades. Si fem servir les dades en l'ordre original, aleshores hem de classificar cada ratlla de fulles separadament.

Alguns exemples de gràfics de tiges i fulles

Mireu l'escala sencera de la variable a més del nombre de valors disponibles a l'hora de decidir sobre la tija i les fulles. Per exemple, considereu aquest model dels coeficients d'intel·ligència de seixanta estudiants:

Taula I

120	101	118	116	108	96
110	102	115	103	91	88
107	94	104	97	95	101
103	105	100	94	124	90
106	107	106	98	96	100
87	112	95	98	103	89
119	96	90	104	105	125
110	98	102	108	98	131
85	104	93	93	94	87
97	100	92	89	100	96

Coeficients d'intel·ligència dels estudiants.

L'escala d'aquests 60 valors va de 85 a 131. D'aquesta manera podríem fer un gràfic de tiges i fulles en què les centenes i les desenes fossin la tija, i les unitats, les fulles:

Taula II

```
 8 |577899
 9 |00123344455666677888
10 |00001122333444556667788
11 |0025689
12 |045
13 |1
```

Això ens dóna 6 ratlles en el gràfic de tiges i fulles. Ja que tenim un bon nombre d'observacions, podríem incrementar el nombre de ratlles i d'aquesta manera aconseguir un gràfic de tiges i fulles més detallat simplement fent que cada ratlla es correspongui amb cinc valors potencials del coeficient d'intel·ligència en comptes de fer-ho amb deu: 85-89, 90-94, 95-99, 100-104, etc. D'aquesta manera cada tija es podria fer servir dos cops:

Taula III

```
 8 |577899
 9 |001233444
 9 |55666677888
10 |00001122333444
10 |556667788
11 |002
11 |5689
12 |04
12 |5
13 |1
```

El centre d'una distribució

En un conjunt de valors que formin una distribució podem identificar un valor que més o menys és el centre de la distribució, un valor que té aproximadament la meitat de les observacions a sota i l'altra meitat a sobre. Precisarem aquesta idea més endavant.

Vegeu l'apartat 3 d'aquesta assignatura.

Activitats

1. Una enquesta de professorat universitari inclou informació sobre els ingressos. Els ingressos que tenen els 50 professors i professores enquestats són els següents (en milions de pessetes):

4.8	5.2	4.0	6.2	4.3	5.5	5.4	6.5	7.1	4.5
3.9	4.2	4.4	6.1	4.6	5.0	4.8	6.2	4.6	4.2
5.3	4.4	4.4	4.8	3.8	4.7	4.2	5.2	4.9	3.9
5.1	6.3	4.3	4.7	5.4	5.2	5.7	4.5	4.2	4.1
4.8	4.2	4.3	4.0	5.0	4.2	4.7	5.3	4.5	4.0

Feu un gràfic de tiges i fulles d'aquestes dades i comenteu-ne el resultat.

Ara ens fixarem en una manera alternativa de visualitzar la distribució d'una variable, anomenada **histograma**. Es poden fer servir els histogrames per a un gran nombre d'observacions quan no és necessari veure els valors individuals d'una manera detallada, com en el gràfic de tiges i fulles, sinó que simplement es vol veure l'aspecte de conjunt de la distribució.

La freqüència d'una classe...

... és el nombre d'observacions o valors de la variable que estan compresos entre els límits inferior i superior de la classe.

L'altura de cada barra representa la freqüència de cada classe.

Histogrames de grans conjunts de dades

Un histograma és semblant a un gràfic de tiges i fulles, però no mostra els valors individuals de les fulles. En canvi, s'hi dibuixa una barra vertical per mostrar el nombre de valors en les nostres dades que es troben dins cada classe de l'histograma. Per aquesta raó, els histogrames són molt més convenients quan es treballa amb un gran nombre de valors en les dades.

La tria de classes en els histogrames

Les classes d'un histograma, com les classes d'un gràfic de tiges i fulles, cobreixen tota l'escala de valors de la variable. A l'hora de decidir les classes per a un gràfic de tiges i fulles esteu limitats pel tipus de valors que teniu; per exemple, per a les dades que són edats sovint triaríeu les desenes per al tronc, de manera que les classes serien 0-9, 10-19, 20-29, etc. A l'hora de decidir les classes d'un histograma teniu més llibertat; per exemple, podríeu triar les classes 6-15, 16-25, 26-35, etc. No obstant això, hi ha dues consideracions importants a fer a l'hora de triar les classes d'un histograma:

L'amplitud o longitud de classe és l'extensió d'un interval de classe.

1) Totes les classes haurien de tenir la mateixa amplada. Per exemple, per a les dades sobre l'edat no trieu classes com ara 0-20, 21-30, 31-45, etc. Penseu detingudament la definició de les classes, per exemple: en algunes mesures les observacions pot ser que sovint tendeixin a ser nombres sencers, posem per

cas 11,0, 15,0, 13,0, etc. En aquest cas seria millor triar els límits de les classes en valors com ara 10,5, 11,5, 12,5, etc., de manera que les classes siguin 10,5-11,5, 11,5-12,5, 12,5-13,5, etc. Normalment, quan definim classes per mitjà d'aquests intervals, la classe inclou el valor superior de l'interval: així un valor d'11,5 s'inclouria en la classe 10,5-11,5 i no en l'11,5-12,5.

2) El nombre de classes depèn de la quantitat de dades que tingueu i el detall que us interessi veure de la distribució. Aquesta qüestió es resol amb seny i sentit comú, i no hi ha cap regla per a fer-ho.

Penseu-hi

Què passaria si el nombre de classes triat fos molt gran? I si fos molt petit?

Patrons en l'histograma

Com als gràfics de tiges i fulles, mireu els patrons generals de l'histograma i després busqueu desviacions d'aquests patrons: les quantitats petites de valors que se separen de la distribució s'anomenen **valors allunyats o insòlits**. Si l'histograma no és simètric, diem que hi ha una asimetria. La part llarga i arrossegada de la distribució asimètrica s'anomena **cua**. Una distribució pot ser asimètrica per l'esquerra o asimètrica per la dreta. En la pràctica és més freqüent trobar la asimetria per la dreta.

Les distribucions d'ingressos...

... són asimètriques per la dreta perquè la majoria de les persones reben sous més baixos. En altres paraules, la distribució es concentra en els valors més baixos, mentre que molt poques persones reben sous alts o molt alts, de manera que la distribució s'estén enllà dins la cua superior.

Dibuixar un histograma

Els punts que cal tenir en compte a l'hora de fer un histograma són:

1) Les classes han de ser de la mateixa amplada.

2) Cal repassar les dades i assignar cada valor a una classe. Per exemple, per a les dades de coeficient d'intel·ligència en la taula I es fa un recompte de cada classe de la manera següent:

Taula IV

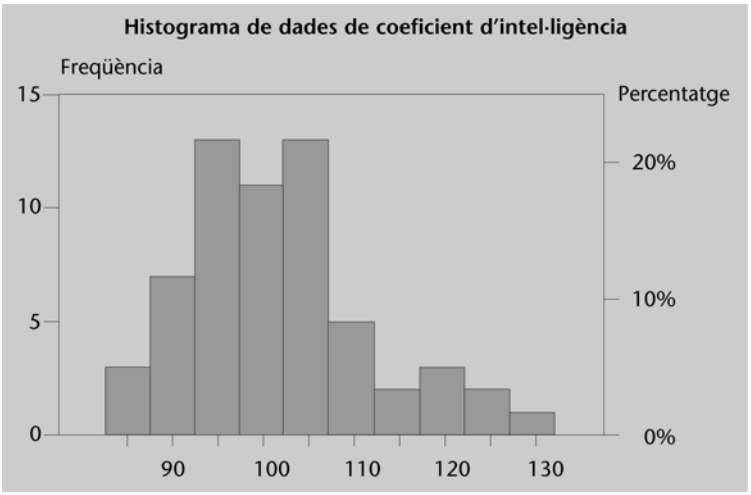
Interval	Recompte	Freqüència
83-87	III	3
88-92	IIIII II	7
93-97	IIIII IIIII III	13
98-102	IIIII IIIII I	11
103-107	IIIII IIIII III	13
108-112	IIIII	5
113-117	II	2
118-122	III	3
123-127	II	2
128-132	I	1

3) Cal marcar l'escala horitzontal en unitats de la variable observada. Cada barra cobreix una classe de valors, sense espais entre les barres adjacents*.

*** Excepte quan no hi ha cap observació en una classe, és clar.**

4) Cal marcar l'escala vertical en recomptes (o percentatges, o totes dues coses alhora, com es pot veure en el gràfic següent).

Gràfic I



Activitats

2. Es mesura el temps que triguen 45 nens per a córrer 50 m:

11.6	14.4	13.3	15.9	15.4
11.6	12.1	14.2	13.8	13.4
11.3	13.1	12.2	13.3	11.2
14.6	12.7	12.1	12.0	10.5
12.6	11.5	10.8	11.2	11.0
11.5	10.7	12.5	13.0	12.9
14.5	13.7	12.0	11.3	10.9
12.3	12.0	11.9	10.0	13.9
11.1	11.8	6.8	12.3	12.8

Dibuixeu un histograma amb aquestes dades. Comenteu la forma de l'histograma.

La distribució de dades categòriques

Finalment volem esmentar la manera usual de dibuixar la distribució d'una variable categòrica. Considereu la taula següent de fons d'inversió a Europa:

Àustria	366	Grècia	84	Portugal	123
Bèlgica	200	Irlanda	270	Espanya	639
Dinamarca	147	Itàlia	344	Suècia	568
Finlàndia	39	Luxemburg	989	Suïssa	218
França	4802	Holanda	175	Regne Unit	1417
Alemanya	493	Noruega	123		

Les dades ja es troben en forma de freqüència i la variable categòrica és "país", amb els 17 països de l'Europa de l'oest com a categories. Podem representar aquestes dades de la manera següent:

```
França       4802   XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...XXXXXXXXXX
Regne Unit   1417   XXXXXXXXXXXXXXXXXXXXXXXXXXXX
Luxemburg     989   XXXXXXXXXXXXXXXXXXX
Espanya       639   XXXXXXXXXXXXX
Suècia        568   XXXXXXXXXXX
Alemanya      493   XXXXXXXXXX
Àustria       366   XXXXXXX
Itàlia        344   XXXXXXX
Irlanda       270   XXXXX
Suïssa        218   XXXX
Bèlgica       200   XXXX
Holanda       175   XXXX
Dinamarca     147   XXX
Noruega       123   XX
Portugal      123   XX
Grècia         84   XX
Finlàndia      39   X
```

Nota: Cada X representa 50 unitats (arrodonim a 50 quan la diferència és menor o igual a 25 unitats).

Diagrama de barres

L'histograma i el diagrama de barres són dos sistemes molt similars que permeten visualitzar la distribució d'una variable. L'única diferència entre tots dos és que l'histograma es construeix per a una variable numèrica després de decidir un conjunt de classes adequades, mentre que el diagrama de barres es construeix per a una variable categòrica en què les classes ja són fetes.

Una diferència entre aquests dos tipus de dades és que les barres de l'histograma es toquen, mentre que les d'un diagrama de barres estan separades per espais. Una altra diferència és que, si les categories no es troben en cap ordre específic, nosaltres podem ordenar de nou les categories perquè el diagrama de barres sigui més fàcil d'interpretar.

En l'exemple anterior, nosaltres hem tornat a ordenar els països des del nombre més alt de fons d'inversions (França) fins al més baix (Finlàndia) perquè el diagrama de barres sigui més fàcil d'interpretar. Noteu que el valor per a França és molt més alt que els altres, tot i que no hem dibuixat tota l'extensió de la barra, cosa que hem indicat amb uns punts suspensius.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Distribució d'una variable: patró dels valors que pren la variable en un estudi, que obtenim resumint quins valors agafa i la freqüència que té cada valor.

Gràfic de tiges i fulles: manera de visualitzar la distribució d'una variable numèrica, usant els dígits superiors del nombre com a tija i els dígits inferiors com a fulles.

Histograma: manera de visualitzar la distribució d'una variable numèrica dividint el rang de valors en classes de la mateixa amplada i després dibuixant el nombre de valors que es troben dins cada classe.

Asimetria: propietat de les distribucions que no són simètriques.

Dada allunyada: valor insòlit que no s'ajusta al patró general d'una distribució.

Diagrama de barres: semblant a un histograma, excepte que la variable és categòrica i les classes són les categories.

3. Les mesures del centre: la mediana i la mitjana aritmètica

Els gràfics de tiges i fulles i els histogrames donen una descripció general d'un conjunt de dades numèriques. Ara fem un cop d'ull a les maneres més específiques de resumir dades numèriques en nombres que ens permetran comparar amb facilitat diferents conjunts de dades. En aquest apartat veiem dues maneres diferents de resumir un valor típic o mitjà d'un conjunt de dades, que mesura el centre d'una distribució.

Paràmetres estadístics

Els paràmetres estadístics són nombres obtinguts amb càlculs a partir de les dades que permeten caracteritzar la variable que s'estudia.

La mediana i la mitjana en són dos exemples.

En aquest apartat sobre mesures del centre aprendreu:

- com es calcula la mediana, o valor central, d'un conjunt de dades;
- com es calcula la mitjana aritmètica, o mitjana, d'un conjunt de dades;
- quines diferències hi ha entre la mediana i la mitjana aritmètica.

La mediana o l'observació central

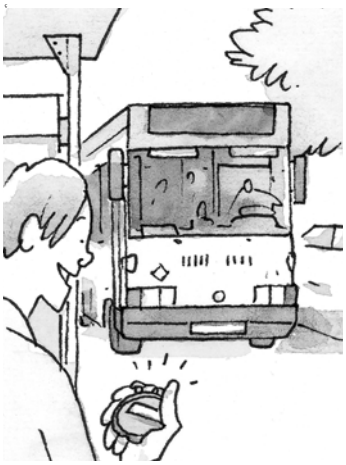
Una manera fàcil d'aconseguir un valor per al centre d'una distribució és trobar quina observació queda exactament al mig. Amb això volem dir que la meitat de les observacions quedin per sota d'aquest valor i l'altra meitat, per sobre. Aquest valor s'anomena **mediana** de la distribució.

Vegem-ne un altre exemple. Suposem que al llarg d'un període de 27 dies anoteu l'estona que heu d'esperar fins que l'autobús arriba al matí. Les dades, en minuts, són les següents:

Taula I

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

Temps d'espera fins que arriba l'autobús, en minuts.



Si tenim un recull de dades podem saber *a posteriori* quina és la mediana corresponent.

Ara pregunteu quin valor podeu usar com a típic per a descriure l'estona que heu esperat. El gràfic de tiges i fulles d'aquestes dades és:

Taula II

```
0 | 2334444
0 | 55667777888899
1 | 11223
1 | 7
```

Gràfic de tiges i fulles de dades sobre el temps d'espera.

Noteu que per a elaborar aquest gràfic de tiges i fulles usem els intervals 0-4, 5-9, 10-14 i 15-19.

A l'hora de construir el gràfic de tiges i fulles hem posat totes les observacions en ordre ascendent, de l'observació més petita (2 minuts) a la més gran (17 minuts). Com que hi ha 27 observacions, l'observació central serà la catorzena en la llista ordenada, ja que hi haurà 13 valors abans que el catorzè i 13 després. El valor catorzè és, de fet, 7 minuts. El fet que hi hagi un cert nombre d'observacions de 7 minuts no té importància (de fet, en la llista ordenada, el valor dotzè, tretzè, catorzè i quinzè és 7).

Una regla per a aconseguir la mediana

La lletra *n* s'usa convencionalment per al nombre d'observacions en un conjunt de dades. La regla general per a trobar la posició de l'observació central en una llista de *n* valors que ha estat ordenada de més petit a més gran és:

$$\frac{n+1}{2}.$$

En el nostre exemple, amb $n = 27$ valors, el valor central era el valor en la posició $(27 + 1) / 2 = 14$ de la llista.

Quan *n* és un nombre senar, el nombre de l'observació per a la mediana és un enter exacte. Ara bé, quan *n* és un nombre parell, no hi ha cap observació exactament central en la llista ordenada. Per exemple, si hi havia 26 observacions, aleshores la nostra fórmula ens dóna el número $(26 + 1) / 2 = 13{,}5$. El que fem ara és prendre com a mediana el punt mitjà entre els números que ocupen el tretzè i catorzè lloc de la llista ordenada. Això encara ens dóna un valor en què la meitat de les observacions queden a sota i l'altra meitat a sobre, de manera que satisfà la definició de la mediana.

Els valors resum com la mediana fan que les comparacions entre diferents grups d'observacions siguin més fàcils.

La mitjana aritmètica o valor mitjà

La mitjana aritmètica d'un conjunt de dades numèriques és la mateixa que el seu valor mitjà. Per a calcular la mitjana aritmètica no cal començar organitzant els valors de les dades ordenadament. Simplement sumem tots els valors i dividim pel nombre total de dades *n*.

El valor mitjà és la mitjana de tots els valors de la variable.

Per a les dades de la taula I els càlculs són els següents:

1) Sumeu els 27 valors: $9 + 5 + 6 + \ldots + 7 + 4 = 200$.

2) Dividiu la suma per 27: $200/27 = 7{,}41$.

Recordeu

El valor mitjà no sempre és igual al valor central.

La mitjana aritmètica d'aquests valors és, per tant, 7,41 minuts –al llarg dels 27 dies heu hagut d'esperar que l'autobús arribés una mitjana de 7,41 minuts.

Penseu-hi

Quina és la mitjana d'hores per dia que heu estudiat aquesta setmana?

Unes quantes notacions

Nosaltres considerem un conjunt de n observacions numèriques d'una variable X. Denotem els valors genèrics amb els símbols x_1, x_2, x_3, etc., fins a x_n. Denotem aquest conjunt d'observacions amb x_1, x_2, x_3, ..., x_n, o amb x_i, $i = 1$, ..., n, on el símbol i utilitzat en els subíndexs s'anomena *índex*. Així, per a les dades de la taula I, $x_1 = 9$, $x_2 = 5$, $x_3 = 6$, ..., $x_{27} = 4$. A l'hora d'ordenar les observacions de més petita a més gran denotarem el nou conjunt de quantitats amb els símbols $x_{(1)}$, $x_{(2)}$, $x_{(3)}$, etc., fins a $x_{(n)}$. Per tant $x_{(1)}$ és el valor més petit i $x_{(n)}$ és el més gran. En el nostre exemple, $x_{(1)} = 2$ i $x_{(27)} = 17$.

Si n és un enter senar, l'observació central és en la posició $(n + 1) / 2$, la qual podem denotar per m. La mediana és d'aquesta manera igual a $x_{(m)}$. Si n és un enter parell, $m = (n + 1) / 2$ és a mig camí entre els dos enters, $m - (1/2)$ i $m + (1/2)$. La mediana és així igual al valor mitjà entre $x_{(m-0,5)}$ i $x_{(m+0,5)}$. Per exemple, en la nostra exemplificació, quan considerem $n = 26$ observacions, $m = 13,5$, és a dir, la mediana és el valor mitjà de les observacions tretzena i catorzena de la llista ordenada.

La mitjana aritmètica d'un conjunt de valors x_i, $i = 1$, ..., n, normalment es denota amb $\bar{x}$**.** Usant la notació introduïda, la mitjana aritmètica és igual a:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,.$$

Recordeu

x = variable.

$x_1\ x_2\ x_3\ \ldots\ x_n$ = observacions (dades).

n =grandària de la mostra o població.

$x_{(1)}\ x_{(2)}\ x_{(3)}\ \ldots\ x_{(n)}$ = dades ordenades.

m = posició de la mediana.

$x_{(m)}$ = la mediana.

$\bar{x}$ = la mitjana.

Efecte de les dades allunyades o insòlites en la mediana i la mitjana aritmètica

Tant la mediana com la mitjana aritmètica mesuren el centre de la distribució, però ho fan de maneres diferents. Solament quan la distribució és simètrica, les dues mesures són iguals. La diferència més important entre l'una i l'altra és com es veuen afectades per asimetries i dades allunyades. Quan la distribució és asimètrica, la mitjana aritmètica sempre es desplaça cap a la cua de la distribució. En el cas més comú d'una distribució que és asimètrica cap a la dreta, aleshores la mitjana aritmètica és més alta que la mediana.

La presència d'un valor molt gran no afecta la mediana, però influeix moltíssim sobre la mitjana aritmètica. Diem que la mediana **resisteix** les dades allunyades. Per exemple, imaginem-nos que, en comptes de 17 minuts, el valor més gran en la taula II fos 45 minuts, cosa que és una espera molt llarga per a un sol dia. Aquest canvi no afecta la mediana, de fet romandria igual, fins i tot si el canviàvem per un valor molt més gran.

Vegeu la taula II d'aquest apartat.

La mitjana aritmètica, però, s'afectaria, ja que la suma de totes les observacions ara seria 228, la qual dividida per 27 dóna el valor 8,44 minuts. Aquest increment d'una observació fa pujar la mitjana aritmètica del temps d'espera en un minut, malgrat que els altres 26 valors romanguin intactes. En una situació com aquesta, la mitjana aritmètica no té la condició de ser un valor típic.

Activitats

1. Considereu novament les dades sobre els sous del professorat universitari donades en l'apartat 2. Calculeu la mediana i la mitjana aritmètica d'aquestes dades. Comenteu-ne els resultats.

Vegeu l'activitat 1 de l'apartat 2.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Mediana: observació central; observació numèrica que divideix les dades en dues parts iguals, de manera que una meitat queda sota la mediana i l'altra, a sobre.

Resistent: propietat de la mediana que significa que els valors extrems de la distribució no afecten la mediana.

Mitjana aritmètica: terme mitjà d'un conjunt de dades numèriques, calculat sumant tots els valors de les dades i dividint-los pel nombre total.

Notacions:

a) Conjunt de n observacions: x_i, $i = 1, ..., n$.

b) Mateixes observacions en ordre ascendent: $x_{(i)}$, $i = 1, ..., n$.

c) Mitjana aritmètica de x_i, $i = 1, ..., n$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

4. Mesures de dispersió: els quartils i la desviació estàndard

En l'apartat 3 hem après diverses maneres de mesurar el centre d'una distribució. Però per a descriure una distribució adequadament no n'hi ha prou de conèixer el centre. Volem, també, resumir d'una manera concisa fins a quin punt les observacions es reparteixen al voltant del centre. En aquest apartat ens presenten diverses maneres de resumir la dispersió d'una distribució. Una manera simple és identificar la més petita i la més gran de les observacions. Després podem identificar els quartils de la distribució, els quals són el punt mitjà de les meitats superior i inferior del conjunt de dades. Finalment, definim una mesura ben coneguda de dispersió d'una distribució anomenada *desviació estàndard*.

En aquest apartat sobre mesures de dispersió aprendreu:

- com es resumeix la dispersió d'una distribució mitjançant cinc quantitats: el mínim, el primer quartil, la mediana, el tercer quartil i el màxim;
- com es representa el resum d'aquestes cinc quantitats en un diagrama de caixa;
- com es calcula la desviació estàndard d'un conjunt de valors.

Mesurar l'extensió

En l'apartat 2 hem considerat diverses maneres de dibuixar la distribució d'una variable. En el 3 hem definit diverses maneres de calcular els nombres que mesuren el centre d'una distribució, sabent que el centre no és suficient per a descriure una distribució adequadament. També necessitem mesurar fins a quin punt les observacions es reparteixen a banda i banda del centre. Hi ha maneres diverses de mesurar la dispersió. Aquestes maneres diverses també depenen de si la distribució és simètrica o no, i de si hi ha presència de dades insòlites.

Els valors mínim i màxim

La manera més simple de mesurar la dispersió és identificar el valor més petit i el més gran d'un conjunt de dades. La diferència entre els valors mínim i màxim s'anomena **rang** (o **recorregut**) de les observacions. En termes de la notació definida en l'apartat 3, en què $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$ són l'ordre estadístic d'una distribució:

Valor mínim $= x_{(1)}$
Valor màxim $= x_{(n)}$
Rang $= x_{(n)} - x_{(1)}$

Els quartils

Mentre que la mediana divideix la distribució en meitats, els quartils d'una distribució són una variació de la idea de mediana. Els quartils són els valors que divideixen la distribució en quarts.

Hi ha tres quartils:

1) El **primer quartil** té un quart de les observacions a sota seu i tres quarts a sobre.

2) El **segon quartil** té dos quarts a sota i dos quarts a sobre –per tant, el segon quartil és idèntic a la mediana.

3) El **tercer quartil** té tres quarts de les observacions a sota i un quart a sobre.

Nota

El primer quartil, d'una manera similar a la mediana, divideix la meitat inferior de les dades en dues parts iguals –és a dir, és la mediana de la meitat inferior de les dades, mentre que el tercer quartil és la mediana de la meitat superior–. És clar aleshores que entre els quartils primer i tercer hi ha la meitat de les dades.

Una altra manera de pensar en això és que la mediana, o segon quartil, divideix les dades en dos grups de la mateixa grandària, que anomenarem *meitat inferior de les dades* i *meitat superior*. Sovint el primer i el tercer quartils s'anomenen **quartils inferior i superior** respectivament.

Calcular els quartils

Calculem el quartils exactament de la mateixa manera que calculem la mediana, excepte que apliquem el càlcul a les meitats inferior i superior de les dades separadament.

Si mireu les dades en la taula I de l'apartat 3, per a les 27 observacions havíem vist que la mediana era el valor catorzè en la llista ordenada, és a dir, 7. La meitat inferior de les dades és, per tant, el conjunt d'observacions des de la primera a la tretzena, i la meitat superior és el conjunt des de la quinzena a la vint-i-setena. Preneu nota que el valor catorzè és el tercer 7 en la llista ordenada i que hi ha quatre 7 en les dades, de manera que la meitat inferior de les dades inclou dos 7:

2 3 3 4 4 4 4 5 5 6 6 7 7

Vegeu el gràfic de tiges i fulles en la taula II de l'apartat 3.

i la meitat superior inclou un 7:

7 8 8 8 8 9 9 11 11 12 12 13 17

Per a trobar els quartils, calculem les medianes d'aquestes meitats de les dades per separat. Hi ha 13 valors en cada meitat, per tant la mediana és el valor amb el nombre de seqüència (13 + 1) / 2 = 7 en cada llista. El setè valor en la meitat inferior és 4, i el setè en la meitat superior és 9.

El rang interquartílic

La diferència que hi ha entre els quartils primer i tercer s'anomena *rang interquartílic*.

En el nostre exemple anterior, el rang interquartílic és igual a $9 - 4 = 5$ minuts.

Els cinc nombres resum de les dades

Els cinc nombres resum d'una distribució és el conjunt següent:

- el mínim,
- el primer quartil,
- la mediana (o segon quartil),
- el tercer quartil,
- el màxim.

En el nostre exemple de temps d'espera, els cinc nombres resum són 2, 4, 7, 9 i 17. En altres paraules, la mediana del temps d'espera és 7 minuts, la meitat de les esperes queden entre 4 i 9 minuts (amb un rang interquartílic de 5 minuts), el temps mínim d'espera és 2 minuts, i el màxim, 17 minuts (amb un rang de 15 minuts). Això és un resum global de la distribució.

Diagrames de caixa

El diagrama de caixa és un gràfic simple dels cinc nombres resum de les dades.

Es dibuixa una escala vertical o horitzontal que es correspon amb l'escala de la variable. Després es dibuixa un quadre amb els nivells inferior i superior en els quartils primer i tercer respectivament. Es dibuixa una línia en el quadre que correspon a la mediana. Després es dibuixen dos braços al capdamunt i al capdavall del quadre fins als valors màxim i mínim respectivament.

El que tenim ara és una descripció gràfica compacta de tota la distribució de la variable. Es poden dibuixar els diagrames de caixa corresponents als diversos conjunts d'observacions sobre la mateixa variable l'un al costat de l'altre i després comparar-los visualment.

Activitats

1. Els índexs d'atur, mesurats com un percentatge de la població activa, per a 27 països del Primer Món són els següents:

7.0	13.5	10.7	17.9	12.3	8.2	4.7	14.4	11.8	2.6
7.2	5.3	6.8	23.5	8.2	4.6	8.4	5.4	2.9	12.7
3.1	11.4	15.8	11.1	14.5	14.3	2.2			

Calculeu els cinc nombres resum d'aquestes dades.

A continuació veurem una definició numèrica alternativa de dispersió anomenada *desviació estàndard*.

La **desviació estàndard** és un nombre únic que es pot usar per a quantificar la dispersió d'un conjunt de dades, més que no pas diversos nombres com en el cas dels cinc nombres resum.

Penseu-hi

Quin valor obtenim si sumem totes les diferències dels valors d'una distribució respecte de la mitjana?

Calcular la desviació estàndard

Recordeu que ara estem interessats a trobar un únic nombre que resumeixi la dispersió de dades, i ens interessa molt particularment la dispersió al voltant de la mitjana aritmètica. Aquest càlcul es faria de la manera següent:

1) El primer pas és calcular totes les diferències entre cada observació i la mitjana aritmètica del conjunt. És clar que com més grans són les diferències, més gran és la dispersió de les dades, però necessitem combinar totes aquestes desviacions en una figura global.

2) Calcular la variància és el pas següent. Fem el quadrat de cada una de les desviacions, les sumem i després dividim la suma que en resulta per $n - 1$ (el nombre de les observacions menys 1). Cal dividir per $n - 1$ i no per n, malgrat que pugui semblar més intuïtiu dividir per n per a obtenir la mitjana del quadrat de les desviacions. El resultat d'aquests càlculs és la variància. Ara donem la fórmula per al càlcul de la variància de n valors de les dades $x_1, x_2, \ldots, x_n$:

La variància és la mitjana aritmètica dels quadrats de les desviacions de les dades d'una sèrie respecte a llur mitjana aritmètica.

$$\text{variància} = s^2 \frac{1}{n-1} \sum (x_i - \overline{x})^2.$$

Heus aquí un altre exemple senzill. Imagineu-vos que tenim els preus d'un cert producte comprat en sis comerços diferents: 260, 240, 250, 210, 230 i 250 pessetes. La suma d'aquests valors és 1.440, de manera que el valor mitjà és 1.440/6 = 240. Les desviacions del valor 240 són 20, 0, 10, −30, −10 i 10 (preneu nota que les desviacions respecte de la mitjana aritmètica sempre sumen 0). Els qua-

drats d'aquestes desviacions són: 400, 0, 100, 900, 100 i 100, i sumen 1.600. Finalment, dividim per $n - 1 = 5$ per a obtenir la variància 1.600/5 = 320.

La **desviació estàndard** és simplement l'arrel quadrada positiva de la variància.

Càlcul de la desviació estàndard

x = variable.

$x_1\ x_2\ x_3\ \ldots\ x_n$ = valors de la variable.

n = nombre d'observacions o valors.

Càlcul de la mitjana $\bar{x}$.

Càlcul de les desviacions $x_i - \bar{x}$.

Càlcul de les desviacions quadràtiques $(x_i - \bar{x})^2$

Càlcul de la variància:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- Càlcul de la desviació estàndard: $\sigma = \sqrt{s^2}$.

Fixeu-vos que la variància es calcula en unitats que són els quadrats de les unitats de les dades originals. Per tant, prenent l'arrel quadrada de la variància per a obtenir la desviació estàndard la mesura d'extensió torna a les unitats originals. En el nostre exemple, la desviació estàndard dels preus és l'arrel quadrada de 320, és a dir, 17,9, posem-hi 18 pessetes.

La interpretació de la desviació estàndard

Donarem una interpretació més exacta de la desviació estàndard més endavant. De moment, simplement adoneu-vos que moltes de les desviacions respecte de la mitjana aritmètica queden dins una desviació estàndard. Per exemple, de les sis desviacions 20, 0, 10, −30, −10 i 10 calculades abans, quatre tenen valors absoluts més petits que 18.

Veurem la desviació estàndard amb més detall a l'apartat 9.

Quan hem d'usar desviacions estàndard i els cinc nombres resum

Significat de la desviació estàndard

Un cop calculada la desviació estàndard d'una distribució, cal veure el percentatge de les observacions o dades que queden dins els intervals següents:
$(\bar{x} - s, \bar{x} + s)$.
$(\bar{x} - 2s, \bar{x} + 2s)$.

Tant els cinc nombres resum (i la seva versió gràfica, el diagrama de caixa) com la desviació estàndard mesuren la dispersió, però de maneres diferents.

La desviació estàndard té l'avantatge de ser un nombre únic, però realment s'hauria d'usar solament quan les distribucions són més o menys simètriques. Quan les distribucions són asimètriques, la dispersió sota el centre i la dispersió sobre el centre no són les mateixes, i ho indicaran els cinc nombres resum i no pas la mitjana aritmètica. També com la mitjana aritmètica, la desviació estàndard és altament sensible a les observacions allunyades. No obstant això, la desviació estàndard és, de lluny, l'estadístic d'ús més comú per a mesurar la dispersió, i nosaltres la usarem sovint al llarg de la resta d'aquesta assignatura.

Parlarem de les transformacions en l'apartat 5.

Sovint, quan les dades són asimètriques, es fa un esforç per transformar les dades de manera que aquests valors transformats siguin més simètrics. En aquest cas també és possible utilitzar la desviació estàndard per a resumir la dispersió de les observacions transformades.

Activitats

2. Un grup de consumidors comproven l'asseveració dels fabricants d'unes noves piles de llarga durada. Sotmeten 20 piles a una càrrega estàndard fins que són totalment buides. Les durades de les piles són les següents (en minuts):

65.1	58.4	64.9	76.0	67.8	75.1	76.7	64.2	74.9	77.6
58.0	68.0	73.3	75.4	76.0	59.4	65.4	74.7	76.6	81.3

Calculeu la mitjana aritmètica i la desviació estàndard d'aquestes dades.

Un comentari sobre els càlculs

No sempre s'espera que feu tots els càlculs per a determinar la mitjana aritmètica i les desviacions estàndard. Més endavant us iniciem en els programes informàtics que us faran els càlculs molt més fàcils. Malgrat que normalment usareu un ordinador com a ajuda, hauríeu d'estar familiaritzats en la manera de fer els càlculs.

Vegeu l'apartat 6.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Primer quartil (o quartil inferior): valor de la dada que té un quart de les observacions a sota i tres quarts a sobre.

Tercer quartil (o quartil superior): valor de la dada que té tres quarts de les observacions a sota i un quart a sobre.

Rang interquartílic: diferència entre els quartils inferior i superior.

Els cinc nombres resum d'una distribució: mínim, quartil inferior, mediana, quartil superior i màxim d'un conjunt de dades.

Diagrama de caixa: versió gràfica dels cinc nombres resum, que mostra els quartils en un quadre i dos braços que s'estenen cap amunt i cap avall fins als valors mínim i màxim.

Variància: mena de valor mitjà de les desviacions al quadrat de les observacions respecte de la seva mitjana aritmètica.

Desviació estàndard: arrel quadrada positiva de la variància, una mesura d'extensió útil per a distribucions aproximadament simètriques.

5. Mesures de relació: la correlació

Un del termes usats amb més freqüència a l'hora de parlar de la relació entre variables és *correlació*. Diem que dues variables estan correlacionades quan en algun sentit estan connectades o associades. Si dues variables es correlacionen, saber el valor d'una variable ens donarà una bona idea del valor de l'altra variable.

En aquest apartat expliquem aquest concepte de correlació i una manera específica de mesurar la força de la relació entre dues variables, usant el coeficient de correlació.

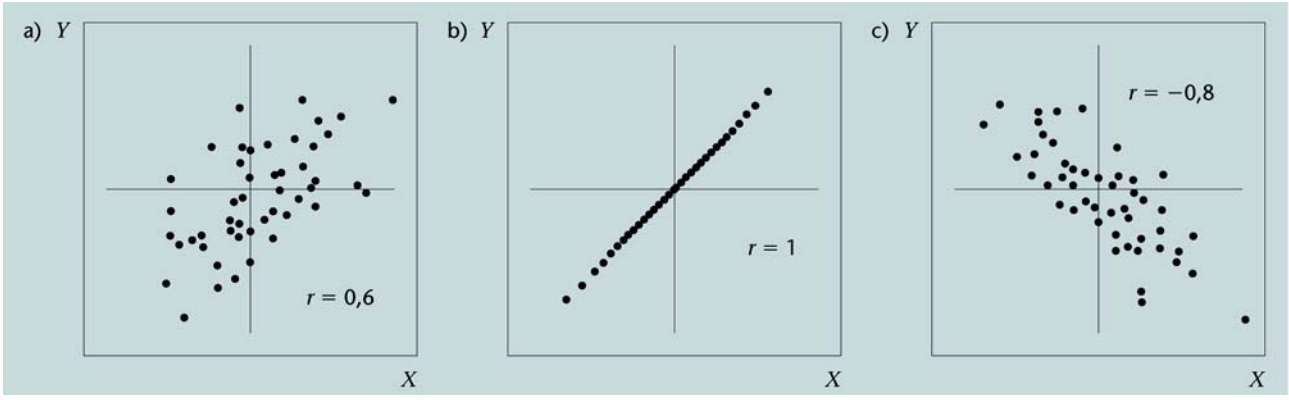
En aquest apartat sobre relacions entre variables aprendreu:

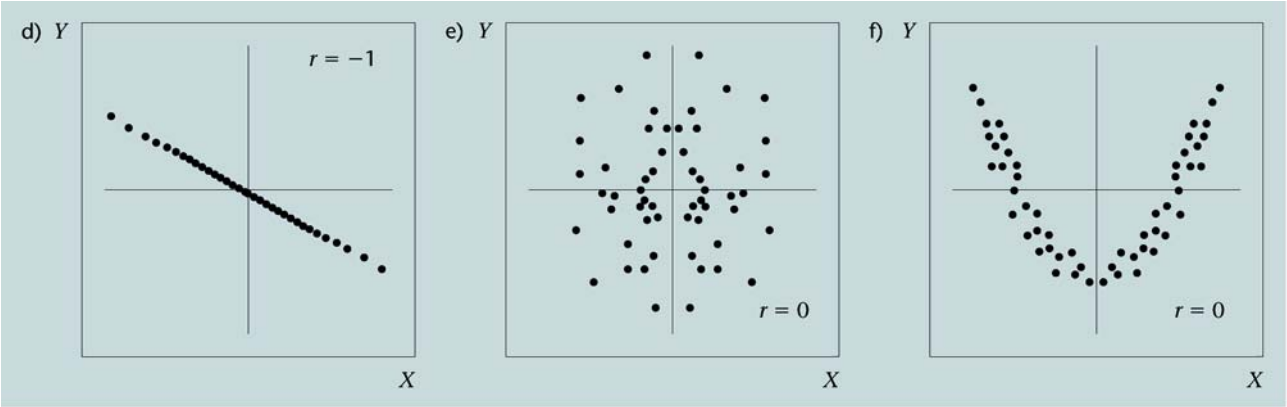
- el concepte de correlació com una mesura d'associació lineal;
- com es calcula un coeficient de correlació;
- com es contrasta un coeficient de correlació per a una significació estadística.

La mesura de l'associació lineal

La correlació és una mesura de la força de l'associació entre dues variables. El nostre interès per l'associació entre dues variables es limita a l'associació lineal que tenen, és a dir, a quina proximitat d'una recta queden els punts en un gràfic de dispersió. És clar que aquesta no és l'única mena d'associació que podem tenir entre dues variables.

A continuació mostrem diversos gràfics de dispersió diferents i els valors corresponents dels coeficients de correlació. Observeu en l'últim gràfic de dispersió que les dues variables mostren una relació corba molt forta, però la correlació és zero –això il·lustra el fet que la correlació solament és útil per a mesurar relacions lineals.





Calcular el coeficient de correlació

En termes de les quantitats que ja hem definit en apartats anteriors, podem definir el **coeficient de correlació *r*** de dades aparellades (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) de la manera següent:

$$r = \frac{\text{cov}(x,y)}{s_x s_y},$$

Karl Pearson (1857-1936)

Matemàtic, estadístic i filòsof anglès. Va trobar la fórmula per a quantificar la relació estadística entre dues variables: el coeficient de correlació lineal *r*.

on cov(*x*, *y*) és la covariància entre els valors *x* i *y*:

$$\text{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \cdot$$

La covariància de dues variables...

... és lleugerament més general que la variància: implica sumar productes que tenen la forma $(x_i - \bar{x})(y_i - \bar{y})$, mentre que la variància suma aquells que tenen la forma $(x_i - \bar{x})^2$. Per tant la covariància d'una variable *x* i ella mateixa és el mateix que la variància: cov (x,x) s_x^2.

La covariància s'assembla molt a la variància, que hem vist a l'apartat anterior. Però en lloc de quadrar les desviacions de la mitjana d'una sola variable, multipliquem les desviacions de les dues variables. Aleshores, la covariància cov(*x*, *x*) entre una variable *x* i la mateixa variable *x* és igual a la variància de *x*.

Així doncs, el coeficient de correlació és la covariància entre les dues variables dividida pel producte de les seves desviacions estàndard.

Una altra manera de pensar en la correlació és primer transformar els valors *x* i *y* dividint-los per les seves respectives desviacions estàndard; ara els anomenem, doncs, **valors transformats**:

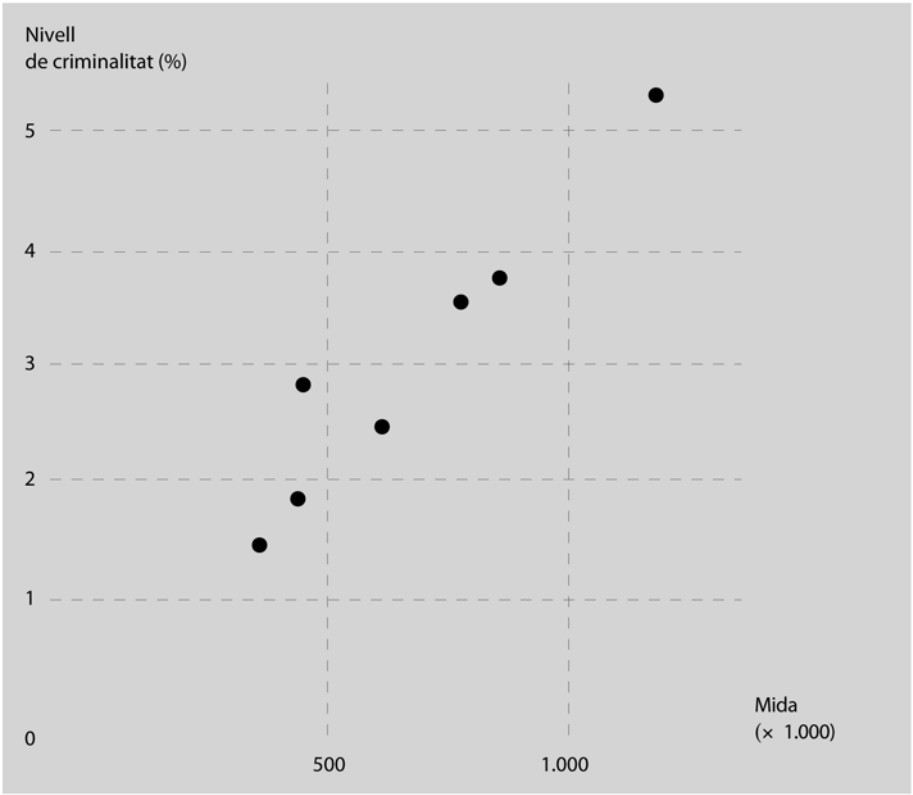
$$x_i^* = \frac{x_i}{s_x} \quad \text{i} \quad y_i^* = \frac{y_i}{s_x}.$$

Aleshores el coeficient de correlació és la covariància entre els valors transformats: $r = \text{cov}(x^*, y^*)$.

Vegem ara un exemple de la correlació entre el nivell de criminalitat i el total de la població en 7 ciutats. Les dades són les següents:

Nivell de criminalitat (%) (x)	Mida de la població (x 1.000) (y)
2,9	490
3,5	720
1,4	410
5,1	1.270
3,7	840
1,9	450
2,5	580

El gràfic següent mostra que hi ha una relació positiva entre les dues variables i que la relació és a prop de ser lineal:



Calculem ara el coeficient de correlació. Primer hem de calcular les mitjanes de cada variable:

$$\text{mitjana de } x\text{: } \bar{x} = \frac{(2,9 + 3,5 + \ldots + 2,5)}{7} = 3,0,$$

$$\text{mitjana de } y\text{: } \bar{y} = \frac{(490 + 720 + \ldots + 580)}{7} = 680.$$

Ara fem una taula que facilitarà els càlculs:

x_i	y_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
2,9	490	−0,1	0,01	−190	36.100	19
3,5	720	0,5	0,25	40	1.600	20
1,4	410	−1,6	2,56	−270	72.900	432
5,1	1.270	2,1	4,41	590	348.100	1.239
3,7	840	0,7	0,49	160	25.600	112
1,9	450	−1,1	1,21	−230	52.900	253
2,5	580	−0,5	0,25	−100	10.000	50
			9,18		**547.200**	**2.125**

Les variàncies són:

$$\text{variància de } x = \frac{9,18}{6} = 1,53,$$

$$\text{variància de } y = \frac{547.200}{6} = 91.200\ ;$$

i les desviacions estàndard:

$$\text{desviació estàndard de } x = \sqrt{1,53} = 1,237,$$

$$\text{desviació estàndard de } y = \sqrt{91.200} = 302,0;$$

i la covariància entre les variables:

$$\text{covariància} = \frac{2.125}{6} = 354,2.$$

Així doncs, el coeficient de correlació és la covariància dividida pel producte de les desviacions estàndard:

$$\text{correlació} = \frac{354,2}{1,237 \cdot 302,0} = 0,948.$$

Aquest coeficient és molt alt, i significa que la relació entre el nivell de criminalitat en una ciutat i el total de la seva població és molt forta. Fixeu-vos que no diem que el total de la població sigui una causa d'una criminalitat alta, només que hi ha una relació entre les dues variables.

Activitats

1. Un grup de 10 estudiants tenen les notes següents en l'examen de matemàtiques de COU:

```
6.1   7.0   5.5   6.5   7.1   6.4   7.4   6.8   7.2   7.4
```

i les notes següents en el primer examen d'estadística a la Universitat Oberta (en el mateix ordre):

5	8	6	6	6	7	8	7	8	9

Quina és la correlació entre els dos conjunts d'observacions?

La correlació entre variables discretes amb dues categories

Sovint tindrem variables discretes amb només dues categories, per exemple sí o no, certa o falsa, suspès o superat, etc. En aquest cas particular, podem mesurar la relació entre dues variables utilitzant el coeficient de correlació. Per a poder fer el càlcul, hauríem d'assignar valors a cada categoria –el més habitual és codificar una categoria amb el valor 1 i l'altra amb el 0–. Aquesta selecció no té cap influència sobre el valor de la correlació, però la utilització dels valors 1 i 0 té altres avantatges, per exemple en la interpretació de la mitjana.

Consulteu els annexos 1, 2, 3 i 4 per a ampliar aquesta informació.

Tornarem a aquest tema en els apartats 12, 15 i 16, quan tractem específicament les variables discretes.

El **concepte principal** que hem vist en aquest apartat és el següent:

Coeficient de correlació: el coeficient de correlació r és una mesura d'associació lineal entre punts (x_i, y_i), $i = 1, \ldots, n$, definit de la manera següent:

$$r = \frac{\operatorname{cov}(x,y)}{s_x s_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

6. Càlcul estadístic: introducció al programa MacAnova

En aquest apartat aprenem a fer servir un programa d'estadística que farà que tots els càlculs que heu de fer per a aquesta assignatura us siguin més fàcils. Ja heu de tenir aquest programa instal·lat en el vostre ordinador.

En aquest apartat sobre càlcul estadístic aprendreu, amb l'ajut de l'ordinador:

- com s'inicia el programa MacAnova i com s'atura;
- com s'usa el programa com a calculadora;
- com s'obté ajuda quan es fa servir el programa;
- com es llegeix en un conjunt de dades;
- com es fan càlculs elementals com ara la mitjana aritmètica, la variància i la desviació estàndard d'un conjunt de dades.

Ara ja podeu engegar el vostre ordinador, en el qual heu instal·lat el programa MacAnova. És preferible estudiar aquest apartat sencer en una sola sessió, sense parar l'ordinador.

W1/00520.01
Si no heu instal·lat encara el MacAnova, recordeu que el podeu trobar en aquesta web.

El programa MacAnova

El MacAnova és un programa interactiu que treballa amb ordres. Això vol dir que mentre el programa funciona podeu introduir una ordre, després el programa hi dóna una resposta, d'acord amb la qual després podeu introduir una altra ordre, i així successivament. Aquest és un entorn de treball molt senzill i aviat veureu que fàcil és executar càlculs corrents, com si el programa fos una calculadora, a més d'executar càlculs estadístics i fins i tot dissenyar programes d'ordinador sofisticats.

Iniciar i tancar el programa

Si heu reeixit a instal·lar el programa, ja el deveu haver provat i hi deveu haver entrat i en deveu haver sortit.

En màquines de DOS, s'inicia el programa simplement introduint l'ordre

```
macanova
```

Icona de MacAnova que ens apareix en un Macintosh i sobre la qual hem de fer doble clic.

i es tanca el programa escrivint

```
quit
```

En un Macintosh, feu doble clic a la icona MacAnova.

En tots dos casos surt un missatge d'inici en la pantalla, i després apareix l'indicador del sistema:

```
Cmd>
```

Sempre heu de teclejar les ordres després de l'indicador de sistema `Cmd>`.

Usar el MacAnova com a calculadora

L'ús més simple del MacAnova és com a calculadora: introduïu els càlculs que voleu executar, i el programa hi dóna la resposta. Per exemple, si voleu calcular la diferència entre 34,5 i 23,7 i dividir-la per 6,2, senzillament introduïu l'ordre (34,5 − 23,7) / 6,2 i hi rebreu la resposta:

Nota

No us amoïneu per l'`(1)` de l'esquerra de la resposta `1,7419`; més tard ja veureu què signifiquen aquests nombres.

```
Cmd> (34.5-23.7)/6.2
 (1)          1.7419
```

Fixeu-vos que si s'introdueix l'ordre sense parèntesis, s'obtindrà la diferència entre 34,5 i 23,7/6,2 perquè la divisió té prioritat sobre la resta:

```
Cmd> 34.5-23.7/6.2
 (1)        30.677
```

La prioritat dels operadors matemàtics és l'habitual: primer s'executa l'exponent (^), després la multiplicació i la divisió (* i /), i finalment la suma i la resta (+ i −). Quan dubteu, poseu-hi uns parèntesis extres. Aquí en teniu uns quants exemples més:

```
Cmd> 1.96*23.5+10.0 ; (12.54-7.86)*0.93^2
(1)        56.06
(1)        4.0477
```

Fixeu-vos aquí que es pot introduir més d'un càlcul en la mateixa línia, separant-los amb un punt i coma (;). Es reben totes dues respostes. El primer càlcul executa $(1{,}96 \cdot 23{,}5) + 10{,}0$, i el segon $(12{,}54 - 7{,}86) \cdot 0{,}93^2$.

Assignar valors a les variables

El terme *variable* s'usa en terminologia informàtica per a designar un lloc d'emmagatzematge en la memòria. Es poden definir variables amb noms de

fins a dotze caràcters de llargada, per exemple `ingressos`, `pes` i `UOC1995` són tots noms vàlids. Els noms han de començar amb una lletra, de manera que el nom `2any` no és vàlid. Els noms són sensibles a l'escriptura, de manera que els noms `ingressos`, `Ingressos` i `INGRESSOS` són diferents per al programa. Per exemple, es pot assignar el valor 2,54 a la variable `InchtoCent` usant l'operador <- (un símbol 'menys que' < seguit d'un guionet –) d'aquesta manera:

```
Cmd> InchtoCent <- 2.54
```

Probablement sabeu que el valor 2,54 és el factor de conversió entre polzades i centímetres, de manera que un cop s'ha assignat el valor 2,54 a la variable `InchtoCent` es pot convertir qualsevol valor en polzades a centímetres multiplicant per `InchtoCent`, per exemple per a convertir 2 peus o 24 polzades a centímetres:

```
Cmd> 24*InchtoCent
 (1)         60.96
```

Si voleu veure el valor que una variable emmagatzema, simplement n'heu d'introduir el nom:

```
Cmd> InchtoCent
 (1)       2.54
```

Hi ha algunes constants implementades dins el MacAnova, per exemple: les constants π i e, anomenades `PI` i `E` respectivament (recordeu que el programa diferencia les majúscules de les minúscules). Introduïu les ordres següents per veure'n els valors i per executar el càlcul:

$$\frac{1}{2\sqrt{\pi}} e^{-\frac{1}{2}}$$

```
Cmd> PI
(1)        3.1416

Cmd> E
(1)        2.7183

Cmd> (0.5/PI^0.5)* E^(-0.5)
(1)        0.1711
```

Funcions

El MacAnova conté totes les funcions matemàtiques i trigonomètriques habituals, com ara el logaritme natural o neperià, `log()`; el logaritme en base 10, `log10()`; el sinus, `sin()`; l'exponencial, `exp()`, i l'arrel quadrada, `sqrt()`. Quan fem referència a la funció, posem dos parèntesis després del nom per indicar que és una funció, no pas una variable, i que l'argument de la funció queda expressat entre els parèntesis. Aquí en teniu alguns exemples d'ús:

```
Cmd> log(3)+5*sqrt(6)
(1)        13.346

Cmd> sin(30)
(1)        -0.98803
```

Com que sabem que sin(30) = 0,5, deduïm que el MacAnova espera que els angles siguin en radians. Per a canviar-los a graus, cal introduir l'ordre `setoptions(angles:"degrees")`; després d'això totes les funcions trigonomètriques són en graus:

```
Cmd> setoptions(angles:"degrees")
Cmd> sin(30)
(1)          0.5
```

Aquesta opció roman activa durant tota la sessió de MacAnova fins que o bé la canvieu, o bé sortiu del programa. Per tant, recordeu que, quan torneu a iniciar el MacAnova, les unitats per defecte de mesura d'angles tornaran a ser els radians.

Aconseguir ajuda mentre feu anar el programa MacAnova

El MacAnova té una funció d'ajuda, `help()`, implementada en el programa. Podem demanar ajuda sobre qualsevol de les ordres.

Per exemple, per saber com es passen els angles a graus, podem demanar sobre la funció del sinus, `sin()`, introduint l'ordre següent:

```
Cmd> help(sin)
```

Allò que el programa contesta és una llista completa de totes les funcions de transformació accessibles i una indicació que les funcions trigonomètriques són per a angles en radians. També explica com es canvien els angles a graus.

Activitats

1. Introduïu l'ordre `help(sin)` –també podeu introduir `help ("transformations")`– i vegeu les diverses transformacions que el MacAnova proporciona. Després executeu els càlculs següents (tots els angles són en graus):

   a) $\sqrt{1{,}77^2 \cdot 3{,}59^2}$

   b) $e^{\frac{0{,}556}{2}}$

   c) In (9,86)

   d) $\cos\left(\frac{12{,}4}{7{,}3}\right)$

   e) $\tan^{-1}(0{,}7)$

Variables vectorials i funció `vector()`

Sovint treballem amb conjunts de valors d'una variable estadística particular –per exemple alçades i ingressos– i necessitem mantenir-los junts i executar operacions en tot el conjunt de dades.

Per a emmagatzemar un conjunt de valors podem usar la variable vectorial –vector– en el MacAnova. Això ens permet d'assignar més d'un valor a un sol nom. Hi ha diverses maneres de fer-ho. Una manera és usar la funció `vector()`:

```
Cmd> x<-vector(1.2,4.3,2.2,5.1)
```

Això posa tots quatre valors 1,2, 4,3, 2,2 i 5,1 dins la variable `x`. Introduïu l'ordre de més amunt i després mireu què hi ha dins `x` de la manera habitual:

```
Cmd> x
(1)        1.2        4.3        2.2        5.1
```

Els arguments de `vector()` poden ser una variable vectorial mateix, per exemple:

```
Cmd> x2<-vector(x, -x)
Cmd> x2
(1)      1.2   4.3   2.2   5.1   -1.2
(6)     -4.3   -2.2   -5.1
```

Això posa una còpia de `x` i els negatius de tots els valors de `x` dins la variable vectorial `x2`. També podeu veure que els nombres a l'esquerra són els índexs dels nombres dins el vector `x2` –el valor `-4,3` és el sisè valor–. Abans solament teníem resultats que eren nombres simples, de manera que solament teníem un `(1)` a l'esquerra.

Si voleu veure quants valors hi ha emmagatzemats en la variable vectorial, useu la funció `nrows()`, per exemple:

```
Cmd> nrows(x2)
(1)             8
```

Aquesta funció s'anomena `nrows()` perquè els vectors es prenen com a vectors columna*, de manera que el nombre de files es correspon amb el nombre de valors dins el vector.

*** En anglès, les columnes de nombres s'anomenen *rows*.**

Llegir en un conjunt de dades

Podeu emmagatzemar petits conjunts de dades en una variable vectorial usant la funció `vector()`, tal com s'ha descrit abans. Si ja teniu algunes dades emmagatzemades en un arxiu de l'ordinador, aleshores és possible llegir les dades directament dins una variable vectorial fent servir la funció `vecread()`. Per exemple, un dels arxius de dades que us proporciona el disquet del programa (i que ha estat instal·lat en el disc dur del vostre ordinador) és el fitxer `NOTES` (ens podem referir als arxius pels seus noms en majúscules o minúscules). Aquest arxiu conté les puntuacions finals dels exàmens d'estadística d'un grup de 78 estudiants universitaris. Observeu que donem el nom de l'arxiu com un argument entre cometes. Executeu l'ordre que acabem de presentar i després comproveu els valors que hi ha en la variable `notes`:

```
Cmd> notes<-vecread("NOTES")
```

Podeu llegir aquestes dades dins una variable, la qual anomenem *notes*, de la manera següent:

```
Cmd> notes
 (1)   4.15  1.2   7.4   4.95  7.25
 (6)   3.65  4.6   6.9   6.25  2.65
(11)   7.25  4.25  4.15  3.6   6.6
(16)   5.6   5.75  7.9   8.2   7.25
(21)   6.6   5.75  6.4   7.55  9.85
(26)   6.1   7.2   4.6   4.25  4.8
(31)   7.9   8.05  4.45  7.4   7.9
(36)   8.55  4.75  6.6   5.1   8.35
```

```
(41)   5.1   5.9   5.6   4.9   3.15
(46)   2.85  3.65  6.4   7.4   5.1
(51)   5.4   7.05  6.1   8.55  6.1
(56)   4.3   5.1   1.75  7.7   2.15
(61)   5.6   7.25  5.75  8.85  7.25
(66)   5.6   8.55  5.95  5.6   7.4
(71)   7.7   2.95  5.3   5.95  8.55
(76)   4.95  6.25  6.6
```

Funcions simples d'estadística

Un cop heu introduït un conjunt de valors dins una variable vectorial, sigui usant la funció `vector()` o la funció `vecread()` tal com s'acaba de descriure, podeu executar fàcilment alguns càlculs simples d'estadística fent servir les funcions que el MacAnova proporciona. Les que nosaltres considerem aquí són les funcions `describe()`, per a obtenir estadístics bàsics; `stemleaf()`, per a obtenir un gràfic de tiges i fulles, i `boxplot()`, per a obtenir un diagrama de caixa.

1) La funció `describe()`

La funció `describe()` calcula la mitjana aritmètica, la variància, la mediana, el màxim, el mínim i els quartils inferior i superior. Per exemple, si la variable vectorial `notes` conté les puntuacions dels estudiants, nosaltres podem obtenir tots aquests resultats resumits de la manera següent:

```
Cmd> describe(notes)
   component: n
   (1)            78
   component: min
   (1)           1.2
   component: q1
   (1)           4.8
   component: median
   (1)          5.95
   component: q3
   (1)          7.25

   component: max
   (1)          9.85
   component: mean
   (1)        5.9231
   component: var
   (1)        3.2206
```

Els resultats s'expliquen per si sols: el nombre de valors és 78; després tenim els cinc nombres resum que consisteixen en el mínim, 1,2; el primer quartil, 4,8; la mediana, 5,95; el tercer quartil, 7,25, i el màxim, 9,85; finalment, la mitjana aritmètica, 5,9231, i la variància, 3,2206.

2) La funció `stemleaf()`

La funció `stemleaf()` proporciona el gràfic de tiges i fulles d'un conjunt de valors, per exemple:

```
Cmd> stemleaf(notes)
      1     1*|2
      2     1.|7
      3     2*|1
      6     2.|689
      7     3*|1
     10     3.|666
     16     4*|112234
     23     4.|6678999
     29     5*|111134
    (11)    5.|66666777999
     38     6*|1112244
     31     6.|66669
     26     7*|02222224444
     15     7.|577999
      9     8*|023
      6     8.|55558
      1     9*|
      1     9.|8

           1*|1 represents 1.1 Leaf digit unit = 0.1
```

Aquest gràfic de tiges i fulles té una columna extra a l'esquerra, la qual mostra el nombre d'observacions a la cua de la distribució. Per exemple, en la primera línia, el número `1` significa que hi ha una observació inferior a 1,5; en la segona línia, el `2` significa que hi ha dues observacions més petites que 2,0; i així continuaríem (fixeu-vos que aquest gràfic de tiges i fulles té dues ratlles per cada enter de la tija, el primer duu la marca d'un asterisc (*)).

Quan arribem a la novena línia podem veure que hi ha 29 valors més petits que 5,50. En la desena línia, el nombre a l'esquerra és entre parèntesis –això és

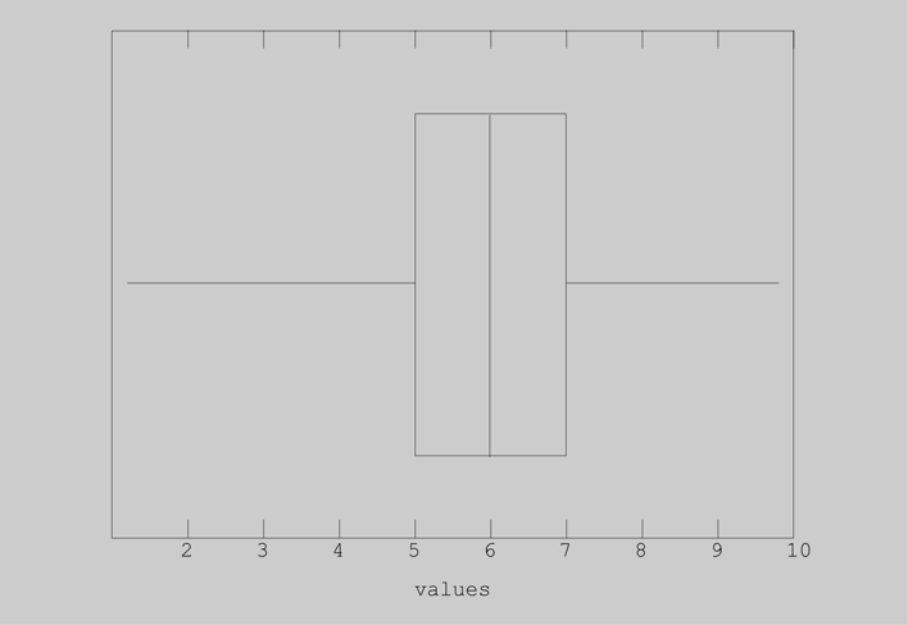
perquè la mediana és en les fulles d'aquest tronc (havíem vist que era 5,95)–, i en aquest cas el nombre de valors que es donen a l'esquerra és el nombre d'observacions en aquesta línia determinada. Per sobre de la mediana tornem a obtenir els nombres d'observacions a la cua de la distribució, la qual és ara el nombre de valors per sobre del valor corresponent de la tija.

Per exemple, en la línia 12 el número `38` significa que hi ha 38 valors més grans o iguals que 6,0; i així successivament. Finalment, al capdavall del diagrama hi ha una clau molt breu per a l'escala de les tiges i les fulles.

3) Funció `boxplot()`

La funció `boxplot()` proporciona el diagrama de caixa d'un conjunt de valors, per exemple:

```
Cmd> boxplot(notes)
```



Executar càlculs amb vectors

Si voleu executar la mateixa operació en cada element d'un vector, ho podeu fer usant una instrucció. Per exemple, restem la mitjana aritmètica de totes les dades dins `notes` de la manera següent:

```
Cmd> notes_centr <- notes-5.9231
```

Noteu que `notes` és un vector, mentre que la mitjana aritmètica és un nombre simple o escalar.

Ara bé, el MacAnova fa l'operació pretesa i aplica la resta a tots els elements del vector `notes`. El resultat del càlcul, assignat a `notes_centr`, és també una variable vectorial, amb el mateix nombre de files que té `notes`.

Activitats

Introduïu l'ordre d'abans i després introduïu `notes_centr` per veure tots els valors "centrats" de les puntuacions dels estudiants.

Després apliqueu la funció `describe()` al vector `notes_centr` per comprovar que té la mitjana aritmètica 0 i la mateixa variància com abans.

2. Ara dividiu `notes_centr` per la desviació estàndard:
   `notes_stand<-notes_centr/sqrt(3.2206)`
   i apliqueu la funció `describe()` al vector `notes_stand`. Fixeu-vos que `notes_stand` té la variància 1.

Calcular la mitjana aritmètica i la desviació estàndard

Ja hem vist que la funció `describe()` proporciona com una part dels resultats la mitjana aritmètica i la variància d'un conjunt de valors. També podem calcular-ho directament des de les fórmules bàsiques:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{i} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Usem la funció `sum()`, la qual suma tots els valors del seu argument. Estudieu cadascuna de les ordres següents amb cura i proveu-les.

```
Cmd> n <- nrows(notes) ; notes_mean <- sum(notes)/n
Cmd> notes_var <- sum((notes-notes_mean)^2)/(n-1)
Cmd> notes_sd <- sqrt(notes_var)
Cmd> notes_mean ; notes_sd
(1)        5.9231
(1)        1.7946
```

- La primera línia guarda el nombre de valors (els quals sabem que són 78 en aquest cas) en la variable `n`, i després calcula la mitjana aritmètica com la suma de tots els valors dins `notes` i la divideix per `n`. La mitjana aritmètica és guardada en una variable que nosaltres anomenem `notes_mean`.

Fixeu-vos que podem usar el caràcter subratllat (_) en un nom, una cosa útil per a fer més llegible un nom.

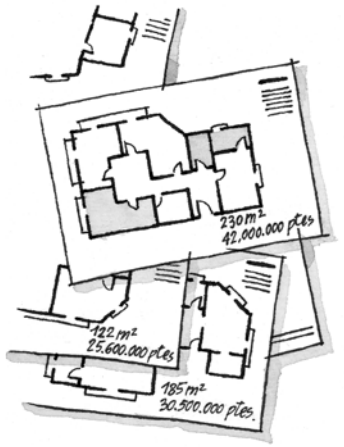
- La segona línia calcula la variància restant primer la mitjana aritmètica de tots els valors dins `notes`, fent el quadrat d'aquestes diferències i sumant-les totes, i finalment dividint per `n-1`. Noteu que quan prenem el quadrat d'una variable vectorial d'això en resulta el vector dels quadrats de tots els elements individuals. El resultat s'emmagatzema en la variable `notes_var`.

- En la tercera línia calculem la desviació estàndard com l'arrel quadrada de la variància i emmagatzemem el resultat dins `notes_sd`.

Calcular el coeficient de correlació

Vegem ara un exemple més complicat: calcularem amb MacAnova el coeficient de correlació entre els preus de vuit pisos i les seves mides:

Preu (en milions de pessetes)	Mida (en metres quadrats)
25,6	122
27,8	121
27,7	140
30,4	156
30,5	185
37,0	240
41,8	200
42,0	230



El càlcul de la correlació ens permet avaluar, entre altres coses, la relació entre la mida dels pisos i el seu preu.

Primer mostrarem les ordres del MacAnova per a calcular la correlació a partir de la fórmula original (tenim les dades del preu i la mida en dos arxius anomenats `price` ('preu') i `size` ('mida'), i es poden llegir utilitzant la funció `vecread()`):

```
Cmd> x<-vecread("size")
Cmd> y<-vecread("price")
Cmd> describe(x,mean:T,var:T)
component: mean
(1)       174.25
component: var
(1)       2183.1
Cmd> describe(y,mean:T,var:T)
component: mean
(1)        32.85
component: var
(1)       42.451
Cmd> sum((x-174.25)*(y-32.85))/((7*sqrt(42.451*2183.1))
(1)      0.87143
```

Els dos primers arguments calculen les mitjanes i les variàncies de les dues variables (les quals posem en els vectors `x` i `y`). El tercer argument calcula la fórmula.

Recordeu

Per definir el coeficient de correlació r ho fem de la manera següent:

$$r = \frac{\mathrm{cov}(x,y)}{s_x s_y}.$$

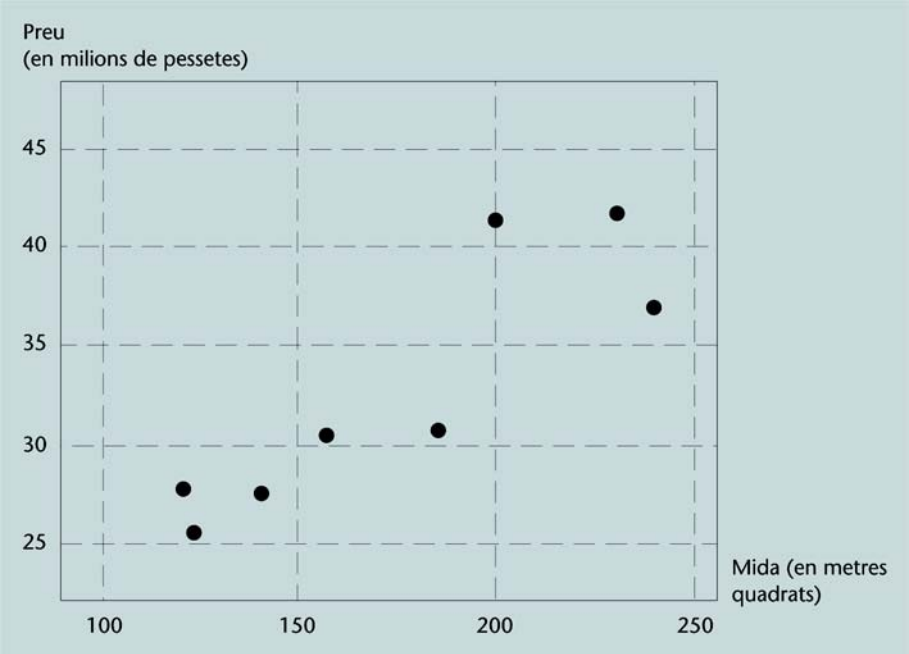
Per sort, el MacAnova té una funció especial per a calcular correlacions, `cor()`. Aquesta funció de fet calcula el que anomenem *matriu de correlacions* per a un conjunt de variables observades en les mateixes persones o unitats

mostrals. En el nostre cas, aquí solament tenim dues variables que ens donaran una matriu de correlacions amb dues files i dues columnes:

```
Cmd> cor(x,y)
(1,1)             1       0.87143
(2,1)       0.87143             1
```

La matriu mostra les correlacions entre totes les parelles de variables, incloenthi les variables amb si mateixes. Per tant, veiem correlacions d'1 sobre la diagonal d'aquesta matriu, les quals són correlacions entre la variable preu i si mateixa, i entre la variable mida i si mateixa. Fora de la diagonal veiem la correlació entre la variable preu i la variable mida en la posició del capdamunt i a la dreta, i la correlació entre la variable mida i la variable preu en la posició del capdavall i a l'esquerra –és clar que aquestes correlacions són idèntiques.

Per tant, veiem que la correlació entre el preu dels pisos i la mida que tenen és alta, 0,871. Ho podem confirmar representant les dades en un gràfic de dispersió.



Fixeu-vos que el valor del coeficient de correlació diu solament quina és la proximitat dels punts pel que fa a una recta, i el signe del coeficient de correlació indica si la recta puja o baixa. El coeficient no us diu com és el pendent de la recta, si és molt inclinat o no ho és.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Iniciar i tancar el MacAnova: teclejar `macanova` i `quit` respectivament.

Constants matemàtiques: `PI`, `E`.

Funcions matemàtiques: `log()`, `sin()`, `exp()`, `sqrt()`.

Canviar angles de radians a graus:
`setoptions (angles: "degrees").`

Aconseguir ajuda: `help()`.

Posar un conjunt de dades en un vector: `vector()`, `vecread()`.

Altres funcions d'un vector de valors: `nrows()`, `sum()`.

Funcions d'estadístics descriptius d'un vector de valors:
`describe(), stemleaf(), boxplot()`.

Coeficient de correlació: `cor()`.

7. Recollida de dades (I): cens i mostreig

En els pròxims apartats veurem diverses maneres d'obtenir dades. En alguns casos molt especials és possible obtenir dades de cada unitat en la població d'interès. En la pràctica, però, en la majoria de situacions solament és factible la recollida de dades d'un conjunt més petit d'unitats, anomenat **mostra**. Argumentarem diverses maneres d'obtenir una mostra d'una població.

En aquest apartat sobre la recollida de dades aprendrem:

- què és un cens de la població;
- què és una mostra;
- la manera com les empreses industrials usen el mostreig per a controlar la qualitat dels productes;
- què és una mostra aleatòria simple;
- com se selecciona una mostra aleatòria simple d'una població coneguda, usant una taula de nombres aleatoris;
- com se selecciona una mostra aleatòria sistemàtica d'una població coneguda.

Població i cens

Una situació comuna que afrontem és provar de recollir informació d'un gran conjunt d'unitats, per exemple: totes les persones residents a Catalunya o tots els bancs d'Espanya. Quan s'emprèn un projecte de recerca, normalment va dirigit a un conjunt total d'unitats; per exemple: podríem estar interessats a estudiar la rendibilitat de les empreses industrials d'Espanya. Si es tinguessin uns recursos il·limitats per a tal estudi, es podria considerar contactar amb cadascuna de les empreses industrials d'Espanya per a esbrinar les xifres clau del seu moviment d'efectiu, inversions, facturació, etc.; després d'això es podria confeccionar un informe definitiu sobre l'estat financer d'aquestes empreses.

Anomenem **població** la totalitat de les unitats que estem interessats a estudiar i **cens** un estudi de tota la població.

El context usual d'aquests termes és quan una població és una població humana i un cens és el procés de recollida de dades de cada membre de la població. Però aquí l'usem en un sentit més general.

Per exemple, una població podria ser qualsevol de les que presentem a continuació:

- tots els terratrèmols que hi ha hagut a la regió mediterrània;
- tots els estudiants de la Universitat Oberta de Catalunya en el curs 1997-1998;
- tots els informatius de TV3 durant el període de la campanya d'eleccions municipals.

El cens a Espanya

Els **censos de població** a Espanya s'efectuen cada deu anys, aproximadament, seguint les normatives de l'ONU i la Comunitat Europea.

El cens recull, elabora, valora i analitza les dades referents a la demografia i els trets culturals i socials de tots els habitants d'un país en un moment determinat del temps.

Qüestionari censal

Les preguntes del qüestionari censal les contesten confidencialment els ciutadans. Les dades obtingudes s'informatitzen d'una manera anònima sense inclusió de noms, cognoms o adreces.

En el cas de Catalunya (i Espanya), l'últim cens es refereix a les persones que tenien la residència fixada o es trobaven a Catalunya a les zero hores del dia primer de març de 1991. Aquest cens va seguir el procés d'elaboració següent:

- Entre març i juny de 1991 es van recollir les dades del cens i es va renovar el padró municipal d'una manera simultània.
- En un primer estadi es van obtenir les dades provisionals recomptant els quaderns de treball de camp dut a terme per entrevistadors i entrevistadores porta a porta.
- L'Institut Nacional d'Estadística (INE) va comunicar, com ho fa sempre, els resultats provisionals als ajuntaments, els quals els sotmeten a l'aprovació dels plens municipals.
- Els ajuntaments van deixar les dades exposades, i els habitants van comprovar el padró i van esmentar altes i baixes.
- Finalment es va elaborar la proposta oficial que l'INE, després d'una comprovació prèvia, usa per a publicar les dades finals i definitives.

Lectura complementària

Si voleu més informació sobre les dades i el procés d'elaboració d'aquest darrer cens, podeu llegir:

Cens de Població 1991/(1992), vol. 2: *Xifres oficials. Dades municipals*. Barcelona: Institut d'Estadística de Catalunya, Generalitat de Catalunya (col·lecció Estadística Demogràfica, Censos).

Mostres

Una **mostra** és una col·lecció parcial d'unes quantes unitats de la població.

Per a les poblacions esmentades abans podríem tenir les mostres següents:

- escoles primàries a Catalunya;
- centres esportius dins l'àrea metropolitana de Barcelona;
- cada desè estudiant de la UOC en el curs 1997-1998, agafat d'una llista alfabètica;
- els informatius del capvespre de TV3.

És clar que hi ha moltes maneres de triar una mostra, i algunes són millors que d'altres per a estudiar una certa situació. També hi ha consideracions pràctiques a l'hora de triar una mostra: estudiar la població sencera costaria massa diners i ens hi passaríem massa temps. El cens espanyol és molt car, i dur-lo a terme requereix una vasta mà d'obra. Encara que tenir a l'abast tota la informació sobre la població sencera és molt útil, en la pràctica podem aconseguir la informació que necessitem a partir d'una mostra més petita de la població.

Mostres representatives

Com que la informació de la mostra servirà per a treure conclusions sobre la població, és extremament important que la mostra sigui representativa de la població. Idealment, la mostra hauria de ser com la població en tots els aspectes –excepte, és clar, el fet que és més petita–. Si la mostra no reflecteix acuradament la població, aleshores els errors són fàcils de cometre.

Per exemple, si únicament tenim dades de les empreses industrials de Barcelona, aleshores les nostres conclusions no són aplicables a les empreses espanyoles en general. Aquest punt sembla obvi, però us sorprendrà veure quantes vegades les conclusions extretes d'unes mostres no representatives s'estenen a unes poblacions més àmplies.

Quan la mostra no és representativa de la població, diem que les nostres conclusions pot ser que siguin *esbiaixades*.

Parlarem sobre el concepte de mostra esbiaixada més detalladament en l'apartat 8 d'aquest mòdul.

Mostres aleatòries

Podem fer que una mostra sigui representativa de maneres ben òbvies. Per exemple, si sabem amb antelació que el 60% dels estudiants de la UOC són dones i el 40% són homes, aleshores podem triar una mostra dels estudiants amb aquesta mateixa proporció de dones i homes. La representativitat, però, no és suficient per a assegurar una mostra bona. Els mecanismes pels quals triem cada unitat per a incloure-la en la mostra haurien de ser totalment fora del nostre control. Això ens porta a un dels conceptes més importants de l'estadística: la **mostra aleatòria**.

Una mostra aleatòria ha de complir les dues propietats següents:

1) Cada unitat de la població té la mateixa probabilitat de ser representada dins la mostra.

2) Les unitats de la mostra es trien independentment les unes de les altres.

La primera propietat és necessària per a assegurar que tots els membres de la població reben el mateix tracte. Això garanteix que no hi haurà cap tendència a discriminar en favor o en contra de cap unitat de la població. La segona propietat és més subtil, i és necessària perquè la mostra contingui tanta informació útil com sigui possible. També és un requisit per als tests estadístics que presentarem en apartats posteriors.

La tria d'una mostra aleatòria

La millor manera de triar una mostra aleatòria és usar una font externa d'aleatorietat, com ara a cara o creu, una taula de nombres aleatoris o una selecció aleatòria garantida per mitjà d'un ordinador.

En l'annex 2, al final d'aquest mòdul, trobareu nombres aleatoris que podeu usar per a generar una petita mostra aleatòria.

Suposem que volem treure una mostra de 20 estudiants de la UOC. Tenim una llista amb els noms i les adreces de tots els estudiants, i a la nostra llista n'hi ha 2.150. Per al propòsit d'aquest exercici numerem els estudiants en aquesta llista de l'1 al 2.150. Necessitarem nombres aleatoris de quatre dígits, per tant, usem els nombres aleatoris de la nostra taula de quatre dígits simultanis. Primer tirem un dau per veure on comencem. Suposem que traiem un 4, i comencem en el quart dígit de la taula, que és 2.

A partir d'una llista completa dels estudiants de la UOC, per exemple, podem obtenir una mostra aleatòria.

19223	95034	05756	28713	96409	12531	42544	82853

▲
punt de partida

El primer nombre que formem amb quatre dígits començant per aquest 2 és, per tant, 2395. Com que és fora del camp dels nombres dels estudiants, el saltem. El pròxim conjunt de quatre dígits és 0340, de manera que seleccionem l'estudiant número 340 com el primer estudiant de la nostra mostra. Després saltem els números 5756, 2871 i 3964. El següent nombre és dins el camp numèric que necessitem; per tant, el segon estudiant de la nostra mostra és el número 912. Després saltem els números 5314, 2544, 8285, 3736 i 7647 abans d'aconseguir el número 1509; per tant, el tercer estudiant de la nostra mostra és el número 1.509. Continuem d'aquesta manera fins a tenir 20 estudiants a la mostra.

Activitats

1. Continueu de la mateixa manera fins a trobar els pròxims tres estudiants o estudiantes de la mostra.

L'ús del MacAnova per a triar una mostra

W1/00520.01
Poseu l'ordinador en marxa i inicieu el programa MacAnova.

El sistema d'abans, d'ús d'una taula de nombres aleatoris, no és gaire eficient, com segurament ja heu notat. Hi ha moltes seqüències de quatre dígits en la taula que queden fora del camp numèric que necessitem i, si havíem de triar una mostra gran, per exemple de 100 estudiants, aleshores seria un procés realment pesat. Podem usar el programa MacAnova per a generar nombres aleatoris per a nosaltres i calcular els nombres seqüencials dels estudiants a la mostra.

El MacAnova té una funció `runi()` que genera els anomenats **nombres aleatoris uniformes** entre 0 i 1. També podríeu tenir una calculadora específica amb una tecla de funció que proporciona el nombre aleatori. Nosaltres ara obtindrem els nombres seqüencials de la nostra mostra d'una manera diferent de la descrita abans. Primer veurem com funciona `runi()`.

`runi` és l'abreviació de *random uniform*.

Necessitarem almenys 20 nombres aleatoris uniformes, que podem obtenir usant l'ordre `runi(20)`; però també els volem retenir per a càlculs posteriors, de manera que teclegem l'ordre següent:

```
Cmd> rand <- runi(20)
NOTE: random number seeds set
to 848561694 and 904809327
```

Per a veure els nombres emmagatzemats dins `rand`:

```
Cmd> rand
 (1)      0.27323  0.030781  0.2935   0.7375   0.18599
 (6)      0.42633  0.89986   0.55211  0.26319  0.95351
(11)      0.12608  0.49387   0.21837  0.39074  0.086703
(16)      0.42135  0.24914   0.78911  0.30629  0.91501
```

Després de la primera ordre veiem una resposta sobre quines han estat les bases de nombres aleatoris en aquest cas –fixeu-vos que les vostres bases seran diferents d'aquestes i també que el vostre conjunt de 20 nombres aleatoris també serà diferent d'aquests–. Els nombres dins `rand` s'estenen aleatòriament entre 0 i 1. Nosaltres volem triar nombres aleatoris de l'1 al 2.150 per seleccionar la mostra d'estudiants. El primer pas és multiplicar rand per 2.150:

Cada persona que faci servir l'ordre `runi()` al MacAnova obtindrà un resultat diferent.

```
Cmd> rand*2150
 (1)   587.44    66.18     631.03    1585.6    399.88
 (6)   916.6     1934.7    1187      565.87    2050
(11)   271.06    1061.8    469.49    840.08    186.41
(16)   905.91    535.64    1696.6    658.51    1967.3
```

Això ens proporciona nombres entre 0 i 2.150. Per a obtenir enters de l'1 al 2.150, hauríem d'arrodonir cada nombre al proper enter més gran: 587,44 esdevé 588; 66,18 esdevé 67, i així successivament. Hi ha una funció dins el MacAnova, anomenada `ceiling()`, que arrodoneix a l'enter més gran:

La paraula *ceiling* vol dir 'sostre'.

```
Cmd> ceiling(rand*2150)
 (1)     588       67        632       1586      400
 (6)     917       1935      1188      566       2051
(11)     272       1062      470       841       187
(16)     906       536       1697      1697      1968
```

Els nombres seqüencials d'abans són els de la nostra mostra. Podríem obtenir-ne la llista sencera amb una sola ordre:

```
ceiling(runi(20)*2150)
```

si no volíem emmagatzemar els nombres aleatoris. Si repetíem aquesta ordre, és clar, obtindríem una mostra diferent.

Activitats

2. Suposem que tenim una llista de 92 persones que treballen en una empresa i en volem obtenir una mostra aleatòria de 15. Digueu l'ordre del MacAnova que usaríeu per a obtenir la seqüència de nombres de la mostra.

Aquí teniu un altre exemple de l'obtenció d'una mostra, aquest cop d'una llista de valors que ja heu emmagatzemat en un fitxer d'ordinador. Les dades de la taula I del segon apartat, els coeficients intel·lectuals de 60 estudiants, són en un fitxer anomenat `IQ` i aquest és en el vostre programa MacAnova. Suposem que volem seleccionar 10 d'aquests valors aleatòriament. Les ordres següents ens permetrien de fer-ho:

Vegeu la taula I de l'apartat 2 d'aquest mòdul.

```
Cmd> rand <- runi(10)
Cmd> ind <- ceiling(rand * 60)
Cmd> iq <- vecread("IQ")
Cmd> sample <- iq[ind]
```

- La primera ordre posa 10 nombres aleatoris uniformes dins `rand`.
- La segona posa els nombres seqüencials entre 1 i 60 dins `ind`, tal com ja hem descrit abans.
- La tercera ordre permet de llegir els 60 valors de coeficient intel·lectual del fitxer `IQ`.
- La quarta ordre inclou una nova estructura, els claudàtors després del nom del vector: `iq[ind]`. És la manera amb què fem referència als elements específics d'un vector; per exemple: `iq[1]` és el primer valor dins `iq`, el qual és 120; `iq[23]` és el vint-i-tresè valor, el qual és 124, i així successivament. El nombre entre claudàtors és l'**índex del vector**. Si, en comptes d'un sol nombre, posàvem un vector de valors indicadors en claudàtors, aleshores obtindríem un vector de tots els valors corresponents. Per tant, si `ind` és un vector que conté els tres valors 24, 7 i 41, aleshores `iq[ind]` és el vector `iq[24]`, `iq[7]` i `iq[41]`. La millor manera de comprendre això és provant-ho.

Activitats

3. Executeu les ordres del MacAnova que acabem d'explicar i comproveu els continguts de `rand`, `ind`, `iq` i `sample` respectivament després de cada ordre.

Ordres d'ús del MacAnova

a. `runi()` genera nombres aleatoris uniformes; per exemple: `runi(10)` genera 10 nombres aleatoris entre 0 i 1.

b. `ceiling()` arrodoneix cap a l'enter més gran.

c. `x[i]` proporciona l'element `i`-èssim d'un vector `x`; si `i` és un vector, aleshores obtenim el vector dels elements corresponents de `x`.

Mostra aleatòria sistemàtica

Quan tenim una llista que conté la nostra població, aleshores hi ha una manera més senzilla i convenient d'obtenir una mostra aleatòria que en la pràctica s'usa sovint. Una guia telefònica és un bon exemple de la llista que podríem tenir quan la població d'interès consisteix en tots els abonats al servei telefònic.

Suposem que hi ha 834.781 abonats en les guies de telèfons de Barcelona i que volem una mostra aleatòria de 400 abonats per a trucar i preguntar-los si estan satisfets amb el servei d'informació telefònic que proporciona el 003. Com que 834.781 dividit per 400 és aproximadament 2.086, podem agafar cada 2.086è nombre de la guia, i això ens donarà una mostra de 400 abonats estesos al llarg de totes les entrades de la guia. Per a començar la selecció triem un nombre a l'atzar entre 1 i 2.086, usant l'ordre del MacAnova `ceiling(runi(1)*2086)`; suposem que aquest nombre és el 731. Busquem el nombre 731è en la guia i després l'entrada nombre 731 + 2.086 = 2.817, després 2.817 + 2.086 = 4.903, i així successivament.

Això és el que s'anomena **mostra aleatòria sistemàtica** –comencem en algun punt a l'atzar a prop del començament de la llista i després continuem sistemàticament al llarg de tota la llista a intervals fixos–. Els intervals es calculen per a permetre l'entrada del nombre que es desitja dins la mostra, després de repassar tota la llista.

S'hi poden introduir algunes dreceres de sentit comú per a fer la tasca una mica més senzilla. Comptar 2.086 entrades en la guia cada vegada és pesat, i un petit canvi en el disseny del mostreig anterior no hi resta validesa, sempre que el canvi s'estableixi a l'inici, abans que el mostreig comenci. Per exemple, suposem que, en comptar quantes entrades hi ha en unes quantes pàgines de la guia, trobem que la mitjana és de 205 entrades per pàgina, és a dir: 2.086 entrades fan unes 10 pàgines, amb un restant de 81. Després, des del punt inicial del mostreig, simplement compteu 10 pàgines en la mateixa posició de la pàgina i després compteu 81 entrades per arribar a la unitat següent de la mostra. Encara seria més senzill si usàveu un disseny de mostra en múltiples etapes.

Les mostres en múltiples etapes s'exposen a l'apartat 8.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Població: nombre total d'unitats (per exemple: persones, productes, etc.) que ens interessa estudiar.

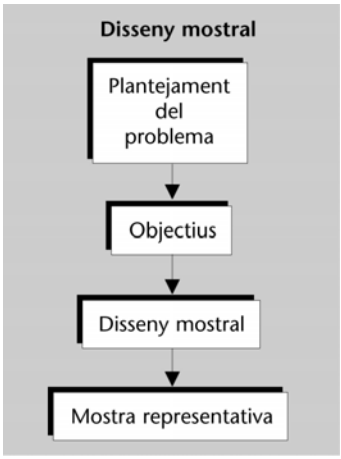
Cens: estudi de la població.

Mostra: col·lecció parcial d'unitats d'una població.

Mostra aleatòria: mostra obtinguda d'una població en què cada unitat d'aquesta població té la mateixa oportunitat de ser dins la mostra i en què cada unitat del mostreig es tria independentment de les altres.

Nombres aleatoris uniformes: nombres aleatoris distribuïts homogèniament entre 0 i 1.

8. Recollida de dades (II): enquestes per sondatge



En aquest apartat parlem de diferents dissenys de mostreig per a obtenir dades d'una gran població. Cadascun té el mostreig aleatori com a principi fonamental, però els dissenys difereixen d'acord amb les consideracions pràctiques i d'acord amb la quantitat d'informació que tenim sobre la població. L'objecte d'aquestes estratègies de mostreig és sempre assegurar que la mostra és representativa de tota la població a l'hora de fer la tria de l'aleatorietat de cada unitat. També veiem breument la manera com les dades es recullen durant les enquestes socials i com els estils de les entrevistes influeixen sobre la veracitat de les respostes.

En aquest apartat sobre la recollida de dades aprendreu:

- què és una mostra en etapes múltiples;
- què és una mostra estratificada;
- què és una mostra de quota;
- la importància de la formació del personal de camp abans de dur a terme una enquesta social o comercial.

Mostres en múltiples etapes

Si la població és petita i ben definida, per exemple la població dels estudiants de la UOC, aleshores triar una mostra aleatòria de la manera que hem descrit en l'apartat anterior és força senzill. En situacions més complexes –per exemple, prenent una mostra dels residents de Catalunya– no és fàcil recollir una mostra aleatòria simple. Primer hi ha el problema de tenir una llista de tots els residents perquè puguem generar nombres aleatoris per a seleccionar la mostra, però les dificultats pràctiques quant a contactar amb les persones enquestades encara són més problemàtiques. En aquestes situacions preferim desglossar l'estructura de la població en unitats més petites, més manejables, i després dur a terme el mostreig aleatori sobre aquestes unitats.

Per exemple, les diferents etapes del mostreig serien:

1) Dividir Catalunya en comarques i després triar-ne una mostra aleatòria (això encara fa que l'enquesta sigui més fàcil en la pràctica, ja que no haurem de visitar cada comarca per a dur-la a terme).

2) De les comarques seleccionades en aquesta primera etapa triem una mostra aleatòria de poblacions.

3) Si tenim accés a una llista de noms i adreces de les persones residents en aquests pobles i ciutats, podrem agafar una mostra aleatòria de residents i dur a terme l'enquesta.

D'altra banda, podríem simplificar la nostra tasca seleccionant a l'atzar unes quantes zones de cada població, sobre el plànol, i després, en l'última etapa, seleccionar llars d'una manera aleatòria. Aquest tipus de disseny mostral s'anomena **mostra en múltiples etapes**.

Tornem al mostreig sistemàtic de la guia de telèfons de què parlàvem en l'apartat anterior. Una manera d'agafar una mostra en múltiples etapes seria fer el mostreig de les pàgines en la primera etapa i després dins les pàgines en la segona etapa. Per exemple, per a obtenir una mostra aleatòria de 400 números de telèfon, en la primera etapa prendríem una mostra aleatòria sistemàtica de 100 pàgines de la guia. Després obtindríem una mostra aleatòria de 4 números de telèfon de cada pàgina. Això demostra com el pragmatisme del sentit comú, combinat amb l'ingredient essencial –el mostreig aleatori–, porten a dissenyar un mostreig que és alhora vàlid i pràcticament factible.

Mostres estratificades

Malgrat que el mostreig aleatori assegura la representativitat, no la garanteix totalment. Suposem que tornem a fer un mostreig de la població d'estudiants de la UOC i que el 75% d'ells tenen feina i el 25% no. Prenem una mostra aleatòria de 100 estudiants i veiem que en conté el 40% sense feina. Pot ser que tinguem una pregunta en l'enquesta sobre el nivell d'ingressos de cada estudiant –és clar que obtindrem un mesurament esbiaixat d'ingressos basat en la nostra mostra, la qual sobrerepresenta els estudiants sense feina–. És possible aplicar un reajustament estadístic per a corregir aquest biaix o error sistemàtic un cop les dades són recollides, però és de bon tros preferible recollir una mostra més representativa des del principi.

Exemple

Dins cada comarca de Catalunya podríem tenir una classificació de les ciutats i pobles en termes del seu nombre de residents: menys de 5.000 habitants i més de 5.000. Aleshores podríem escollir una mostra aleatòria de ciutats i pobles més petits i més grans separadament per assegurar-nos que, per atzar, no obtenim una mostra que conté massa ciutats o pobles petits.

Si sabem amb antelació que tenim subgrups importants en la nostra població, com ara amb feina i sense, i aquests subgrups són rellevants per al nostre estudi, aleshores els podem tenir en compte en el nostre disseny del mostreig. Això s'anomena **estratificació**.

Els subgrups amb feina i sense de la nostra població d'estudiants s'anomenen **estrats de la població**. Podem assegurar-nos que els estrats són prou representats en la mostra prenent una mostra aleatòria de cada estrat. Per tant, per a arribar a una mostra de 100 estudiants hauríem de prendre una mostra aleatòria de 75 estudiants del grup amb feina i 25 del grup sense feina.

L'estratificació es pot aplicar dins un disseny de mostreig en múltiples etapes.

Mostres de quota

Fins i tot el mostreig en múltiples etapes a vegades és massa difícil i car per a algunes empreses. Moltes empreses dedicades a fer sondejos prefereixen usar un disseny de mostreig encara més simple anomenat **mostra de quota**. Aquest disseny també requereix saber algunes característiques bàsiques de la població, com ara la distribució de les edats, la zona de residència i el nivell d'educació.

Per exemple, suposem que sabem que el 18% de la nostra població ha acabat l'educació primària solament; el 65%, l'educació secundària, i el 17%, algun nivell més alt d'educació. Si necessitem una mostra de 1.000 persones enquestades, aleshores caldrà que obtinguem al voltant de 180, 650 i 170 persones en els respectius grups d'educació. El personal de camp rep quotes específiques que haurà de complir per a aconseguir aquestes proporcions.

Exemple

L'investigador de camp potser haurà d'obtenir respostes únicament de persones amb un nivell d'educació secundari. Si selecciona una llar de manera aleatòria i troba un entrevistat potencial que és en un altre grup d'educació, aleshores aquest no serà inclòs a l'estudi.

El mostreig a la pràctica

Un altre exemple n'il·lustrarà la idea. Es va dur a terme un estudi entre l'alumnat de dret a la Universitat Pompeu Fabra. Vam decidir estratificar la població segons el curs universitari i el sexe. Pel que fa a aquestes dues variables, la població tenia l'estructura següent:

Una pràctica de mostreig es va dur a terme entre els estudiants de dret de la Universitat Pompeu Fabra.

	home	dona
1r curs	67	105
2n curs	102	136
3r curs	137	194
4t curs	44	75

Com que volíem al voltant de 200 alumnes en la mostra, vam decidir seleccionar una cinquena part de l'alumnat de cada cel·la de la taula, cosa que donava les característiques següents a la mostra:

	home	dona
1r curs	13	21
2n curs	20	27
3r curs	27	39
4t curs	9	15

En aquest punt teníem dues eleccions:

1) Des d'un punt de vista teòric, el millor plantejament era fer servir una llista de tot l'alumnat i seleccionar a l'atzar una cinquena part de cada grup. Per exemple, necessitàvem la llista dels 67 alumnes de sexe masculí de primer curs i després en vam fer una selecció aleatòria de 13 d'aquests. Si fèiem això, ha-

víem de contactar amb cada alumne de la mostra individualment per a obtenir les seves respostes per al nostre qüestionari.

2) Des d'un punt de vista pràctic, però, fer servir una mostra de quota era molt més senzill. Necessitàvem 13 alumnes de sexe masculí de primer curs i vam enviar el personal de camp amb l'objectiu exprés de trobar-los. Com que se sabia el lloc on els alumnes de primer curs tenien les classes, hi havia la possibilitat d'acostar-s'hi i obtenir enquestats que s'ajustessin a la descripció de la quota. Malgrat que aquest era un sistema molt més convenient de mostreig, era obert a tota mena de biaixos.

Per exemple, el personal de camp havia de tenir en compte que, si s'acostava a un grup d'alumnes, aleshores era preferible triar-ne solament un per a la mostra. Els alumnes d'un mateix grup podien tenir tendència a uns mateixos punts de vista, i, així, les seves respostes podien mancar d'independència.

La validesa de l'estudi depèn moltíssim de la validesa de la selecció de la mostra. Com més defensem les unitats de mostreig perquè siguin seleccionades aleatòriament i independentment, més vàlida serà la mostra. L'observació que provem de fer aquí, però, és que hi ha consideracions pràctiques en cada situació, i a vegades hem de transigir per a fer la recerca factible.

Activitats

1. Com a projecte, us demanem dur a terme un estudi dels estudiants de la UOC per tal de saber si estan satisfets amb el seu curs d'estadística. La UOC consent a fer-vos accessible certa informació sobre la població d'estudiants, si la necessiteu. Descriviu el tipus d'informació que voldríeu tenir i esbosseu el disseny d'una mostra que us sembli apropiada a aquesta situació.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Mostra en múltiples etapes: mostra sobre una població dividida en grups successius; per exemple: una població dividida en regions; després, en ciutats i pobles; pobles i ciutats, en barris, i barris, en llars. La mostra s'obté fent una selecció aleatòria de cada grup successiu.

Mostra estratificada: mostra sobre una població classificada segons una variable categòrica o més –per exemple: grups d'edat i nivells d'educació–, en la qual la representació equitativa dels grups està assegurada en la mostra.

Mostra de quota: disseny mostral més pragmàtic en què el personal de camp mostreja els estrats individuals específicament.

9. La distribució normal (I): corbes de densitat normal

Fins ara hem parlat de la distribució d'un conjunt de dades i hem vist diverses maneres de descriure una distribució. Ara fem atenció a un punt més abstracte, és a dir, la distribució teòrica d'una població de la qual provenen les nostres observacions. El tipus més comú de distribució teòrica s'anomena **distribució normal**. Al llarg d'aquesta assignatura veurem els diversos usos i avantatges de la distribució normal.

En aquest apartat sobre la distribució normal aprendreu:

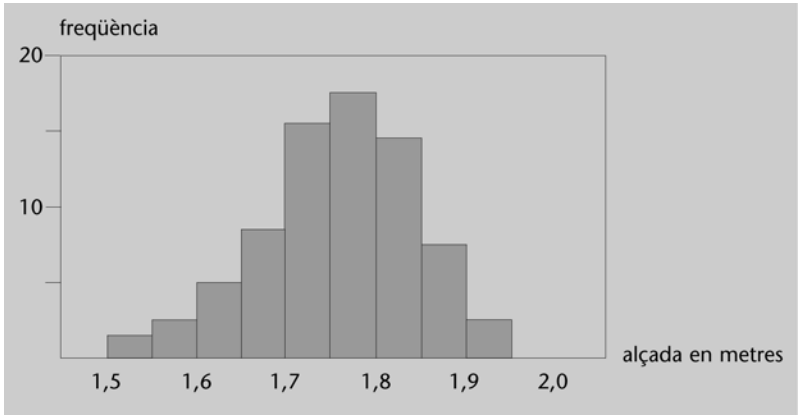
- com es poden definir els histogrames com a densitats de freqüència relativa;
- què és una densitat de probabilitat;
- què és una corba de densitat normal;
- com es calculen algunes àrees sota la corba de densitat normal corresponents a una, dues i tres desviacions estàndard de la mitjana aritmètica.

Histogrames que mostren la freqüència relativa

Tots els exemples d'histogrames que hem vist han estat representacions gràfiques d'un nombre relativament petit d'observacions. Per exemple, si consideréssim un conjunt d'alçades d'home, l'histograma podria tenir l'aspecte següent:

Amb el conjunt de dades de les diferents alçades d'un grup d'homes podem construir un histograma.

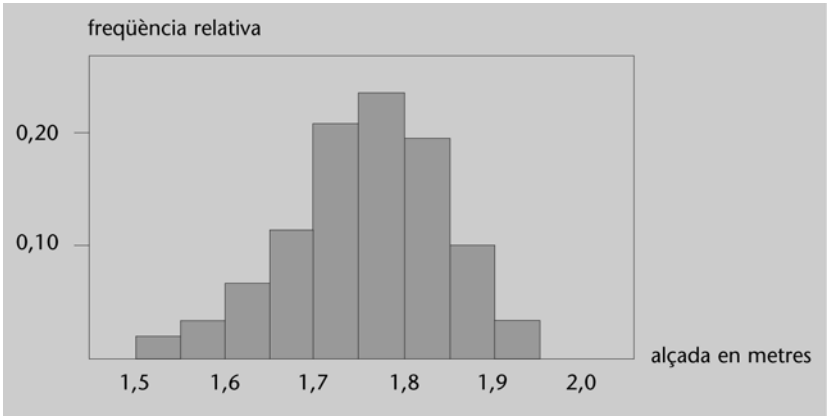
Gràfic I



L'escala vertical és en unitats de freqüències absolutes, per exemple: podem veure que hi ha 15 homes amb una alçada entre 1,70 i 1,75 m. És més convenient expressar l'escala vertical en unitats de freqüència relativa, o proporci-

ons. Això vol dir que calculem les proporcions dels homes en cada classe de l'histograma, i això determina l'alçada de les barres:

Gràfic II

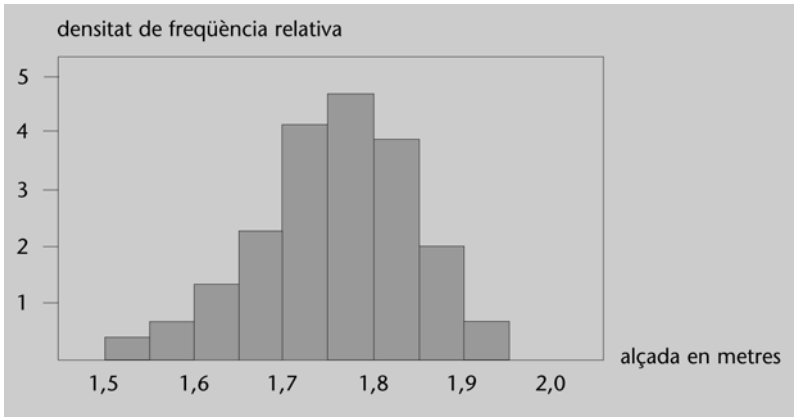


Fixeu-vos que la forma de l'histograma és idèntica, l'única cosa que ha canviat és l'escala vertical. En comptes de ser en freqüències absolutes, ara és en proporcions del recompte total *n*.

Histogrames que mostren la densitat de freqüència relativa

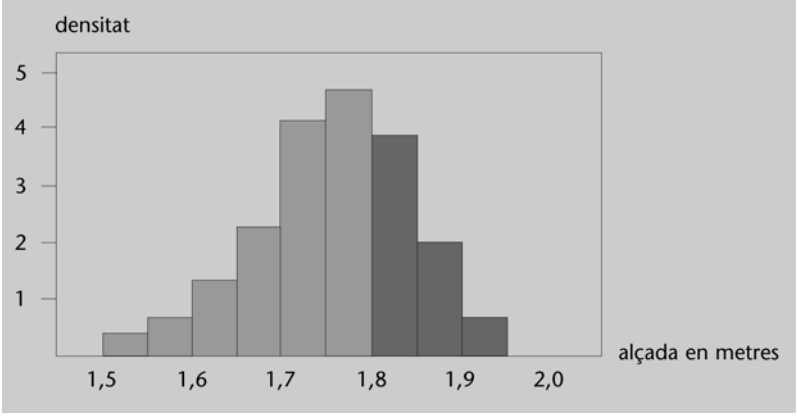
Ara volem que l'àrea de totes les barres juntes sigui 1. Sabem que la suma de les altures de les barres en el gràfic II és 1, de manera que, si l'amplada de cada barra fos 1 unitat sobre l'escala horitzontal (metres), aleshores l'àrea total de les barres seria 1. Ara bé, l'amplada de cada barra és de 0,05 m (5 cm), de manera que l'àrea de totes les barres és 0,05. Per tant, fer que l'àrea d'aquest histograma sigui 1 és senzill: canviem l'escala vertical multiplicant-la per 20.

Gràfic III



Això s'anomena **histograma de densitat de freqüència relativa**, o histograma de densitat. En comptes de llegir les altures de les barres per a avaluar la proporció d'homes en certes classes d'alçada, la nostra regla ara és calcular l'àrea de les classes corresponents. Imaginem-nos que en el gràfic III volem calcular quina proporció d'homes tenen alçades superiors a 1,8 m; avaluarem l'àrea de la zona ombrejada en el gràfic IV.

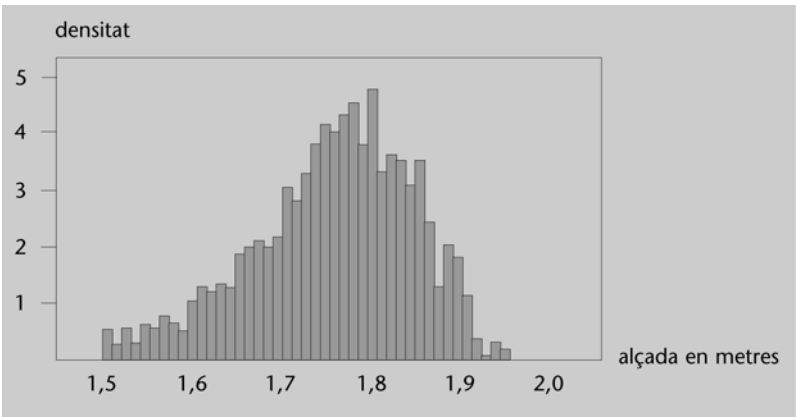
Gràfic IV



Densitats de probabilitat

Ara imaginem-nos que tinguéssim més i més observacions d'alçades d'homes. Com més dades tindrem, més petits podrem fer els intervals de classe. Suposem que tinguéssim uns quants milers d'observacions i poguéssim definir unes classes molt estretes, per exemple: intervals d'1 cm (0,01 m). La densitat de freqüència ara podria tenir aquesta aparença:

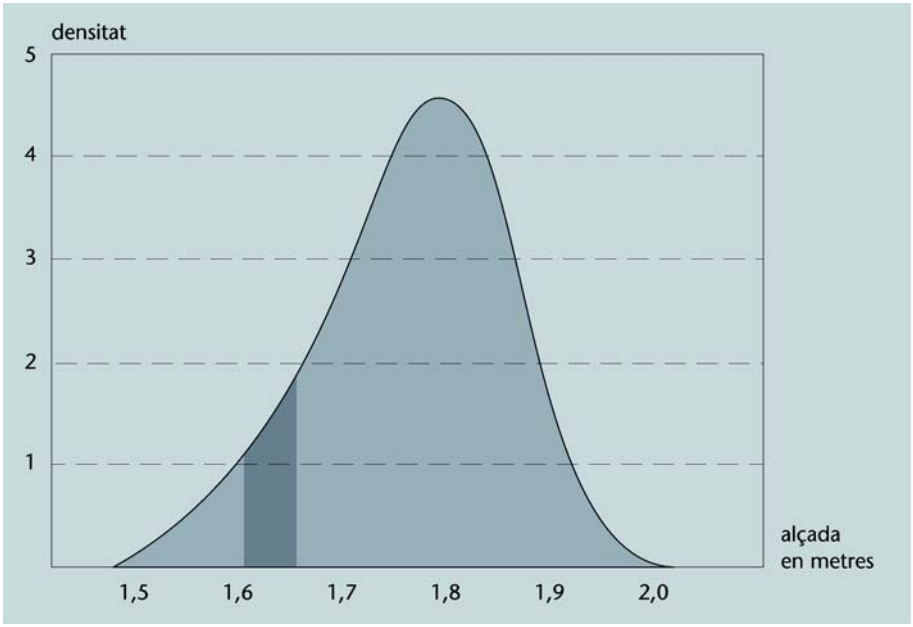
Gràfic V



La regla per a calcular proporcions d'homes en certes classes d'alçada seria la mateixa que abans, calcular l'àrea de la densitat que correspon a les classes.

Ara imagineu-vos que tinguéssim moltíssimes observacions; per exemple, les alçades de tots els homes de Catalunya, i suposem que haguéssim pres les alçades al mil·límetre. Aleshores la densitat de freqüència esdevindria encara més suau, ja que definim longituds de classe més estretes, i podria resultar aproximadament així:

Gràfic VI



Anomenem aquest tipus de corba tan suau **corba de densitat de probabilitat**. La seva forma suau és una corba teòrica que resumeix les proporcions (o probabilitats) tal com són en una certa població (per exemple, tots els homes de Catalunya). Des d'una corba així podem avaluar la proporció (o probabilitat) de qualsevol interval d'alçades calculant l'àrea sota la corba per a aquest interval. Per exemple, l'àrea ombrejada en el gràfic VI mostra la proporció relativa d'homes amb una alçada entre 1,60 i 1,65 m. Com que l'amplada de l'interval és de 0,05 i l'àrea formada té una alçada d'aproximadament 1,4, l'àrea és aproximadament $1,4 \cdot 0,05 = 0,07$; és a dir, més o menys el 7% de la població té una alçada entre 1,60 i 1,65 m.

Penseu-hi

A l'exemple de les alçades, té sentit demanar quina proporció de la població mesura exactament 1,65 m?

La resposta hauria de ser que no es pot saber, i hauríem de canviar la pregunta per: quina proporció de la població té una alçada arrodonida a 1,65 m?

La densitat normal

Fixeu-vos que la densitat de les alçades d'home no cal que resulti tan simètrica com hem mostrat en les figures. Ara bé, hi ha moltes situacions en estadística en què considerem corbes de densitat que semblen aproximadament simètriques i amb forma de campana com les figures anteriors.

Hi ha una densitat de probabilitat teòrica que és la més útil que tenim a l'abast: s'anomena **densitat normal**. En els apartats de més endavant mostrarem fins a quin punt la distribució normal és corrent i per què és tan important. De moment mirem-ne les propietats i com podríem relacionar dades que tenen histogrames aproximadament d'aquest perfil amb la densitat normal.

Penseu-hi

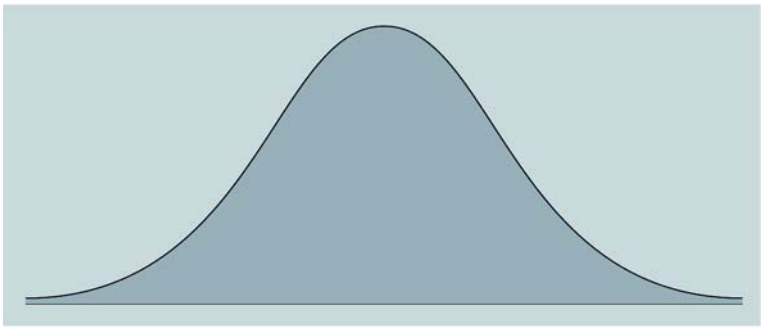
Què és el que determina la forma de la campana d'una distribució normal?

Els errors que es produeixen en mesurar moltes vegades una mateixa magnitud segueixen una distribució normal. Gauss ho va estudiar i va trobar la fórmula matemàtica que descriu aquesta distribució.

De Moivre va obtenir aquest tipus de corba abans que Gauss, a partir d'estudis sobre alguns jocs d'atzar, però no en va donar la fórmula matemàtica.

Propietats de la densitat normal

Les corbes normals tenen la forma característica i simètrica de campana:



Nota

Quan parlem de densitat normal, volem dir específicament la corba de densitat normal. Les corbes de densitat normal tenen el mateix perfil general, l'única diferència entre si és que es poden centrar en diverses posicions (d'acord amb la seva mitjana aritmètica) i poden tenir diverses dispersions (d'acord amb la seva desviació estàndard).

Com que la corba és simètrica, la mitjana aritmètica i la mediana són les mateixes, just en el centre de la corba. La densitat normal és un concepte teòric, i usem símbols grecs especials per a representar-ne la mitjana aritmètica i la desviació estàndard:

- Es representaria la mitjana aritmètica d'un conjunt d'observacions amb $\bar{x}$, però per a una corba de densitat normal ideal nosaltres representem la seva mitjana aritmètica teòrica, situada just en el punt mitjà de la corba, amb la lletra grega μ (mi).

- Es representaria la desviació estàndard d'un conjunt d'observacions amb *s*, però per a la densitat normal nosaltres representem la seva desviació amb la lletra grega σ (sigma).

Noteu...

... que sovint usem el terme *distribució normal* en aquest context i diem que les dades estan distribuïdes normalment.

Una propietat molt atractiva de la densitat normal és que la corba es descriu totalment amb la seva mitjana aritmètica μ i amb la desviació estàndard σ. Anomenem μ i σ els paràmetres de la distribució normal.

Definició d'una densitat normal

La definició matemàtica d'una densitat normal és la següent:

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

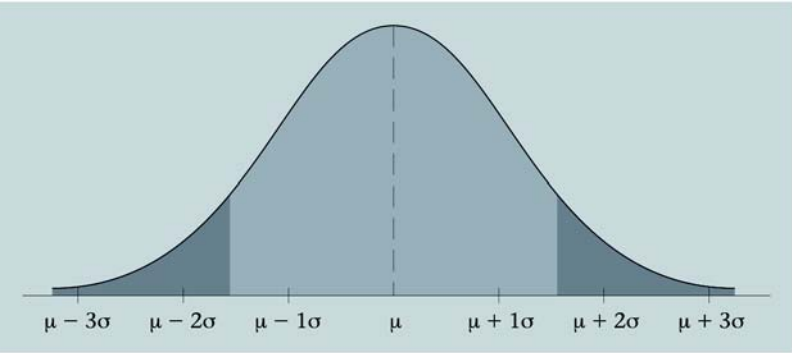
Densitat normal

La funció de densitat normal va ser descoberta per Karl Friedrich Gauss, un matemàtic autodidacte que va viure a Alemanya des de 1777 fins a 1855. La distribució normal sovint s'anomena *distribució de Gauss* en honor seu. La fórmula matemàtica que la defineix i la corba de densitat s'il·lustraven en el bitllet de deu marcs alemany.

Algunes àrees sota la corba normal

Una propietat molt útil de la densitat normal és que, en termes de desviacions estàndard, els càlculs de l'àrea sota la corba són els mateixos per a totes les densitats normals. Per exemple, per a qualsevol densitat normal, l'àrea sota la corba per a l'interval descrit per una desviació estàndard a una banda o a l'altra de la mitjana aritmètica és la mateixa: 0,68. En altres paraules, el 68% de les unitats de població tenen valors entre la mitjana aritmètica menys una desviació estàndard i la mitjana aritmètica més una desviació estàndard.

Si anem a dues desviacions estàndard a una banda o a l'altra de la mitjana aritmètica, el resultat és 0,95 –o el 95%–. I si anem tan enllà com ara tres desviacions estàndard a una banda o a l'altra de la mitjana aritmètica, aleshores hem cobert gairebé tota la població: 0,997 –o el 99,7%–; és a dir, solament el 0,3% de la població queda fora d'aquest interval. Per exemple, per a qualsevol densitat normal, si preníem 1,645 desviacions estàndard a una banda o a l'altra de la mitjana, tindríem 0,90 –o el 90%– de l'àrea coberta:



Suposeu que les alçades dels homes segueixen una corba de densitat normal amb una mitjana aritmètica d'1,69 m i una desviació estàndard de 0,15 m. Aleshores, per la regla de 68-95-99,7 podem deduir això:

- L'alçada del 68% dels homes és entre 1,54 m i 1,84 m.
- L'alçada del 95% dels homes és entre 1,39 m i 1,99 m.
- L'alçada del 99,7% dels homes és entre 1,24 m i 2,14 m.

A més, sabent que 1,645 desviacions estàndard al voltant de la mitjana aritmètica inclouen el 90% de l'àrea sota la corba, podem deduir també que:

- L'alçada del 90% dels homes és entre $1{,}69 - 1{,}645 \cdot 0{,}15 = 1{,}443$ m i $1{,}69 + 1{,}645 \cdot 0{,}15 = 1{,}937$ m.

Amb taules estadístiques (o fent servir un ordinador) podem obtenir l'àrea sota qualsevol part de la corba normal. D'això en parlarem detalladament en el següent apartat.

Penseu-hi

Com és el diagrama tramat d'una distribució normal?

Estandarditzar

El primer pas a l'hora de relacionar un valor donat amb una densitat normal és expressar el valor com un nombre de desviacions estàndard des de la mitjana aritmètica. Aquest procés s'anomena **estandarditzar** el valor.

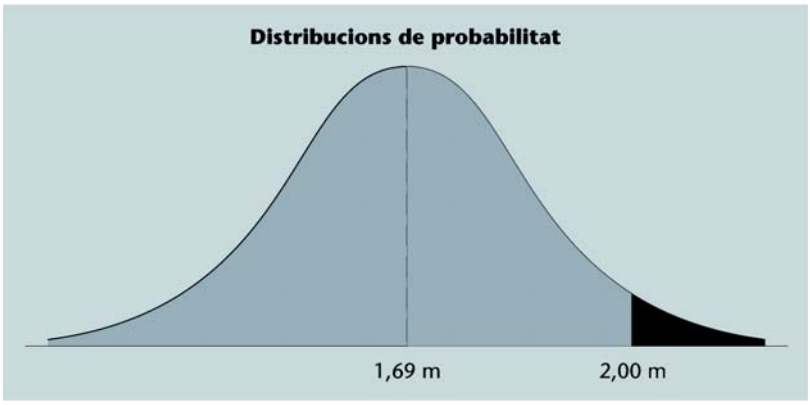
Usant el mateix exemple d'una densitat normal de les alçades dels homes amb una mitjana aritmètica d'1,69 m i una desviació estàndard de 0,15 m, imaginem-nos que volem saber la proporció d'homes amb una alçada superior a 2,00 m. Volem saber en quantes desviacions estàndard aquest valor és sobre la

mitjana aritmètica. La diferència entre 2,00 i la mitjana aritmètica 1,69 és 0,31. Com que la desviació estàndard és 0,15, aquest valor és, per tant, 0,31/0,15 = 2,067 desviacions estàndard sobre la mitjana aritmètica. El valor 2,067 és el valor estandarditzat de l'alçada 2,00 m. Normalment representem el valor estandarditzat amb la lletra *z*.

En general, quan tenim una densitat normal amb mitjana aritmètica μ i desviació estàndard σ, estandarditzem un valor x restant la mitjana aritmètica i després dividint el resultat per la desviació estàndard:

$$z = \frac{x - \mu}{\sigma}.$$

Per a obtenir la proporció concreta d'homes sobre 2,00 m, necessitem conèixer l'àrea sota la corba normal sobre el valor estandarditzat de 2,067. Tot el que sabem de moment és que, sobre dues desviacions estàndard des de la mitjana aritmètica (per exemple: sobre un valor estandarditzat d'$1,69 + 2 \cdot 0,15 = 1,99$ m), l'àrea sota la corba és del 2,5%, tal com s'il·lustra en el diagrama següent:



Per tant, sabem que la proporció d'homes serà just sota el 2,5%, però no sabem exactament quant. Aquest és el tema de l'apartat que ve a continuació.

La paradoxa dels campions de beisbol

Ara podem comprendre el raonament de Stephen Jay Gould a l'hora d'explicar per què avui no hi ha jugadors de beisbol amb mitjanes de batuda tan altes com en altres èpoques d'aquest segle. Noteu que la variable és la mitjana de batuda –és una mitjana de l'habilitat a l'hora de batre per a un jugador en particular–. Veiem que la mitjana aritmètica de mitjana de batuda s'ha mantingut més o menys igual al llarg dels anys, però la desviació estàndard ha minvat.


La mitjana de batuda és una mitjana de l'habilitat a l'hora de batre per a un jugador en particular.

En realitat, això és el resultat de la millora general de tots els jugadors, però sí que vol dir que ara hi ha menys oportunitats perquè un jugador aconsegueixi una puntuació alta. Alhora que la desviació estàndard esdevé més petita, l'àrea

sota la corba de densitat normal per a una mitjana de batuda superior a 0,400 també esdevé més petita.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Corba de densitat: funció contínua no negativa que té una àrea total d'1 sota la funció, la qual se suposa que representa la distribució teòrica d'una variable.

Freqüència relativa: proporció, o freqüència relativa al total; per exemple: si 33 homes de 215 fan entre 1,8 m i 1,9 m, aleshores la freqüència relativa és 33/215 = 0,153 –o el 15,3%.

Densitat normal: corba de densitat específica molt usada com a distribució estadística, simètrica, completament descrita pel seu centre, la mitjana aritmètica μ i la seva dispersió o la desviació estàndard σ.

Estandardització: acció d'expressar un valor x normalment classificat com el nombre de desviacions estàndard respecte de la mitjana aritmètica; per exemple: si una densitat normal té la mitjana aritmètica 245 i la desviació estàndard 61, aleshores el valor 150 té un valor estandarditzat de $(150 - 245) / 61 = -1{,}56$; en altres paraules, el valor 150 és 1,56 desviacions estàndard sota la mitjana aritmètica.

10. La distribució normal (II): càlculs normals i taules

Recordeu

Per a estandarditzar un valor: resteu-hi la mitjana i dividiu el resultat per la desviació estàndard.

Continuem l'estudi de la corba de densitat normal. En aquest apartat calculem àrees sota qualsevol part de la corba de densitat, entre qualssevol valors estandarditzats. Això ens proporcionarà estimacions de la freqüència relativa, o probabilitat, d'un conjunt determinat de valors de la població. Usem una corba de densitat normal particular, coneguda com **densitat normal estàndard**, la qual té mitjana aritmètica 0 i variància 1.

Aquesta corba es defineix en termes d'una variable que està expressada en unitats de desviacions estàndard de la mitjana aritmètica, de manera que es poden trametre els valors estandarditzats directament a aquesta corba de densitat. Les taules que donen l'àrea de la densitat normal s'usen fins a un valor en particular per a obtenir les estimacions de probabilitat que necessita.

En aquest apartat sobre la distribució normal aprendreu:

- què és una distribució normal estandarditzada;
- com es fan servir les taules de la distribució normal estandarditzada per a buscar àrees (probabilitats) sota la corba normal, entre dos valors;
- com es fan servir les taules per a buscar valors entre els quals una àrea específica queda sota la corba normal;
- com es busquen àrees sota la corba de densitat normal estàndard fent servir MacAnova.

Presumpció de normalitat

Aquí hauríem d'emfasitzar que l'ús de la distribució normal com a ideal teòric per a una població d'unitats, com ara les alçades de les noies, les emissions de NOX dels motors dels cotxes o les talles del cap dels soldats, és una presumpció i en la majoria de casos una aproximació de la realitat. Si veiem que l'histograma d'alguns valors reals és més o menys simètric i aproximadament acampanat, aleshores la presumpció de normalitat és raonable. Però no haurí-em d'usar la distribució normal altrament.

En apartats posteriors tractarem de les situacions en què es justifica la distribució normal per altres raons teòriques. De moment simplement suposem que és vàlid de familiaritzar-se amb les propietats de la distribució i amb les taules normals estàndard.

Corba de la densitat normal estàndard

Com que totes les corbes de densitat normal tenen les mateixes propietats i difereixen solament en el seu centre (mitjana aritmètica) i la dispersió (desviació estàndard), nosaltres ens centrem sols en una d'aquestes, la de la densitat normal estàndard.

La **corba de la densitat normal estàndard** és la distribució d'una variable normal estandarditzada *z*, la qual té la mitjana aritmètica 0 i la desviació estàndard 1.

Nosaltres normalment...

... diem mitjana aritmètica 0 i variància 1; però, com que la desviació estàndard és l'arrel quadrada de la variància, la desviació estàndard i la variància són idèntiques.

Suposem que la variable *X* té una distribució normal (penseu que *X* és una variable com ara el nivell d'emissions de NOX, en grams per milla), i suposem que la veritable mitjana aritmètica d'aquesta variable en la **població** d'unitats considerada és un cert valor que representem per μ. També suposem que la desviació estàndard de *X* en la població d'unitats és un valor que representem per σ. Per a estandarditzar aquesta variable hem vist que primer restem la mitjana aritmètica μ de la variable per a obtenir una desviació des de la mitjana aritmètica i després dividim aquesta desviació per la desviació estàndard σ. Aquesta nova variable estandarditzada, la qual representem per *Z*, s'anomena **variable normal estandarditzada**, i es representa així:

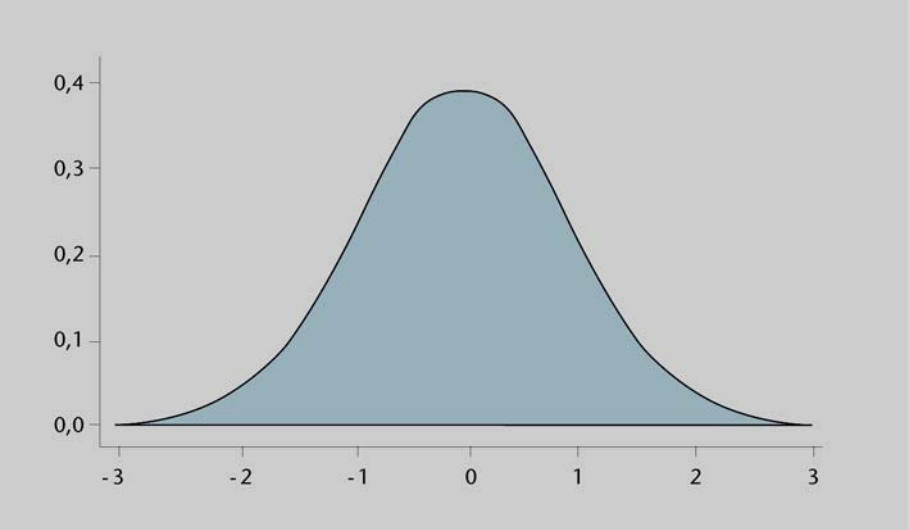
$$Z = \frac{X - \mu}{\sigma}.$$

Fixeu-vos en l'ús de lletres majúscules per a la variable en general. Podríem escriure "emissions NOX estandarditzades" per al símbol *Z* i "emissions NOX" per a la variable *X*. Quan tenim valors específics de *X*, aleshores els escrivim en minúscules, de manera que per a un valor específic *x* tenim una fórmula similar, que porta a valors estandarditzats específics:

$$z = \frac{x - \mu}{\sigma}.$$

La variable *Z* segueix una distribució normal estàndard, amb mitjana igual a 0 (ja que hi hem restat la mitjana aritmètica) i desviació estàndard 1 (ja que l'hem dividida per la desviació estàndard). Sempre estandarditzem les dades normalment distribuïdes, de manera que podrem usar simplement una corba de densitat normal, la de la densitat normal estàndard. Aquesta corba es mostra en el gràfic I.

Gràfic I



Fixeu-vos...

... en l'escala de densitat de l'esquerra i recordeu que l'àrea total sota la corba és igual a 1.

Càlcul d'àrees per a la distribució normal estàndard

Trobareu la taula de les àrees sota la corba normal estàndard a l'annex 1.

Hem subministrat les taules de l'àrea sota la corba de densitat normal estàndard. Les podem usar per a obtenir la freqüència relativa, o probabilitat, de qualsevol interval que triem.

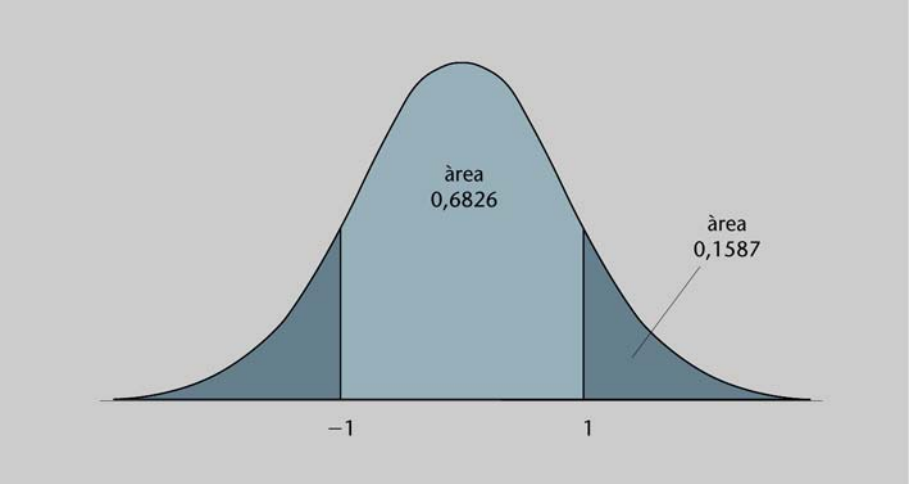
Primer permetem-nos de comprovar la regla 68-95-99,7. Fixeu-vos que les taules normals donen l'àrea sota la corba normal fins a un cert valor *z*, de manera que, a mesura que *z* augmenta, el valor dins la taula puja, començant per una probabilitat molt petita (0,0003) per al valor $-3,40$ a la part superior esquerra de la taula, fins a una probabilitat (0,9998) per al valor 3,49 a la part inferior dreta de la taula.

Vegeu el gràfic II d'aquest apartat.

Per a veure l'àrea dins una desviació estàndard de la mitjana aritmètica, hauríem de buscar el valor $z = -1,0$; i veiem en la filera etiquetada $-1,0$ i la columna 0,00 la probabilitat de 0,1587. Aquesta és la probabilitat d'un valor més petit o igual que -1. Però volem saber la probabilitat entre $-1,0$ i 1,0. Com que la densitat normal és simètrica, sabem que a la dreta de $+1,0$ hi haurà exactament la mateixa àrea de 0,1587. D'aquesta manera hi ha una àrea de $2 \cdot 0,1587 = 0,3174$ fora de l'interval que considerem. Com que l'àrea sota la corba completa és 1, simplement restem 0,3174 del valor 1 per a obtenir el nostre resultat: $1 - 0,3174 = 0,6826$.

Aquesta probabilitat de 0,6826 –o 68,3%– correspon al 68 de la regla 68-95-99,7. Les altres parts de la regla es poden verificar d'una manera similar.

Gràfic II



Usem el símbol *P* per a la probabilitat, o l'àrea sota la corba de densitat, i podem escriure el que acabem de fer de la manera següent:

$$P(Z < -1{,}0) = P(Z > 1{,}0) = 0{,}1587,$$

$$P(-1{,}0 < Z < 1{,}0) = 1 - P(Z < -1{,}0) - P(Z > 1{,}0) = \\ = 1 - 2 \cdot 0{,}1587 = 0{,}6826.$$

- La primera expressió diu que la probabilitat que Z sigui més petita que $-1{,}0$ iguala la probabilitat que la variable Z sigui més gran que 1,0; la qual en la taula és igual a 0,1587.

- La segona expressió diu que la probabilitat que Z quedi dins una desviació estàndard (que és el que nosaltres volem) és 1 menys cadascuna de les probabilitats de sobre, i el resultat és 0,6826.

Permetem-nos de mirar l'ús contrari de les taules una altra vegada. Ara especifiquem per endavant un cert percentatge, o probabilitat. Suposem que hem estandarditzat les puntuacions d'un grup d'estudiants i en volem eliminar el 10% més baix (més avall donem un exemple específic d'això). Necessitem trobar el valor z fins al qual l'àrea sota la corba és 0,10. Ara necessitem cercar en les taules un valor tan aproximat a 0,10 com sigui possible, i el valor a què ens podem acostar més és 0,1003, el qual és en la filera etiquetada $-1{,}2$ i la columna etiquetada 0,08. És a dir, el valor z de $-1{,}28$ separa 0,10 a l'esquerra, o $P(Z < -1{,}28) = 0{,}10$.

Relacionar les dades en les unitats originals amb la distribució normal estàndard

Per a relacionar dades en les unitats originals amb unitats estàndard, estandarditzem les dades. Per a tornar a relacionar dades estandarditzades amb dades en les unitats originals, podrem dir que traiem l'estandardització a les dades.

L'acte d'estandardització d'una variable és, com sabem:

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

Notació

a. Mitjana teòrica d'una distribució: μ.

b. Variància teòrica d'una distribució: σ^2.

c. Desviació estàndard teòrica d'una distribució: σ.

de manera que l'acte de treure l'estandardització és la fórmula inversa:

$$X = \mu + Z\sigma.$$

Per exemple, suposem que assumim que les puntuacions dels exàmens d'una classe d'estudiants d'estadística són normals, amb la mitjana aritmètica 6,4 i la desviació estàndard 1,2. Suposem que volem saber la proporció d'estudiants amb puntuacions superiors a 5,0. Primer estandarditzem el valor 5,0:

Si disposem de les puntuacions dels exàmens d'una classe d'estudiants d'estadística podem saber la proporció d'estudiants amb puntuacions superiors a 5,0.

$$z = \frac{5{,}0 - 6{,}4}{1{,}2} = -1{,}167.$$

Solament tenim taules amb espai per a dos decimals, de manera que busquem el valor més pròxim de $-1{,}17$ mirant la filera etiquetada $-1{,}10$ i la columna etiquetada 0,07. En la intersecció de la filera i la columna trobem el valor 0,1210. Aquesta no és l'àrea que volem; nosaltres volem l'àrea a sobre de z, de manera que restem aquest valor d'1 per a obtenir el nostre resultat 0,879. Per tant, s'estima que el 87,9% de les puntuacions són per sobre de 5,0 (i el 12,1% per sota de 5,0).

Com a exemple de l'ús invers de les taules, permetem-nos de suposar que volem identificar la puntuació per sota de la qual queden exactament el 10% de les puntuacions. Sempre hem vist que el valor z de $-1{,}28$ dóna una probabilitat de 0,1003, la qual és la més pròxima en la taula a 0,10. Necessitem treure l'estandardització d'aquest valor per a tornar a l'escala original. Primer multipliquem $-1{,}28$ per la desviació estàndard 1,2, per a obtenir el valor $-1{,}536$, el qual és la desviació de la mitjana aritmètica. Després hi sumem la mitjana aritmètica 6,4 per a obtenir la puntuació de 4,864, la qual torna a ser en les nostres unitats originals. D'aquesta manera la resposta a la pregunta és que el 10% de les puntuacions són 4,86 o menys.

Activitats

1. En un estudi previ de l'absentisme en una fàbrica hem vist que el nombre de dies en què els treballadors s'absenten segueix una distribució normal, amb una mitjana del nombre d'absències anual de 6,2 i una desviació estàndard d'1,8. Convertiu els valors següents en valors estandarditzats:

0 1 2 3 4 5 6 7 8 9 10

Després, usant les taules normals estàndard, avalueu la proporció de treballadors que estimem absents:

a) durant 1 dia o menys;
b) durant 10 dies o més;
c) entre 4 i 6 dies (incloent-los tots dos).

L'ús de `cumnor()` i `invnor()` per a calcular probabilitats normals

Es pot usar el MacAnova per a calcular les àrees exactes sota una corba de densitat normal estàndard.

La funció `cumnor()` calcula l'àrea sota la corba fins a un valor específic *z*, de la mateixa manera que els gràfics normals que hem estudiat.

El nom `cumnor`...

... és l'abreviació de *cumulative normal* (és a dir: la probabilitat acumulada fins a un cert valor sota la corba normal).

Per exemple, per al valor *z* d'1,50 (una desviació estàndard i mitja sobre la mitjana) obtenim l'àrea sota la corba normal així:

```
Cmd> cumnor (1.50)
(1)          0.93319
```

Si volem calcular l'àrea entre dos valors *z*, aleshores avaluem les probabilitats acumulades per a cada valor *z* i després restem la més petita a la més gran. Per exemple, per a avaluar l'àrea sota la corba normal entre $-0{,}5$ i 1,5 fem aquest càlcul:

```
Cmd> cumnor (1.5)-cumnor(-0.5)
(1)         0.62466
```

Aquesta és l'àrea que mostrem ombrejada en el gràfic III. (Vegeu la pàgina següent.)

Activitats

2. Verifiqueu la regla 68-95-99,7 usant `cumnor()`.

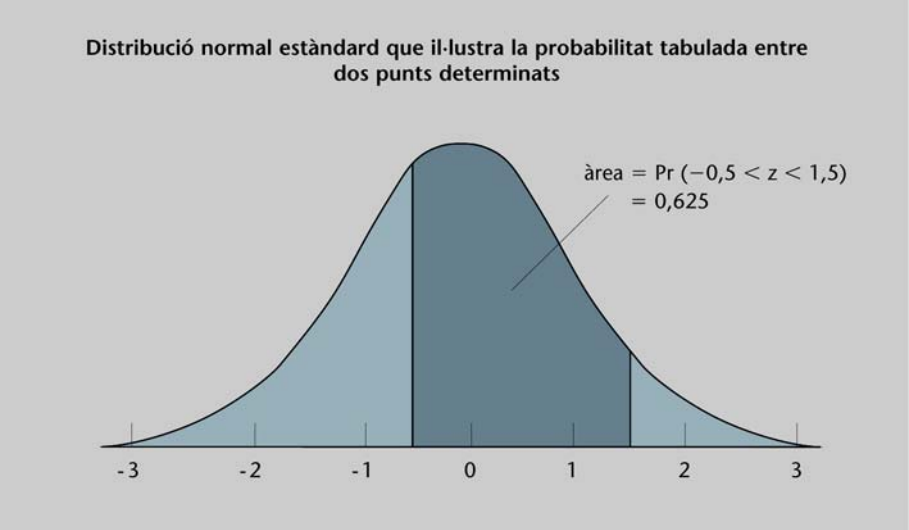
A l'hora d'usar `cumnor()` introduïm el valor *z*, i `cumnor()` proporciona l'àrea, o probabilitat sota la corba normal. Per a fer l'operació inversa haurem d'introduir una probabilitat, o àrea sota la corba, per tal de trobar quin valor *z* correspon a aquesta probabilitat. Per a fer això usem la funció del MacAnova `invnor()`.

Ordres d'ús del MacAnova

a. `cumnor(z)` genera l'àrea sota la corba normal fins a un valor `z`.

b. `invnor(p)` genera el valor normal estandarditzat fins al qual l'àrea sota la corba normal és igual a `p`.

Gràfic III



Per exemple, imaginem-nos que volem saber quins valors *z* corresponen a probabilitats acumulatives de 0,05 i 0,025 (és a dir, el 5% i el 2,5% de l'àrea sota la corba normal):

```
Cmd> invnor(0.05) ; invnor(0.025)
(1)       -1.6449
(1)       -1.96
```

Activitats

3. Useu `invnor()` per a trobar quins valors *z* tallen al 10%, 1% i 0,1% de l'àrea sota la corba normal.

Els **conceptes principals** que hem vist en aquest apartat són:

Distribució normal estàndard: distribució d'una variable normal que ha estat estandarditzada, és a dir, una distribució normal amb mitjana aritmètica 0 i variància 1.

Estandarditzar una observació *x*: $z = \frac{x - \mu}{\sigma}$ (*z* rep el nom de *valor estandarditzat*).

Distribució normal: `cumnor()`, `invnor()`.

11. La distribució normal (III): distribució mostral de la mitjana aritmètica

Fins ara hem vist la distribució d'un conjunt de dades per a una variable en particular. En aquest apartat veurem més de prop la distribució de la mitjana aritmètica d'una variable, més que no pas la distribució de les dades. Veurem que la mitjana aritmètica d'un conjunt d'observacions normalment distribuïdes també és distribuïda d'una manera normal, però amb una desviació estàndard més petita. També trobarem un dels famosos teoremes de l'estadística, el teorema del límit central. Aquest teorema diu que fins i tot quan les dades no són normalment distribuïdes, la mitjana aritmètica calculada sobre una mostra aleatòria d'aquestes dades tendeix a ser normalment distribuïda.

En aquest apartat sobre distribucions mostrals aprendreu:

- que la mitjana aritmètica d'un conjunt de variables normalment distribuïdes també té una distribució normal;
- la manera com la desviació estàndard de la mitjana aritmètica està en relació amb la de les observacions originals;
- que per a mostres àmplies la mitjana aritmètica d'una mostra aleatòria de les observacions sobre qualsevol variable no necessàriament distribuïda normalment és d'una manera aproximada normalment distribuïda (teorema del límit central).

Prendre mostres repetides d'una població

Quan tenim un conjunt de valors de dades, x_1, x_2, ..., x_n, mostrejats d'una manera aleatòria dins una població, un dels estadístics més importants és la mitjana aritmètica $\bar{x}$. La mitjana aritmètica resumeix el centre de la distribució. En la pràctica sols tenim una mostra simple i solament calculem una mitjana aritmètica d'aquesta mostra. Però potencialment hi ha moltes mostres de la població que podríem haver pres, i cadascuna d'aquestes mostres té una mitjana aritmètica diferent. Això suggereix que la mitjana aritmètica té una distribució, i nosaltres l'anomenem **distribució mostral de la mitjana aritmètica**.

Penseu-hi

Les distribucions mostrals es poden calcular de les mitjanes, de les variàncies, de les desviacions, de les medianes...

Un conjunt de dades distribuïdes normalment

Poseu en marxa el vostre ordinador i executeu el programa MacAnova. Intenteu completar aquest apartat en una sola sessió.

Per a il·lustrar aquestes idees cal que tinguem una població de la qual puguem prendre repetides mostres.

Hem subministrat un arxiu anomenat `NORMAL` que conté 1.000 valors d'una distribució normal, amb una mitjana aritmètica 100 i una desviació estàndard d'aproximadament 10. Llegiu aquestes dades de la manera usual:

```
Cmd> pop <- vecread ("NORMAL")
```

Nota

Hem generat aquestes dades usant la funció de MacAnova `rnorm()`, la qual genera observacions normals aleatòries a partir d'una població normal amb una mitjana aritmètica 0 i una variància 1. Per a obtenir 1.000 valors de la distribució, l'ordre seria `rnorm(1000)`; però volíem valors amb una desviació estàndard 10 i una mitjana aritmètica 100, per tant usem els valors transformats `rnorm(1000)* 10 + 100`. Aquest va ser el nostre argument concret, el qual mostrava les bases de nombre aleatori que es van fer servir automàticament: `Cmd> rnorm(1000)* 10 + 100.`

I després introduïu el nom del vector de dades `pop` per a veure els valors. Comproveu el mínim, el màxim, la mitjana aritmètica i la desviació estàndard:

```
Cmd> min(pop) ; max(pop)
(1)        70.321
(1)        141.45

Cmd> sum(pop)/1000
(1)           100

Cmd> sqrt(sum((pop-100)^2)/999)
(1)         9.9447
```

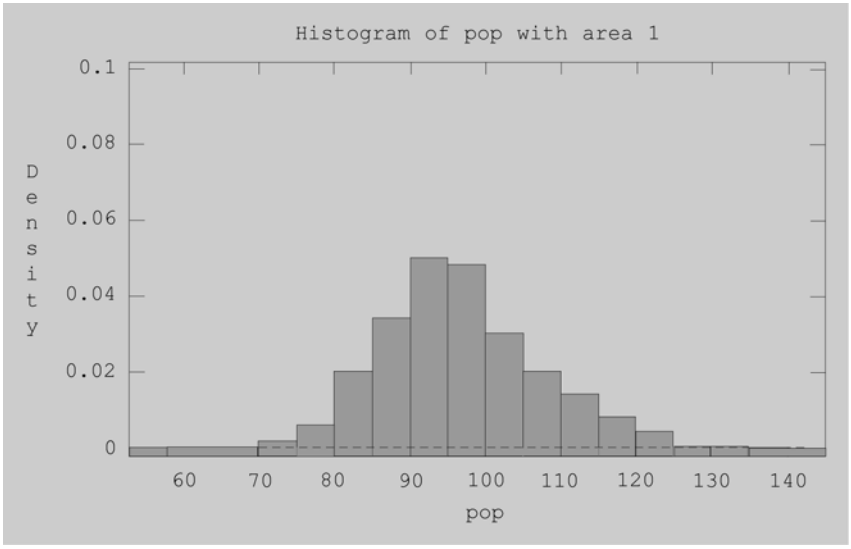
Repetiu aquests càlculs pel vostre compte.

També podem veure la distribució de la població amb la funció `hist()` del MacAnova, i alhora ensenyar-vos com es controla l'amplada i l'alçària de les barres dels histogrames:

Atenció amb els parèntesis

Quan repetiu els càlculs vigileu tots els parèntesis de l'última argumentació, hi ha un conjunt de parèntesis per a la funció `sqrt()`, després n'hi ha un parell a l'interior per a la funció `sum()` i, finalment, un altre parell a l'interior d'aquests darrers per a la desviació `pop-100`.

```
Cmd> bars<-run(55,145,5)

Cmd> bars
(1)                55    60    65    70    75
(6)                80    85    90    95   100
(11)              105   110   115   120   125
(16)              130   135   140   145

Cmd> hist(pop,bars,ymin:0,ymax:0.1)
```



La funció `run(a,b,c)` genera un vector de valors que comença en `a`, acaba en `b` i s'incrementa de `c` en `c`. Per tant, `run(55, 145, 5)` genera un vector de valors 55, 65, ..., 140, 145, els quals guardem en el vector `bars`. Després l'ordre `hist(pop,bars)` ens dóna un histograma en què les barres van del 55 al 60, del 60 al 65 i així fins a l'interval del 140 al 145.

Penseu-hi

Què obteniu si feu servir la funció `run(55, 145, 2)`?
Com és ara l'histograma?

Aquesta és una característica útil, perquè volem dibuixar tots els histogrames per a aquestes dades en la mateixa escala horitzontal i la mateixa amplada de barres. Per a controlar l'alçària de les barres podem especificar un valor mínim i un valor màxim per a l'eix de les ordenades (*y*), fent servir les paraules clau `ymin` i `ymax`. En aquest cas volem imprimir l'histograma actual i els dos que seguiran, tots en la mateixa escala vertical, i l'amplitud de 0,0 a 0,1 és suficientment ampla per a acomodar-los tots tres.

La il·lustració de la distribució mostral

Considerem que el nostre conjunt de 1.000 valors normalment distribuïts són la nostra població, amb una mitjana aritmètica μ igual a 100 i una desviació estàndard σ gairebé exactament igual a 10. Ara fem veure que no tenim aquesta població sencera, sinó que necessitem fer-ne un mostreig amb el propòsit de deduir-ne les característiques. En particular volem saber el comportament de la mitjana aritmètica en les mostres aleatòries extretes de la població.

Mitjançant una enquesta podem obtenir dades d'una mostra aleatòria de la població.

Començarem mirant les mostres aleatòries de mida 4. Ja sabem com es pren aquest tipus de mostra. N'obtindríem la llista dels índexs amb l'ús de l'expressió `ind <- ceiling(runi(4) * 1000)` i després els valors amb l'ús de `pop[ind]`. Preneu, doncs, una mostra d'aquest tipus, emmagatzemeu el resultat en un vector anomenat `sample` ('mostra') i després trobeu la mitjana aritmètica de `sample`.

A l'apartat 7 hem vist com s'usa el MacAnova per a triar una mostra.

```
Cmd> ind <- ceiling(runi(4) * 1000)

Cmd> sample <- pop[ind]

Cmd> sample
(1)92.389    94.233    109.36    108.28

Cmd> sum(sample)/4
(1)        101.07
```

Quan repetiu aquestes expressions, no obtindreu el mateix resultat, ja que els nombres aleatoris generats per `runi()` seran diferents i mostrejareu un conjunt diferent de quatre valors procedents dels 1.000 valors dins `pop`.

Resum del que es fa amb l'ordinador

a. Tenim una població N (μ, σ).

b. Prenem mostres probabilístiques de mida n.

c. Calculem les mitjanes x de les mostres.

d. Fem la distribució mostral de les mitjanes.

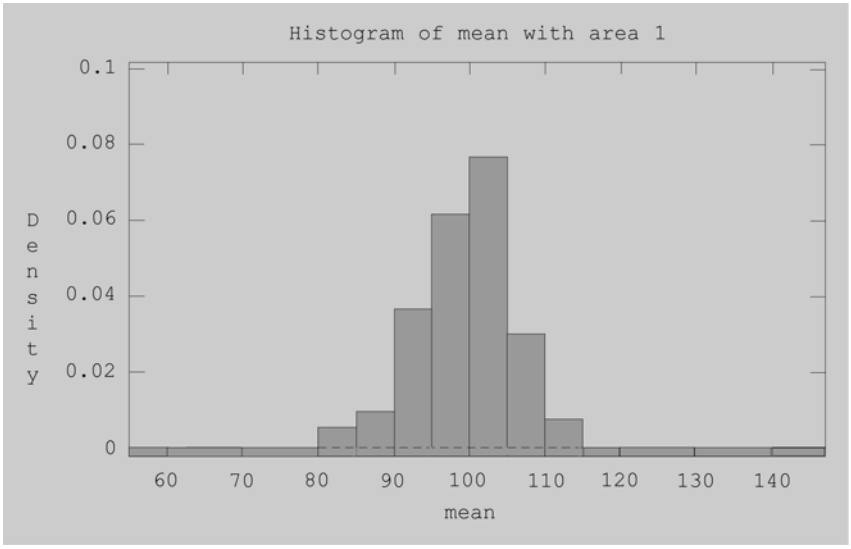
e. Comparem amb la mitjana i la desviació poblacional.

Aquest és exactament el punt que provem d'il·lustrar: la mitjana aritmètica cada vegada serà diferent. Però, quines diferències hi haurà entre si?

Ho podem estudiar generant moltes mitjanes aritmètiques diferents basades en mostres aleatòries de mida 4 procedents de la població. Ja hem repetit aquest exercici 400 vegades i hem emmagatzemat els resultats en un fitxer anomenat `MEANS4` (és a dir, mitjana aritmètica de mida 4). D'aquesta manera podem veure aquestes mitjanes aritmètiques, la seva distribució, i la seva mitjana aritmètica i les desviacions estàndard de la manera següent:

***Mean* en anglès significa 'mitjana aritmètica'.**

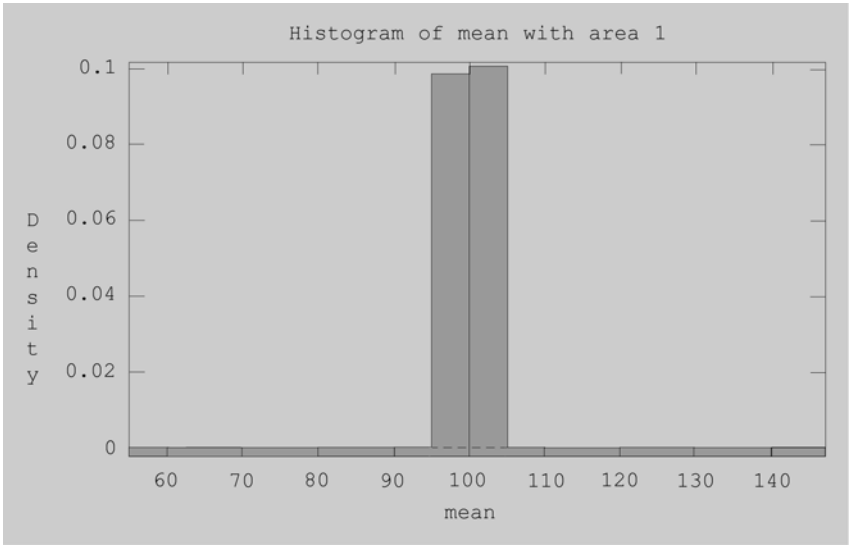
```
Cmd> mean <- vecread("MEANS4")

Cmd> sum(mean)/400
(1)        100.15

Cmd> variance<-sum((mean-100.15)^2)/399

Cmd> sqrt(variance)
(1)        4.8137

Cmd> hist(mean,bars,ymin:0,ymax:0.1)
```



Recordeu que el que estudiem aquí és un conjunt de mitjanes aritmètiques, com si aquestes mitjanes fossin les dades. Primer observeu que el valor mitjà de 400 mitjanes aritmètiques és 100,15, el qual és a prop de la veritable mitjana aritmètica 100 de la població. Aleshores la desviació estàndard de les 400 mitjanes aritmètiques és 4,8137, la qual és aproximadament la meitat de la desviació estàndard de la població.

Mirem, ara, mostres grans d'una població prenent-ne una sèrie de grandària 64. Com abans, ja hem preparat el fitxer, aquest cop anomenat `MEANS64`, el qual conté les respectives mitjanes de 400 mostres aleatòries de mida 64, preses de la població de 1.000 valors emmagatzemats en el fitxer `NORMAL`. Observeu aquestes "dades" de la mateixa manera que abans:

```
Cmd> mean <- vecread("MEANS64")

Cmd> sum(mean)/400
(1)        100.07

Cmd> variance<-sum((mean-100.07)^2)/399

Cmd> sqrt(variance)
(1)        1.1789

Cmd> hist(mean,bars,ymin:0,ymax:0.1)
```



Ara el valor mitjà de les mitjanes aritmètiques encara s'acosta més a la veritable mitjana aritmètica, i veiem que la desviació estàndard ha baixat a 1,1789, al voltant d'un quart de la desviació estàndard de les mitjanes aritmètiques de les mostres de grandària 4.

La desviació estàndard de la mitjana aritmètica

Hem il·lustrat un resultat ben conegut en estadística, és a dir, la mitjana aritmètica varia cada cop menys a mesura que la mida de la mostra augmenta. Si σ indica la desviació estàndard de la població, i si $\sigma_{\bar{x}}$ indica la desviació estàndard de la mitjana aritmètica basada en una mostra de grandària *n*, aleshores la relació exacta entre $\sigma_{\bar{x}}$ i σ és la següent:

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

Per tant, la desviació estàndard minva en proporció inversa a l'arrel quadrada de la mida de la mostra. Per això les mitjanes aritmètiques de les mostres de mida 4 tenien unes desviacions estàndard aproximades d'una meitat de la desviació estàndard de la població, de 10, i per això quan multiplicàvem la grandària de la mostra per 16 (de 4 a 64) la desviació estàndard de la mitjana es dividia per quatre, és a dir, la mitjana aritmètica de les mostres de mida 64 té una desviació estàndard aproximada d'1/8 de la desviació estàndard de la població.

Hi ha un terme especial per a designar la desviació estàndard de la mitjana aritmètica: l'**error estàndard**. El resultat que hem obtingut, doncs, ens mostra que l'error estàndard de la mitjana aritmètica és la desviació estàndard de la població dividida per l'arrel quadrada de la grandària de la mostra.

El teorema central del límit

Tot el que hem fet fins ara ha estat per a una població normal, és a dir, quan les observacions són normalment distribuïdes.

El teorema central del límit diu que, fins i tot si la distribució d'una observació no és normal, la distribució de la mitjana basada en una mostra de mida *n* serà aproximadament normal, també amb l'error estàndard igual a la desviació estàndard de població d'una observació dividida per l'arrel quadrada de *n*.

Aquest teorema esdevé més i més cert a mesura que *n* augmenta; en altres paraules, per a una *n* "petita" (per exemple menys de 10), la distribució de la mitjana aritmètica només és aproximadament normal, mentre que per a una *n* "gran" (per exemple de 100), la distribució és gairebé normal.

Deixarem que comproveu vosaltres mateixos el teorema central del límit en l'activitat següent.

Activitats

1. Assageu això que ve a continuació usant el MacAnova:

```
Cmd> normal<-rnorm(400)
Cmd> lognormal <- exp(normal)
Cmd> hist (lognormal)
```

La primera argumentació posa 400 observacions normals estàndard dins el vector `normal`. La segona, aplica la funció exponencial a aquestes dades i les posa dins el vector `lognormal`. La tercera, fa un histograma d'aquestes observacions transformades, i comprovareu que són molt asimètriques cap a la dreta. Aquestes dades s'anomenen *dades lognormal* perquè els seus logaritmes són normalment distribuïts.

Us hem proporcionat algunes ordres fetes del MacAnova per a mostrejar amb èxit una distribució lognormal, i després calcular la mitjana aritmètica de cada mostra i emmagatzemar-les en un vector anomenat `mean` ('mitjana aritmètica'). Si voleu veure un exemple de les ordres del MacAnova, feu un cop d'ull al contingut de l'arxiu `CLT10.MAC` fent servir un editor o simplement teclegeu `TYPE CLT10.MAC` des del DOS. Per a executar el contingut d'aquest fitxer dins el MacAnova introduïu l'ordre següent:

Nota

En un Macintosh hauríeu d'introduir l'ordre `batch()`, i després fer clic sobre el nom de l'arxiu apropiat.

```
Cmd> batch("CLT10.MAC")
```

Les ordres que l'arxiu conté surten en la pantalla, el programa triga uns quants segons a executar-les i després acaba. La màquina ha agafat 400 mostres aleatòries de grandària 10 d'una distribució lognormal, n'ha calculat el valor mitjà i després ha emmagatzemat els 400 valors mitjans en el vector `mean`. Per a veure la distribució d'aquestes mitjanes aritmètiques teclegeu l'ordre:

```
Cmd> hist(mean)
```

Veureu que la distribució de les mitjanes aritmètiques encara és asimètrica, però no tan asimètrica com les dades lognormal originals.

- També podeu calcular la desviació estàndard d'aquestes mitjanes aritmètiques i anotar-la.

Ara repetirem tot l'exercici amb un fitxer anomenat `CLT50.MAC`. Aquest fitxer fa el mateix, però per a les mostres de mida 50. Teclegeu l'ordre:

```
Cmd> batch("CLT50.MAC")
```

i després torneu a fer l'histograma usant `hist(mean)`. Ara veureu una distribució de 400 mitjanes aritmètiques, totes basades en una mostra de mida 50, i el patró de la distribució és definitivament més simètric.

- Calculeu també la desviació estàndard d'aquest nou conjunt de mitjanes aritmètiques i compareu-la amb la desviació estàndard calculada més amunt: quina és la relació aproximada entre totes dues?

Això il·lustra el teorema central del límit. Si preníeu mostres de grandària 100, aleshores la distribució de les mitjanes aritmètiques encara seria més simètrica, i per a mostres més i més grans les mitjanes aritmètiques esdevenen normalment distribuïdes.

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Mitjanes aritmètiques: les mitjanes aritmètiques calculades sobre mostres normalment distribuïdes de mida n també són normalment distribuïdes, però amb una desviació estàndard igual a la desviació estàndard de la població dividida per $\sqrt{n}$.

Teorema central del límit: mitjanes aritmètiques sobre mostres a partir de distribucions que no són normals esdevenen normalment distribuïdes, a mesura que augmenta la mida de la mostra; la desviació estàndard de tals mitjanes també decreix per $\frac{1}{\sqrt{n}}$ a mesura que augmenta la grandària n de la mostra.

Distribució lognormal: variable amb un logaritme normalment distribuït; per tant, podem construir dades lognormal aplicant la funció exponencial de les dades normalment distribuïdes.

Funció `batch("nomdelfitxer")` del MacAnova: llegeix d'un arxiu el codi MacAnova i l'executa.

12. Introducció a les dades categòriques: la distribució d'una proporció

En l'apartat 1 diferenciàvem entre dades numèriques i dades categòriques. En aquest apartat fem un cop d'ull al tipus més simple de variables categòriques, aquella que no té sinó dos valors possibles. Tot sovint es troben aquests tipus de variables en la pràctica, per exemple el sexe amb les possibles categories mascle o femella, o els resultats dels exàmens amb les categories aprovat o suspès. Quan mirem les mitjanes d'aquestes dades, això ens portarà a considerar la distribució d'una proporció o d'un percentatge.

En aquest apartat introductori sobre les dades categòriques aprendreu:

- la forma més simple de variable categòrica, la variable binària;
- què són la mitjana i la desviació estàndard d'una variable binària;
- que la mitjana mostral de les dades binàries és equivalent a un percentatge;
- com s'utilitza el teorema central del límit per a aproximar-se al percentatge.

Les variables binàries

Les **variables categòriques** són aquelles que tenen únicament uns quants valors possibles.

Per exemple, en una enquesta feta a la clientela d'un banc es pregunta: què us sembla el servei que rebeu en el taulell d'informació –molt satisfactori, satisfactori, no ho sabeu, insatisfactori o molt insatisfactori–? Aquesta és una variable categòrica amb cinc categories. Una variable categòrica més simple és la que té únicament dues categories, per exemple: satisfactori o insatisfactori. La manera com es formulen tals preguntes depèn molt dels objectius de l'enquesta i el grau de detall exigit. Altres dades són inherentment categòriques, com ara el sexe (home o dona) o les comunitats autònomes espanyoles (Catalunya, Andalusia, etc.).

En aquest apartat veiem el tipus més simple de variable categòrica, la qual únicament té dues categories. L'anomenem **variable binària** o, com se la sol anomenar sovint en els llibres de text d'estadística, **variable Bernoulli**.

És convenient codificar les dues categories d'una variable binària amb els valors 0 i 1. Tant és el valor que cada categoria rebi, però sovint una de les dues categories és més el centre d'atenció que l'altra, en aquest cas donem el codi 1 a aquesta categoria i el 0 a l'altra.

Jakob Bernoulli (1654-1705)

Teòleg, matemàtic i astrònom. A la seva obra *Ars conjectandi* va enunciar l'anomenat *teorema de Bernoulli* sobre el càlcul de les probabilitats.

Exemples

Abans de les eleccions generals espanyoles de 1996 una gran part de l'atenció se centrava en si l'electorat votaria pel Partit Popular (PP) o no. Podem pensar una variable observable de tot l'electorat espanyol amb les categories "vota PP" o "no vota PP". En una enquesta feta sobre aquesta qüestió seria natural assignar el valor 1 a la primera categoria i 0 a la segona.

Un altre exemple d'una variable binària és en el control de qualitat en què un inspector o una inspectora aprova o rebutja un producte com a satisfactori o defectuós respectivament. Aquí la categoria "satisfactori" seria codificada amb l'1, i "defectuós", amb el 0.

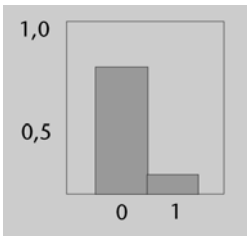
La mitjana i la desviació estàndard d'una variable binària

Com a exemple d'una variable binària considerem ara un joc en què teniu la probabilitat de guanyar d'1/6. Per exemple, tireu un dau i, si traieu per exemple un sis, aleshores guanyeu; si no, perdeu. Cada vegada que es tira el dau s'anomena *assaig* i és una observació, la qual té el valor o bé 1 per a una victòria o bé 0 per a una derrota.

Com a exemple d'una variable binària podem considerar un joc en què teniu una probabilitat de guanyar d'1/6.

Quina és la població en aquest cas? En altres paraules, quina és la distribució teòrica de tots els assajos possibles? La distribució pren una forma molt simple. Primer, els valors possibles de la variable són únicament 0 i 1, i l'àrea, o probabilitat, sota la corba de densitat del valor 0 és 5/6, i l'àrea de l'1 és 1/6.

Gràfic I



Quina és la mitjana d'aquesta distribució? Gràcies als codis 0 i 1 que hem donat a les dues categories, la mitjana és simplement la probabilitat 1/6 de guanyar. La mitjana aritmètica d'una població és la suma de tots els assajos possibles dividits pel total. Com que les dades són o bé 0 o bé 1, la suma és igual al nombre d'uns que hi hagi, i, si dividim pel total, el resultat és el nombre d'uns com una proporció del total, la qual en aquest cas és simplement 1/6.

Recordeu

La mitjana és la suma de totes les dades observades en una població dividida pel nombre total de dades:

$$\mu = \frac{\sum x_i}{n}.$$

Ara introduïm alguna notació estàndard perquè ens ajudi a expressar les nostres idees d'una manera més formal. Generalment s'indica la proporció de la població, per exemple 1/6 de l'exemple del joc del dau, amb la lletra grega π. De moment ens interessa la població i acabem de veure que la mitjana poblacional μ d'una variable binària és la mateixa π:

$$\mu = \pi.$$

Què passa amb la variància d'aquesta variable? Per a calcular la variància, primer hem de restar cada valor en la població, sigui un 0 o un 1, de la mitjana poblacional. Primer mirem-nos el nostre exemple en què la mitjana és 1/6. La variància en la població és el valor mitjà de les diferències al quadrat entre tots els assajos possibles i la mitjana. Recordeu que tenim una infinitat d'assajos dins la població, però aleshores 5/6 tenen el valor 0, i la diferència al quadrat entre 0 i la mitjana és:

$$\left(0-\frac{1}{6}\right)^2=\frac{1}{36}.$$

L'altre 1/6 dels nostres assajos té el valor 1, que correspon a una diferència al quadrat amb la mitjana de:

$$\left(1-\frac{1}{6}\right)^2=\frac{25}{36}.$$

Per tant, la variància de la població és:

$$\frac{5}{6}\cdot\frac{1}{36}+\frac{1}{6}\cdot\frac{25}{36}=\frac{1}{6}\cdot\frac{5}{6}.$$

D'aquesta manera la variància d'una variable binària amb una mitjana 1/6 és igual a:

$$\frac{1}{6}\left(1-\frac{1}{6}\right).$$

Aquest resultat és cert per a la mitjana general π, la variància d'una variable binària amb una mitjana π és $\pi\ (1-\pi)$ i la desviació estàndard és, doncs:

$$\sqrt{\pi(1-\pi)}\,.$$

No és fàcil pensar què significa realment la desviació estàndard d'una variable binària que pren únicament els valors 0 i 1. Si $\pi = 1/6$ aleshores la desviació estàndard és:

$$\sqrt{\frac{5}{36}}=0{,}373.$$

No podem interpretar això com la desviació estàndard d'una distribució normal, ja que seria absurd dir, per exemple, que el 95% de totes les desviacions es troben enmig de dues desviacions estàndard de la mitjana. La distribució del gràfic I no és ni de bon tros una distribució normal.

Ara bé, el coneixement de la desviació estàndard de la població ens ajuda a obtenir un error estàndard respecte de la mitjana de les observacions sobre una variable binària, i aquí podem fer servir la distribució normal gràcies al teorema central del límit.

Recordeu

La variància poblacional és la mitjana de les desviacions quadràtiques de les dades respecte de la mitjana:

$$\sigma^2=\frac{\sum(x_i-\mu)^2}{n}.$$

Activitats

1. Suposem que tireu una moneda enlaire i observeu si surt cara o creu. Assignem l'1 a la cara i el 0 a la creu. Quina és la distribució de totes les tirades? Quina és la mitjana aritmètica i la desviació estàndard d'aquesta distribució?

La distribució mostral de la mitjana aritmètica d'una variable binària

Ara tornem al nostre joc dels daus, en el qual la probabilitat que tenim de guanyar és d'1/6 i la de perdre, de 5/6. Tirar els daus un sol cop és prendre una observació d'una variable binària amb una mitjana 1/6. Ara els tirarem diverses vegades, és a dir que prendrem una mostra de valors a partir de la distribució. Suposem que jugàvem 100 vegades i obteníem 19 victòries i 81 derrotes. Una sisena part de 100 és 16,67; en altres paraules, de 100 tirades esperaríem aconseguir 16 o 17 victòries, però és clar que hi haurà alguna variació al voltant d'aquest valor.

Vegeu el gràfic I d'aquest apartat.

Aquesta és exactament la mateixa situació que teníem abans quan mostrejàvem a partir d'una distribució normal i estudiàvem la distribució mostral de la mitjana aritmètica. L'única diferència és que aquí mostregem a partir d'una distribució molt simple, i la distribució, sens dubte, no és normal.

La mitjana dels nostres 100 assajos o repeticions del joc és la suma de les nostres observacions dividida per 100. Les observacions consisteixen en 19 uns i 81 zeros, de manera que la mitjana és igual a 19/100 = 0,19, cosa que és la proporció observada de victòries. Si tiràvem 100 cops més, obtindríem un altre valor per a la mitjana mostral, per exemple 0,15. Després, si hi tornàvem, potser obtindríem una proporció de 0,17 victòries, i així successivament. Gradualment construïm la distribució mostral de la mitjana.

Quina suposem que serà la mitjana d'aquesta distribució mostral? Serà la mitjana poblacional: 0,167.

I la desviació estàndard? Serà la desviació estàndard de la població dividida per l'arrel quadrada de la mida de la mostra, que en aquest cas serà 0,373 dividit per 10, és a dir: 0,0373.

I el perfil de la distribució? Gràcies al teorema central del límit tenim una mida de la mostra gran, 100; podem dir amb seguretat que la distribució de la mitjana és aproximadament normal.

Hem arribat a un resultat important que podem generalitzar. Suposem que tenim un experiment aleatori amb la probabilitat d'èxit igual a π. Duem a terme un nombre d'assajos independents d'aquest experiment i avaluem la proporció d'èxits p. Aleshores p té una distribució aproximadament normal, amb una mitjana π i una desviació estàndard:

$$\sqrt{\frac{\pi(1-\pi)}{n}}.$$

Ara executeu el programa MacAnova.

Per a il·lustrar el mostreig a partir d'una distribució binària hem preparat un programa que trobareu emmagatzemat en el fitxer `PROPN.MAC`, el qual pren 400 mostres de grandària n a partir d'una distribució binària amb una mitjana poblacional π i determina en cada mostra la proporció d'èxits. El programa pren 400 mostres, per tant, acabem amb 400 proporcions estimades emmagatzemades en el vector `mean`. La mitjana i la desviació estàndard d'aquestes estimacions es calcula i s'emmagatzema dins `sample_mean` i `sample_sd` respectivament. Primer introduïu la mida de la mostra i el valor π que vulgueu fer servir; en la nostra il·lustració usem $n = 100$ i $\pi = 1/6$:

```
Cmd> n<-100 ; pi<-1/6
```

Després executeu el programa que trobareu en el fitxer `PROPN. MAC` per mitjà de l'ordre: `batch ("PROPN. MAC",echo:F).`

Recordeu...

... un cop més que els vostres resultats seran lleugerament diferents.

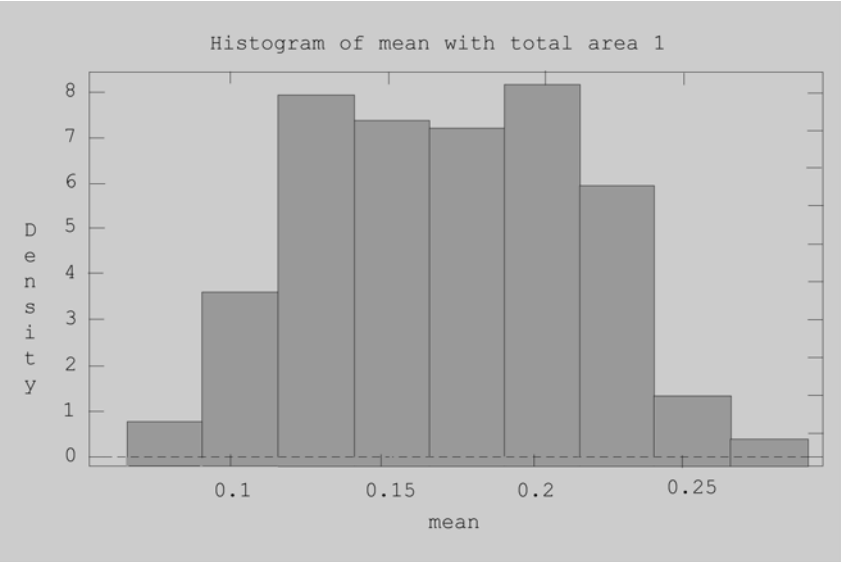
La funció `batch ("nomdelfitxer")` executa un conjunt d'ordres emmagatzemades en un fitxer, i l'opció `echo:F` suprimeix la impressió del programa en la pantalla. El resultat en el nostre cas ha estat el següent:

```
Cmd> batch("PROPN.MAC",echo:F)

Mean of 400 proportions
(1)       0.16802

Standard deviation of 400 proportions
(1)      0.038569
```

La mitjana de totes les mitjanes mostrals s'acosta molt a la mitjana poblacional d'1/6 = 0,16667, mentre que la desviació estàndard també s'acosta molt a la desviació estàndard que hem treballat teòricament abans.

Com que totes 400 proporcions han estat emmagatzemades en el vector `mean`, si fèiem servir la funció `hist()`, en podríem veure la distribució:

```
Cmd> hist(mean)
```



Observeu que la distribució de les mitjanes sembla de perfil normal.

Activitats

2. Torneu a executar el programa MacAnova del fitxer `PROPN.MAC`, fent servir el mateix valor de π (1/6), però amb una mostra de grandària més petita igual a 20. Feu-ne l'histograma i comenteu-ne el perfil.

3. Suposem que sabeu que el 0,48 dels nounats són mascles. Suposem que teniu informació sobre 50 nadons, que podem prendre com una mostra aleatòria. Quin nombre de nounats esperaríem trobar en aquesta mostra aleatòria i quina és la desviació estàndard d'aquest nombre?

4. Verifiqueu els resultats de l'activitat 1 empíricament, executant el programa dins l'arxiu `PROPN.MAC` amb els retocs adequats de n i π.

El **concepte principal** que hem vist en aquest apartat és el següent:

Variable binària: variable categòrica que únicament té dos valors possibles, codificats com a 0 i 1, respectivament. Anomenem la categoria amb el codi 1 un *èxit* i la categoria amb el codi 0 un *fracàs*.

La distribució d'una variable binària és totalment descrita per la proporció d'èxits, indicada per π. La densitat de probabilitat de la distribució col·loca una probabilitat de π en el valor 1 i una probabilitat $(1 - \pi)$ en el valor 0.

La mitjana poblacional de la variable binària és igual a π.

La desviació estàndard de la població d'una variable binària és igual a $\sqrt{\pi(1-\pi)}$.

Si prenem una mostra de grandària n d'una distribució binària amb la probabilitat d'èxit π, aleshores la mitjana d'aquesta mostra té –per a una n gran– una distribució normal amb una mitjana π i una desviació estàndard $\sqrt{\pi(1-\pi)/n}$.

13. Inferència estadística (I): interval de confiança per a la mitjana aritmètica

El procés d'utilitzar l'estadística per a arribar a una conclusió sobre algun aspecte concret de la població s'anomena *inferència estadística*.

En aquest apartat presentem un tipus de deducció molt útil que implica calcular el grau de precisió de les nostres estimacions de les mitjanes poblacionals, anomenat **interval de confiança**.

En aquest apartat sobre els intervals de confiança aprendreu:

- què és un marge d'error;
- què és un interval de confiança i com s'interpreta;
- què és un nivell de confiança;
- com es calcula un interval de confiança per a la mitjana aritmètica d'una distribució normal amb variància coneguda.

La precisió de l'estimació

Hem vist algun exemple de l'estimació de la mitjana desconeguda μ d'una població, com ara quan parlàvem d'estimar la proporció de persones que votaran a un partit específic en unes eleccions, en què la veritable proporció π és una mitjana poblacional. Hem suposat que la mitjana poblacional és un valor fix que solament podríem mesurar amb exactitud si coneguéssim la població sencera. Per tant, prenem una mostra aleatòria d'observacions i fem servir la mitjana de la mostra per a estimar el valor poblacional.

Vegeu l'apartat 12 d'aquesta assignatura.

També hem vist que la mitjana mostral és en si mateixa una variable aleatòria i que té la seva pròpia distribució mostral. Per tant, si preníem una altra mostra, obtindríem una estimació diferent de la mitjana poblacional μ.

En la pràctica, però, tan sols tenim una única mostra i una única estimació de la mitjana. Sabem que, si la nostra mostra hagués estat més àmplia, aleshores la seva variabilitat seria més petita, i això suggereix clarament que una mostra tal seria una estimació més precisa de μ. Però com podem mesurar la precisió de les nostres estimacions?

Els intervals de confiança

Pensem ara en dos diaris diferents que facin prediccions sobre quin serà el percentatge de la població que participarà en unes eleccions. L'un prediu que el percentatge serà del 71%, mentre que la predicció de l'altre és del 76%. Després de les eleccions el veritable percentatge és exactament el 75% –sembla, doncs, que la segona empresa havia fet una predicció més acurada–. Solament ho podem comprovar si sabem el veritable percentatge, i en la pràctica és molt rar que tinguem una situació en què el veritable valor d'un paràmetre poblacional sigui conegut. En gairebé totes les situacions estimem uns valors poblacionals que no podrem confirmar mai.

Podem comprovar l'exactitud d'una predicció sempre que disposem *a posteriori* de les dades reals.

Per tant, com podem quantificar la precisió de les nostres estimacions quan tenim solament una única mostra de dades i cap manera de confirmar-ne el resultat? La manera de fer-ho és no donar una única estimació del valor poblacional, sinó tot un ventall de valors, i després reforçar aquest ventall de valors per mitjà d'una declaració del vostre grau de confiança que el veritable valor es troba dins aquest ventall. Això s'anomena **interval de confiança**.

L'interval de confiança per a la mitjana d'una distribució normal

Considerem un cas simple per començar, quan la població és normal i coneixem la desviació estàndard σ d'aquesta distribució (és molt poc freqüent que coneguem la desviació estàndard de la població, normalment l'estimem també a partir de la mostra).

L'estimació de la desviació estàndard a partir de la mostra s'exposarà en l'apartat 14.

El procediment per calcular un interval de confiança per a la mitjana μ basat en una mostra de grandària n és el següent:

El marge d'error...

... d'un interval de confiança és el radi de l'interval al voltant de la mitjana, és a dir, $z^* \sigma_{\bar{x}}$.

1) Calculeu la mitjana $\bar{x}$ de la mostra.

2) Calculeu l'error estàndard de la mitjana: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Nivell de confiança

El valor $(1 - \alpha)$ representa la probabilitat que el paràmetre estimat estigui inclòs en l'interval de confiança (és a dir, l'àrea de la corba normal inclosa entre $-z^*$ i z^*).

3) Calculeu el marge d'error com a z^* per l'error estàndard: $z^* \sigma_{\bar{x}}$.

4) Tenim així que l'interval de confiança és la mitjana més menys el marge d'error: $\bar{x} \pm \sigma_{\bar{x}}$.

La interpretació d'un interval de confiança

Per a il·lustrar el que volem dir amb un nivell específic de confiança, simulem una sèrie d'intervals de confiança fent servir mostres extretes d'una distribució coneguda i vegem si contenen la veritable mitjana poblacional o no. Posem per cas, doncs, que una empresa comprova la durada de les piles, 20 piles cada cop. Suposem que la distribució de la durada d'una sola pila és normal. Malgrat que habitualment no sabríem la mitjana (això és el que provem de desco-

brir), imaginem que la veritable durada mitjana de les piles és de 52,6 hores. Suposem també que la desviació estàndard poblacional és de 6,2 hores.

Hem preparat un petit programa en l'arxiu `CONFINT.MAC` que selecciona una mostra aleatòria de mida 20 a partir d'una distribució normal amb una mitjana de 52,6 i una desviació estàndard de 6,2; calculeu la mitjana de la mostra, l'error estàndard, el marge d'error (fent servir $z_{0,025} = 1,96$ per a obtenir un nivell de confiança del 95%) i finalment els límits inferior i superior de l'interval de confiança, els quals són impresos. Executem, doncs, aquest programa una vegada:

Ara executeu el programa MacAnova.

```
Cmd> batch "CONFINT.MAC",echo:F)
NOTE: random number seeds set to 724044218 and
320455728
Lower and upper limits of confidence interval
(1)        52.553
(1)        57.899
```

Això simula la mostra de 20 piles que examinem. En mesurem les durades, calculem el marge d'error i l'interval de confiança del 95%, el qual en aquest cas és [52,553; 57,899]. Ara sabem que la veritable mitjana és de 52,6; per tant, veiem que el nostre interval de confiança sí que inclou el veritable valor. En aquest cas el nostre interval de confiança ha funcionat.

Podem, però, continuar simulant mostres addicionals a partir de la mateixa distribució de la mateixa manera. Per fer això, executeu la mateixa ordre amb un ordinador que treballi amb l'entorn de DOS, premeu la tecla de funció F3, i aquesta repetirà la línia de l'última ordre introduïda en l'ordinador. Aquí teniu quatre repeticions d'aquesta ordre:

```
Cmd> batch "CONFINT.MAC",echo:F)
Lower and upper limits of confidence interval
(1)        48.343
(1)         53.69

Cmd> batch "CONFINT.MAC",echo:F)
Lower and upper limits of confidence interval
(1)        49.133
(1)         54.48
```

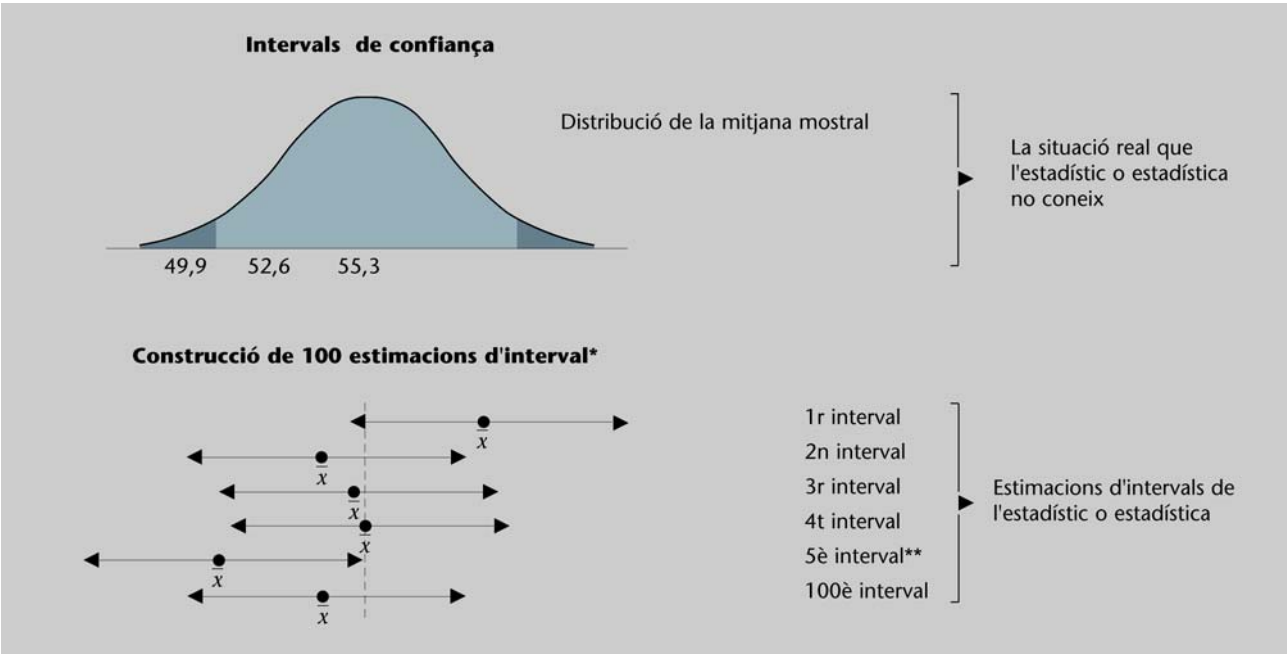
Recordeu...

... un cop més que la vostra execució particular d'aquest programa donarà resultats diferents, perquè el generador de nombres aleatoris produeix mostres diferents.

```
Cmd> batch "CONFINT.MAC",echo:F)
Lower and upper limits of confidence interval
(1)          49.3
(1)        54.647

Cmd> batch "CONFINT.MAC",echo:F)
Lower and upper limits of confidence interval
(1)        46.255
(1)        51.602
```

Els tres primers exemples inclouen la veritable mitjana, però el quart, no: aquest darrer interval de confiança no funciona. Podem continuar executant aquesta simulació tant de temps com vulguem, observant quina conté la veritable mitjana de 52,6 i quina, no. Nosaltres hem continuat fins a completar un total de 100 simulacions i hem observat que 3 d'aquestes no incloïen la mitjana poblacional. El que veuríem si continuàvem l'estudi d'aquesta simulació milers de vegades és que al voltant del 95% dels intervals inclouen 52,6 i al voltant del 5% restant, no.

En altres paraules, al voltant del 95% funcionen, i el 5%, no. Això és el que volem dir mitjançant el nivell de confiança del 95%.



* Al voltant del 95% inclouen correctament l'objectiu μ.
** Aquest interval no inclou la veritable mitjana.

Activitats

1. Continueu executant la vostra pròpia simulació fins a un total de 100 vegades i compteu quants intervals no inclouen la mitjana poblacional de 52,5.

L'intercanvi entre la precisió i el nivell de confiança

Si rebaixàvem el nivell de confiança al 90%, el marge d'error seria més petit (ja que el valor z seria més petit d'1,645 en contraposició a 1,96), i l'interval de confiança seria més curt. Aquest sembla un resultat més precís, però el nivell de confiança naturalment és més baix: ara la possibilitat (1 entre 10) que l'interval no inclogui la veritable μ és més gran. Per tant, res no és gratuït. Hi ha un intercanvi entre la precisió que es pot expressar en un interval de confiança i el nivell de confiança. Per a una mostra en particular, com més curt i precís sigui l'interval de confiança, més baix serà el nivell de confiança.

L'efecte de la grandària de la mostra

L'única manera de millorar tant la vostra precisió com el vostre nivell de confiança és reduir l'error estàndard. Si la desviació estàndard poblacional σ és fixa, aleshores únicament podem reduir l'error estàndard mitjançant l'increment de la mida mostral. Això redueix el marge d'error i així s'escurça l'interval de confiança per a un nivell de confiança en particular. Alternativament, si es manté el marge d'error en un nivell fix, incrementar la mida mostral porta a incrementar el valor z^* i per conseqüent també el nivell de confiança.

Fixeu-vos que, com que l'error estàndard s'obté dividint la desviació estàndard per l'arrel quadrada de n, es necessita una mostra quatre vegades més gran per a reduir l'amplada de l'interval de confiança a la meitat.

Recordeu

L'error estàndard és $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Activitats

2. Es pren una mostra aleatòria de grandària 50 d'una distribució normal. La desviació estàndard d'aquesta distribució és 0,34. La mitjana aritmètica de la mostra és 1,89. Calculeu:
a) un interval de confiança del 95% per a la mitjana m de la distribució;
b) un interval de confiança del 99%.

3. A partir d'unes dades prèvies sabem que el nivell de pol·lució de l'aire urbà, mesurat amb un índex de pol·lució de 0 a 100, és normalment distribuït, amb una desviació estàndard de 13 unitats. En un dia bo la pol·lució a la zona és de 25-30 unitats i en un dia dolent arriba fins a 70. Suposem que prenem 4 mesuraments al llarg d'un dia i obtenim una mitjana de pol·lució d'índex 46. Quin és l'interval de confiança del 95% per al veritable nivell de pol·lució aquest dia?

4. En connexió amb l'activitat 3, suposem que volem estimar un interval de confiança del 95% per a un nivell de pol·lució tal que el marge d'error és com a màxim de 5 unitats. Quants mesuraments independents necessitem prendre?

5. Un banc comprova el temps de resposta de la seva xarxa nacional de caixers automàtics. D'estudis anteriors se sap que el temps de resposta és al voltant de 10 segons amb una desviació estàndard de 2 segons. Els preocupa que el temps augmenti i volen establir el temps de la mitjana actual de resposta amb una precisió de 0,5 segons. Quina grandària mostral hauríem de prendre per a obtenir tal precisió? Suposem que prenen una mostra aleatòria de 10 temps de resposta i troben que la mitjana és 12,4 segons. Això evidencia que el temps de resposta de la xarxa ha augmentat?

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Interval de confiança: estimació d'un paràmetre poblacional en forma d'interval, en el qual confiem que es troba el paràmetre.

Marge d'error: precisió de l'interval de confiança; l'interval de confiança és donat en forma d'una estimació més menys el marge d'error.

Per a un interval de confiança sobre la mitjana μ d'una distribució normal amb una desviació estàndard coneguda σ, es calcula la mitjana aritmètica $\overline{x}$ de la mostra aleatòria de mida n, i aleshores l'interval és:

$$\overline{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

on $z_{\alpha/2}$ és el punt apropiat sobre la corba normal de tal manera que una àrea de $(1 - \alpha)$ s'inclou sota la corba entre $\pm z_{\alpha/2}$.

Nivell de confiança: probabilitat $(1 - \alpha)$ que el nostre interval de confiança inclogui el veritable paràmetre poblacional. Expressat com un percentatge és $100\ (1 - \alpha)\%$.

14. Inferència estadística (II): interval de confiança per a una proporció

En l'apartat 12 vèiem que una proporció, o percentatge, es pot calcular com a mitjana aritmètica d'un conjunt de dades binàries codificades com a 0 o 1. També deduíem la desviació estàndard d'una proporció, o el seu error estàndard. No podem suposar que les dades binàries són normals, però sabem que, pel que fa a mostres grans, la distribució de la mitjana mostral és aproximadament normal. D'aquesta manera podem usar tots els resultats obtinguts fins ara per a aconseguir els intervals de confiança per a una proporció.

Recordeu

Si x es una variable binaria, la mitjana es una proporció π i la desviacio estandard es $\sigma = \sqrt{\pi(1-\pi)}$.

En aquest apartat sobre intervals de confiança aprendreu:

- com es calcula un marge d'error i un interval de confiança per a un percentatge;
- com es calcula la mida mostral que dóna un marge específic d'error en l'estimació d'un percentatge.

Els percentatges presentats en els mitjans de comunicació

Nosaltres sentim diverses estimacions percentuals gairebé cada dia: l'índex d'atur, el percentatge de persones que votaran a un partit polític determinat, el percentatge de consumidors que trien tal sabó o tal diari, el percentatge de mals conductors i conductores en les nostres carreteres, i així successivament. Totes aquestes estimacions es basen en una mostra a partir d'una població, però gairebé mai no se'ns dóna la precisió de l'estimació. En alguns casos és possible que vegem una petita nota a peu de pàgina informant d'algun marge d'error o –com en aquestes rares excepcions extretes del *New York Times*– l'explicació següent quant al mètode utilitzat per a dur a terme un sondeig d'opinió:

> "En teoria, es pot dir que en 95 casos de cada 100 els resultats basats en la totalitat de la mostra no difereixen en més de tres punts percentuals en una i altra direcció d'allò que s'hauria obtingut si s'hagués entrevistat tota la població adulta nord-americana."

La distribució del percentatge, o proporció

Noteu que...

... sempre fem servir proporcions a l'hora de parlar de qüestions teòriques, però en la pràctica normalment donem els resultats sobre una escala percentual.

En l'apartat 12 vèiem que es pot considerar una proporció com la mitjana d'un conjunt de mesuraments 0 o 1. Per tant, per a mostres grans, hem vist que una proporció calculada té una distribució normal aproximada, amb una mitjana igual a la proporció poblacional μ i una desviació estàndard (és a dir: l'error estàndard) igual a

$$\sqrt{\frac{\pi(1-\pi)}{n}}.$$

Vegeu el gràfic I de l'apartat 12.

Però, què vol dir *gran* per a nosaltres? La distribució binària pot ser molt asimètrica quan la proporció de la població no és a prop del 0,5. Cal tenir una mida mostral d'almenys 100 unitats perquè el teorema central del límit sigui aplicable, i en tot cas en necessitareu almenys 100 per a poder estimar el per-

centatge correcte per a un punt percentual. Per tant, solament estudiarem proporcions calculades sobre mostres de 100 o més unitats.

Solament tenim un problema que hem de resoldre abans d'aplicar la teoria que hem desenvolupat: l'error estàndard depèn del veritable valor de π que provem d'estimar; per tant, com calculem el marge d'error? La solució és substituir el valor de la proporció mostral p, que és la nostra estimació de π, en la fórmula per a l'error estàndard. Per exemple, si tenim una estimació per a π de $p = 0{,}37$, basada en una mida mostral de 100, calculem l'error estàndard com a

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}37 \cdot 0{,}63}{100}} = 0{,}0483.$$

Per tant, un interval de confiança del 95% per a la proporció poblacional π seria $0{,}37 \pm 1{,}96 \cdot 0{,}0483 = 0{,}37 \pm 0{,}095$, un interval de 0,275 a 0,465.

Recordeu

L'interval de confiança és $\mu \pm z_{\alpha/2}\sigma_{\bar{x}}$.

Interval de confiança per a una proporció

Els passos per a calcular un interval de confiança per a una proporció són els següents:

1) Calcular la proporció p d'èxits a partir de la mostra de grandària n.

2) Calcular l'error estàndard de la proporció: $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$.

3) Calcular el marge d'error com a $z_{\alpha/2}$ per l'error estàndard: $z_{\alpha/2}\ \sigma_p$, en què $z_{\alpha/2}$ és el valor apropiat de la variable normal estàndard per al nivell de confiança $100\ (1 - \alpha)\%$.

4) L'interval de confiança és la proporció observada p més menys el marge d'error: $p \pm z_{\alpha/2}\ \sigma_p$.

Activitats

1. En una mostra aleatòria de barcelonins, el 10% tenen un cotxe aparcat al carrer. Construïu un interval de confiança del 90% per a la proporció de la població de Barcelona que té un cotxe aparcat fora, si la grandària mostral és:
a) $n = 125$;
b) $n = 500$.

2. Es va dur a terme un sondeig d'opinió a Espanya, i una de les preguntes fetes a una mostra aleatòria de 1.500 persones era: "Us sembla que l'economia millorarà el 1998 o no?". De les persones enquestades 473 (31,5%) van dir que sí, 967 (64,5%) van dir que no, i 60 (4,0%) van dir que no ho sabien. Construïu els intervals de confiança del 95% per a cadascun dels percentatges de les respostes "sí" i "no".

3. Durant unes eleccions municipals en què participaven dos partits, CiU i PSOE, es va dur a terme un sondeig d'opinió en què es preguntava a 1.000 votants seleccionats a l'atzar quin partit votarien. Un total de 615 van indicar la seva preferència per CiU. Construïu un interval de confiança del 95% per a la proporció de vots que s'emetran a favor de CiU. CiU pot pensar que té la victòria assegurada?

Les grandàries mostrals per a un marge d'error prèviament establert

En general, el marge d'error per a estimar una mitjana amb un nivell de confiança 100 $(1 - \alpha)$% a partir d'una mostra de grandària n és:

$$\text{marge d'error} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

en què σ és la desviació estàndard de la distribució poblacional. Per a una determinada grandària de la mostra n podem calcular el marge d'error. D'altra banda, si prèviament establim el marge d'error que requerim per a la nostra estimació, podrem calcular la grandària mostral.

L'activitat 4 de l'apartat anterior era un exemple de la idea que s'exposa en el text central.

Expressem ara la grandària mostral n a partir de la fórmula anterior en termes dels altres factors:

$$\text{grandària mostral} = \left(z_{\alpha/2}\frac{\sigma}{\text{marge d'error}}\right)^2.$$

Això mostra clarament que per a reduir el marge d'error a la meitat, per exemple, cal incrementar la grandària mostral quatre vegades.

Si apliquem aquesta fórmula a la nostra situació actual d'estimació de proporcions, en què $\sigma = \sqrt{\pi(1-\pi}$, obtenim:

$$\text{grandària mostral} = \left(z_{\alpha/2}\right)^2\frac{\pi(1-\pi)}{(\text{marge d'error})^2}.$$

Aquesta fórmula és útil com a anticipació a una enquesta per sondatge per a determinar la mida mostral requerida per a estimar una proporció amb una precisió determinada. Però per a aplicar aquesta fórmula necessitem conèixer π, la proporció que provem d'estimar. Si en realitat $\pi = 0{,}25$, aleshores $\pi\,(1 - \pi) = 0{,}1875$; mentre que, si $\pi = 0{,}10$, $\pi\,(1 - \pi) = 0{,}09$, cosa que és la meitat del valor anterior i implicaria que es requereix la meitat de la grandària mostral.

Quin valor de π hem d'usar? Això depèn de si tenim alguna idea aproximada de la proporció poblacional o no la tenim. Per exemple, podem estar força segurs que la popularitat d'un partit polític és al voltant del 30% de la població, però volem dur a terme una enquesta per a determinar aquest percentatge amb més precisió, diguem que amb un marge d'error de 2 punts percentuals. Podríem usar el valor 0,30 per a determinar la grandària mostral requerida:

$$\text{grandària mostral } = 1{,}96^2\frac{0{,}30\cdot 0{,}70}{0{,}02^2} = 2.017.$$

D'altra banda, si no tenim cap idea inicial de la proporció poblacional, o si duem a terme una enquesta per a estimar proporcions diverses, algunes de les quals poden ser altes i d'altres baixes (per exemple, la població de diferents partits polítics), aleshores haurem d'usar el valor 0,5 per a π. La proporció 0,5 dóna el valor més alt de $\pi\,(1-\pi) = 0{,}5 \cdot 0{,}5 = 0{,}25$ i així ens dóna la grandària mostral màxima necessària per a obtenir el marge d'error per a qualsevol proporció. Per tant, per a obtenir un marge d'error de 2 punts percentuals per a estimar qualsevol proporció, la grandària mostral hauria de ser:

$$\text{grandària mostral } = 1{,}96^2 \frac{0{,}5 \cdot 0{,}5}{0{,}02^2} = 2.401.$$

Fent servir aquesta fórmula podem obtenir les grandàries mostrals màximes necessàries per a qualsevol marge d'error, per exemple des del 5% fins a l'1%:

Marge d'error	Grandària mostral
5%	384
4%	600
3%	1.067
2%	2.401
1%	9.604

Activitats

4. En una gran empresa agrícola separen les pomes de qualitat alta de les de qualitat baixa. En anys passats els percentatges de pomes de qualitat alta i baixa han estat aproximadament de 50:50. Després d'una temporada de molt poca pluja, el director de l'explotació vol comprovar el percentatge de pomes de qualitat baixa i voldria tenir una estimació del percentatge veritable amb una precisió de 5 punts percentuals. Quantes pomes haurien d'inspeccionar a l'atzar per a obtenir una estimació amb un nivell de confiança del 90%?

5. S'ha introduït un nou formulari de les rendes, i el govern vol estimar el percentatge de formularis emplenats incorrectament. Quants formularis haurien de comprovar a l'atzar abans de poder arribar a una estimació amb una precisió d'1 punt percentual? (Si no s'especifica un nivell de confiança, useu el nivell del 95%.)

6. Una empresa de serveis estadístics duu a terme enquestes mensuals per a estimar una àmplia varietat d'opinions sobre qüestions socials. Per a assegurar una precisió general de 2,5 punts percentuals o més en les seves estimacions, quina grandària mostral hauria d'usar aquesta empresa?

Els **conceptes principals** que hem vist en aquest apartat són els següents:

Marge d'error per a una proporció estimada: el marge d'error per a una proporció estimada p és:

$$z_{\alpha/2} = \sqrt{\frac{p(1-p)}{n}},$$

on $z_{\alpha/2}$ és el valor normal estàndard apropiat que talla una probabilitat $\alpha/2$ a la cua de la corba; això correspon a un interval de confiança 100 $(1 - \alpha)$% i s'hauria d'usar solament per a mostres d'almenys 100 unitats.

La grandària mostral: la grandària mostral requerida per a estimar una proporció amb un marge d'error determinat és donada per la grandària:

$$\left(z_{\alpha/2}\right)^2 \frac{\pi(1-\pi)}{(\text{marge d'error})^2}.$$

on π és la veritable proporció poblacional. En la pràctica usem un valor aproximat per a π, basat en l'experiència, o el valor $\pi = 0{,}5$, que ens donarà la grandària mostral màxima requerida.

15. Bondat de l'ajustament: ajustament de les dades a les distribucions teòriques

Una qüestió important en les ciències socials és si les nostres observacions s'ajusten a una distribució de freqüències teòriques donada. Això és important en verificar si una mostra és representativa d'una població determinada en funció d'algunes característiques conegudes com l'edat i la posició socioeconòmica.

Per exemple, suposem que sabem que el 14% de la nostra població té estudis universitaris; en aquest cas hauríem d'obtenir aproximadament el mateix percentatge en una mostra representativa. Si la mostra contingués un percentatge diferent de persones amb estudis universitaris, suposem que el 13%, llavors això seria la prova de manca de representativitat? Sembla que aquesta diferència és massa petita per a preocupar-nos. Però si la mostra contingués només el 3% amb estudis universitaris, llavors estaríem segurs que la mostra no és representativa. Fins a quin punt podem dir que una mostra no s'ajusta a una població?

En aquest apartat sobre bondat de l'ajustament, aprendreu a:

- calcular les freqüències esperades d'un conjunt de classes en una mostra d'informació coneguda sobre la població;
- comparar les freqüències esperades amb les freqüències observades i a calcular una mesura d'ajustament anomenada *estadística* χ^2;
- usar la distribució χ^2 per a jutjar si les freqüències de mostra difereixen significativament de les freqüències esperades.

Suposeu que llanceu enlaire una moneda 100 vegades i compteu el nombre de cares obtingut. Abans de fer-ho, quantes espereu obtenir-ne? Sembla lògic que esperem obtenir 50 cares i 50 creus. Això és, naturalment, una previsió teòrica, basada en la suposició que hi ha la mateixa possibilitat d'obtenir cara que creu. En la pràctica el resultat seria diferent però molt a prop de 50 cares (l'aproximació real és el que ens interessa aquí).

Quan es llança una moneda a l'aire, si la moneda no està trucada, hi ha la mateixa probabilitat d'obtenir cara que d'obtenir creu.

Usem el MacAnova per a llançar enlaire la moneda i per a comptar el nombre de cares:

```
Cmd> rand < runi(100)
NOTE: random number seeds set to 870960853
and 1838139030
```

```
Cmd> rand
(1)  0.25969  0.32721  0.10062  0.45281   0.75994
(6)  0.15048  0.34246  0.52913  0.83059   0.12310
 .      .        .        .        .         .
 .      .        .        .        .         .

Cmd> toss <- floor(rand*2)

Cmd> toss
(1)        0        0        0        0        1
(6)        0        0        1        1        0
 .         .        .        .        .        .
 .         .        .        .        .        .

Cmd> sum(toss)
(1)     53
```

- La primera instrucció genera 100 nombres aleatoris entre 0 i 1 i els emmagatzema en el vector `rand`.
- La segona instrucció llista el contingut de `rand` (aquí només mostrem els primers 10 valors).
- La tercera instrucció, en primer lloc, multiplica els nombres aleatoris emmagatzemats en el vector `rand` per 2 per obtenir nombres aleatoris entre 0 i 2, i aleshores aplica la transformació `floor()` per arrodonir els nombres als enters 0 i 1.
- La quarta instrucció mostra els valors del resultat que emmagatzemem en el vector `toss`. En aquest cas el número 1 correspon a la cara, i el 0, a la creu. Per això, tots els valors de `rand` fins a 0,5 es transformaran en 0, i tots els valors més grans que 0,5 es transformaran en 1.
- La cinquena instrucció suma els valors del llançament; en altres paraules, compta el nombre de cares. Així, en aquest experiment obtenim 53 cares.

Proveu-ho vosaltres mateixos i veieu què obteniu. Podeu fer l'experiment sencer en una instrucció que combini els tres càlculs anteriors: `sum(floor(runi(100)*2))`. Aquí teniu cinc experiments més d'aquest tipus:

```
Cmd> sum(floor(runi(100)*2))
(1)             50

Cmd> sum(floor(runi(100)*2))
(1)             44
```

```
Cmd> sum(floor(runi(100)*2))
(1)             52

Cmd> sum(floor(runi(100)*2))
(1)             44

Cmd> sum(floor(runi(100)*2))
(1)             58
```

En general obtenim valors per sobre o per sota de 50, encara que veiem que en un cas obtenim exactament 50 cares.

Recordeu...

... que no obtindreu exactament els mateixos resultats, excepte si comenceu a generar els vostres nombres aleatoris usant les mateixes llavors.

Activitats

1. Feu aquest experiment 100 vegades (recordeu que en DOS l'última instrucció es pot repetir prement la tecla de funció F3). Compteu quantes vegades el nombre de cares obtingut és de 40 a 60, i quantes vegades sobrepassa aquests límits.

Canviem l'experiment de context i fem una pregunta pertinent. Suposeu que tenim una mostra de 100 persones i que els demanem què pensen de la moneda única europea. Trobem que 61 persones de la mostra hi estan a favor i 39, en contra. La pregunta que volem contestar és si això demostra que una majoria de la població hi està a favor. Compareu aquesta qüestió d'estar a favor o en contra de la moneda única amb el llançament d'una moneda, de manera que, si no hi havia majoria en cap de les dues direccions, el resultat de la nostra enquesta seria com un dels resultats de l'experiment del llançament de monedes efectuat abans. Ara nosaltres volem saber si el resultat real de 61 contra 39 és inusualment diferent de l'esperat 50 contra 50. Podem contestar aquesta pregunta de diverses maneres diferents:

Si en una enquesta determinada 61 persones de la mostra estan a favor de la qüestió enunciada, vol dir que la majoria de la població hi està a favor o no?

1) Solució empírica

Una manera empírica de contestar la pregunta és fer l'experiment anterior moltes vegades (per exemple milers de vegades), i tindrem clar que el resultat obtingut de 61 cares (o més) és molt poc freqüent –gairebé sempre el nombre de cares està entre 40 i 60–. Això ens portaria a creure que 61 és un valor inusual i que la idea que la població està dividida 50:50 no és correcta. Concloure'm que més del 50% de la població està a favor de la moneda única. Això equival a concloure que la moneda no és justa perquè mostra més vegades cara que creu.

2) Interval de confiança

Vegeu l'apartat 14 d'aquesta assignatura.

Una altra manera de contestar la pregunta és fer servir el que ja sabem sobre intervals de confiança per a un percentatge. Tenint en compte que són 61 de 100, un percentatge de 0,61 hi estan a favor, el marge d'error és $1{,}96 \cdot \sqrt{0{,}61 \cdot 0{,}39) / 100} =$ $= 0{,}096$, i es dóna un interval de confiança de 95% per als que hi estan a favor

de [0,514; 0,706]. Com que aquest interval de confiança no inclou 0,5, conclouríem que la proporció de població és més gran que 0,5.

3) Estadística χ^2

Finalment, hi ha un mètode general per a comparar les nostres observacions amb les nostres previsions, i aquest és el tema d'aquest apartat. L'avantatge d'aquest tercer enfocament és que es pot ampliar fàcilment en el cas en què hi hagi més de dues categories. En altres paraules, podem usar aquest mètode per a jutjar respostes a preguntes com "A quin partit polític dóna suport?", on hi ha diverses respostes possibles, i no solament el cas "sí/no" descrit aquí.

El mètode funciona de la manera següent:

- A la primera columna, llisteu-hi totes les categories.
- A la segona columna, llisteu-hi les freqüències esperades en la mostra; aquestes tenen alguna presumpció prèvia sobre la població, i en aquest cas aquesta opinió es divideix en parts iguals sobre la qüestió de la moneda única.
- A la tercera columna, llisteu-hi les freqüències trobades en la mostra.
- A la quarta columna, calculeu-hi les diferències entre les freqüències observades i les esperades.
- La cinquena columna mostra com combinem les diferències en un diagrama que mesura la diferència global entre el que observem i el que esperem. Cada diferència es calcula al quadrat i es divideix per la freqüència esperada corresponent. En aquest cas particular obtenim dos valors idèntics d'$11^2/50 = 2{,}42$, i aquests sumen 4,84.

Categoria	Freqüències esperades	Freqüències observades	Diferències	Diferències al quadrat dividides per freqüències esperades
	E_i	O_i	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
A favor	50	61	11	2,42
En contra	50	39	-11	2,42
Total	**100**	**100**	**0**	**4,84**

Observeu la fórmula de l'estadística χ^2 que mesura la bondat d'ajustament: és la suma de les diferències al quadrat entre les freqüències observades i les esperades, cadascuna dividida per la freqüència esperada.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

El valor 4,84 mesura la diferència entre les freqüències observades i les esperades. Aquesta és la mesura de bondat d'ajustament de les nostres observacions respecte d'una previsió teòrica. Si haguéssim observat 60 a favor i 40 en contra, aquest valor seria 4; si haguéssim observat 55 a favor i 45 en contra, el valor seria 1.

Clarament la bondat d'ajustament és zero quan les observacions i les previsions es corresponen perfectament, i n'augmenta el valor a mesura que les observacions s'allunyen de les esperades (en aquest sentit, s'hauria d'anomenar *mesura de "maldat" d'ajustament*).

En comptes de mirar les observacions originals, mirem l'única mesura d'ajustament. L'única qüestió pendent és saber què és per a nosaltres un valor acceptable d'aquesta mesura per tal que concloguem que les nostres observacions s'ajusten a les nostres previsions. O, per dir-ho d'una altra manera, podem fer la pregunta: quan es dóna un valor de bondat d'ajustament tan alt que comencem a dubtar de la presumpció en la població en la qual es basen les freqüències esperades?

Convenientment la bondat d'ajustament segueix d'una manera aproximada una distribució estadística molt coneguda denominada **distribució χ^2**, una de les distribucions usades més sovint en estadística. Associat amb una distribució χ^2 hi ha un concepte conegut com **graus de llibertat**.

En anglès, khi quadrat, χ^2, s'anomena *chi square*.

No és possible, en aquesta assignatura, fer una explicació completa de la distribució χ^2 i els graus de llibertat, incloent-hi les matemàtiques necessàries. Només heu de recordar que els graus de llibertat associats a l'estadística χ^2 en aquesta situació són iguals al nombre de categories menys una. D'aquesta manera, aquí tenim dues categories, els graus de llibertat són iguals a 1.

Sovint, la mesura de bondat d'ajustament mateixa la denominem *estadística* χ^2 i observem el que s'anomenen *punts crítics* en taules de la distribució χ^2 o usant un programa com el MacAnova.

Si consulteu...

... la taula C de l'annex 3, veureu que el punt crític de la mesura és 3,84. Això vol dir que els valors de la bondat d'ajustament per sota de 3,84 són acceptables, mentre que els que són per sobre d'aquest valor indiquen diferències entre les freqüències observades i les esperades que és improbable que hagin ocorregut purament de casualitat.

La nostra conclusió és llavors, com abans, que 61 de 100 a favor de la moneda única no és conseqüent amb la presumpció que no hi ha majoria en la població, així que decidim que hi està a favor una majoria de la població.

El que hem estat fent és un exemple simple d'un test d'hipòtesi (contrast d'hipòtesi). Proposem un model senzill per a les observacions, en aquest cas, en què hi ha una proporció igual de persones a favor i en contra de la moneda única. A continuació mesurem la diferència entre què esperaríem si el model fos veritat i què observem, per veure si la hipòtesi és creïble o no ho és. L'estadística χ^2 mesura la diferència en un nombre, i la distribució χ^2 ens proporciona una manera de jutjar la credibilitat de la hipòtesi. Aquestes proves de la bondat d'ajustament se solen trobar en el terreny de la recerca de les ciències socials.

Més de dues categories

Estudiem ara amb cura un altre exemple on trobem més de dues categories. Considerem la població de 860 alumnes de ciències empresarials:

Curs	Alumnes
1r curs	172
2n curs	238
3r curs	331
4t curs	119
Total	**860**

Cada dimecres hi ha un programa de música de jazz a l'hora de dinar, i els organitzadors estan interessats a saber si els que hi assisteixen són representatius de la població. A un d'aquests esdeveniments, hi van assistir 56 alumnes, i s'ha constatat que estan distribuïts de la manera següent:

Curs	Alumnes
1r curs	7
2n curs	8
3r curs	30
4t curs	11
Total	**56**

Com abans, usem la informació de població per a determinar què preveuríem en aquesta mostra si fos només una mostra aleatòria de la població. Tenint en compte que són 172 de 860, és a dir 0,20 o el 20% de la població són de primer curs, $0{,}20 \cdot 56$ de la mostra haurien de ser de primer curs, és a dir, 11,20. Adoneu-vos que, tot i que sigui ridícul parlar d'una fracció d'una persona, hem d'usar decimals per a assegurar-nos la precisió en els nostres càlculs. Igualment, tenint en compte que 238 de 860 o 0,2767 de la població són de segon curs, $0{,}2767 \cdot 56 = 15{,}50$ haurien de ser alumnes de segon curs en la mostra. Les altres freqüències esperades es calculen d'una manera similar (columna E_i):

Categoria	E_i	O_i	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
1r curs	11,20	7	-4,20	1,575
2n curs	15,50	8	-7,50	3,629
3r curs	21,55	30	8,45	3,313
4t curs	7,75	11	3,25	1,363
Total	**56,00**	**56**	**0,00**	**9,880**

A la taula C de l'annex 3...

... aquest valor és 7,82, així que aquí tenim l'evidència que les freqüències observades no són representatives de la població. Les diferències entre les freqüències observades i les esperades de la distribució coneguda en la població són massa grans per a ser casuals. Per dir-ho d'una altra manera, si prenem mostres aleatòries de 56 alumnes repetidament d'una població de 860, llavors serà molt poc probable obtenir un conjunt de freqüències observades tan diferent de les esperades.

La resta de càlculs són com els anteriors. Calculem la diferència entre les freqüències observades i les esperades, i aleshores aquestes diferències es posen

al quadrat i es divideixen per les freqüències esperades respectives. La suma d'aquestes últimes quantitats ens dóna una mesura de la diferència global entre les freqüències observades i les esperades.

Per acabar, hem de decidir si el valor de 9,880 és un valor inusualment alt o no ho és. Això s'aconsegueix comparant-lo amb el punt crític d'una distribució χ^2 amb 3 graus de llibertat, un menys que el nombre de categories.

Finalment necessitem interpretar les diferències entre les freqüències observades i les esperades. Podem veure que van venir més alumnes de tercer i quart curs dels previstos al concert. Els organitzadors van concloure que l'esdeveniment era més atractiu per als alumnes més grans, d'acord amb una enquesta independent de preferències musicals entre els alumnes segons la qual als alumnes de segon cicle els agrada més el jazz, mentre que els alumnes més joves prefereixen la música pop i el rock.

Activitats

2. En la mateixa població anterior de 860 alumnes, hi ha 350 homes i 510 dones. En la nostra mostra aleatòria de 56 alumnes, hi ha 25 homes i 31 dones. La mostra és representativa dels sexes?

3. Hi ha una creença popular que diu que hi ha més possibilitats de guanyar la loteria si el número acaba en 7. Suposem que en els últims 36 sortejos de la loteria, 10 dels números guanyadors acabaven en 7. Això confirma aquesta creença?

4. Segons informació de cens sabeu que la distribució d'edat en una certa població és de la manera següent:
fins a 18 anys: 16,5%
18-35 anys: 18,2%
36-49 anys: 30,0%
49-69 anys: 25,5%
+70 anys : 9,8%

Esteu dirigint un estudi sobre la popularitat de diferents tipus de mitjans de comunicació (televisió, ràdio, diaris, etc.) i obteniu dades d'una enquesta dirigida per una agència de publicitat. L'enquesta implica una mostra a escala 1045, suposadament representativa de la població, i les dades inclouen l'edat de cada un dels enquestats. Ara calculeu la distribució d'edat en la mostra i obteniu el següent:
fins a 18 anys: 207
18-35 anys: 259
36-49 anys: 305
49-69 anys: 188
+70 anys: 76

Podeu concloure que la mostra és representativa dels grups d'edat?

Comentari

Quan les mostres tenen grups que no es corresponen als previstos de la població, hi ha maneres de corregir aquesta manca de representativitat en els nostres càlculs. Aquest és un tema de teoria estadística més avançada de què no tractem en aquesta assignatura. Tanmateix, sempre que sigui possible és preferible assegurar una representativitat en un estudi empíric, per exemple com es descriu anteriorment en l'apartat 8, dissenyant l'esquema de mostreig de manera que els grups tinguin una representació proporcional a les seves freqüències de població.

16. Taules encreuades: associació entre dues variables categòriques

En les ciències socials sovint volem mesurar la correlació entre dues variables discretes. Per exemple, hem demanat a una sèrie de persones què opinaven de l'avortament, si hi estaven a favor o en contra. També teníem dades sobre els seus grups d'edat, per exemple, 18-25 anys, 26-35 anys, 36-45 anys, etcètera. Com podem mesurar si les dues variables estan associades?

Això no és més que una variació del mètode χ^2 tractat a l'apartat 15. Proposem un model simple per a les observacions en el qual no hi ha cap associació i a continuació calculem les freqüències esperades en aquest cas. Comparant les freqüències observades amb les esperades, podem mesurar la relació entre les variables i jutjar si és alta o no ho és en un sentit estadístic.

En aquest apartat sobre associació entre dues variables discretes aprendreu a:

- usar un model d'independència entre dues variables discretes;
- calcular freqüències esperades segons aquest model;
- calcular l'estadística d'associació χ^2;
- jutjar el valor d'aquesta estadística χ^2.

Independència entre dues variables discretes

Comencem amb un exemple molt simple. Hem demanat a una mostra de 200 persones si estaven a favor de legalitzar la pena de mort. La mostra consta de 100 homes i 100 dones, i 70 dels homes i 50 de les dones estan a favor de la pena de mort.

Les dades es poden classificar de la manera següent:

Taula I

Dues taules encreuades			
Sexe	**Posició**		
	A favor	**En contra**	**Total**
Home	70	30	100
Dona	50	50	100
Total	**120**	**80**	**200**

Aquest tipus de taula s'anomena **taula encreuada** de les variables "sexe" i "posició". Sol tenir més de dues files i dues columnes.

Sembla que hi ha una freqüència més alta d'homes a favor de la pena de mort que de dones, però volem investigar amb més cura aquest descobriment, ja que sabem que podria ser que aquests resultats particulars s'haguessin donat per casualitat (recordeu l'experiment del llançament de monedes i la variació possible en el nombre de cares que podíem obtenir). Intentarem mesurar l'associació entre les dues variables amb un sol nombre. Primer, necessitem considerar què entenem per *absència d'associació* o què anomenem *independència*.

En tota la mostra veiem que 120 dels 200 enquestats estan a favor de la pena de mort, és a dir el 60%. Si no hi hagués diferències entre els enquestats homes i dones, suposaríem que el 60% dels homes i el 60% de les dones hi han estat a favor. Això ens dóna la clau del que significa independència en aquest exemple: la **independència** és la manca de diferència en percentatges de resposta entre grups.

Quines serien les freqüències si el model d'independència fos veritat? Suposaríem que 60 homes i 60 dones hi han estat a favor, i 40 homes i 40 dones hi han estat en contra. Ara es tracta només de fer una llista d'aquestes freqüències de la mateixa manera que abans, comparar-les amb les observacions reals com abans i calcular l'estadística χ^2 com abans.

Taula II

Categoria	E_i	O_i	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
Homes a favor	60	70	10	1,667
Dones a favor	60	50	-10	1,667
Homes en contra	40	30	-10	2,500
Dones en contra	40	50	10	2,500
Total	**200**	**200**	**0**	**8,334**

Hi ha dues diferències importants entre aquest càlcul i el que hem fet a l'apartat 15:

1) la manera de calcular les freqüències esperades;
2) els graus de llibertat de l'estadística χ^2.

Càlcul de freqüències esperades

No calculem les freqüències esperades segons la informació donada sobre la població, sinó segons els totals de columnes i files de les taules encreuades.

Recordeu...

... com vam obtenir les freqüències esperades: vam agafar els totals de les columnes, 120 i 80, i vam calcular el percentatge previst d'enquestats a favor i en contra. Després, vam multiplicar aquests percentatges pels totals de files de la taula per obtenir les freqüències esperades en cada fila de la taula.

Observeu que no hi ha cap diferència si definiu la independència com la manca de diferència entre les files o entre les columnes. Això ho demostrarem mi-

rant la taula I a l'inrevés. Tenim 100 homes i 100 dones. Si no hi ha cap diferència entre les posicions a favor i en contra de la pena de mort, la meitat d'homes i dones hi estarien previsiblement a favor i l'altra meitat, en contra. Això donaria una freqüència esperada dels que estan a favor de la pena de mort de 60 homes i 60 dones, i de 40 homes i 40 dones en contra. Aquestes freqüències esperades són exactament les mateixes que hem calculat abans.

Hi ha una manera de calcular les freqüències esperades que és vàlida per a totes les taules encreuades: per a cada cel·la de la taula multipliqueu els totals de files i columnes per la cel·la esmentada, i dividiu-los pel total de la taula. Usant la notació R_i per al total de la fila i, C_j per al total de la columna j i n per al total de tota la taula, la fórmula per al valor previst en la fila i i la columna j és:

freqüència esperada: $E_{ij} = \dfrac{R_i \cdot C_j}{n}$.

Graus de llibertat

Els graus de llibertat es calculen d'una manera diferent: el nombre de files menys 1 multiplicat pel nombre de columnes menys 1.

Així, en aquest exemple, els graus de llibertat no són $4 - 1 = 3$, sinó $(2 - 1) \cdot (2 - 1) = 1$. L'estadística χ^2 torna a mesurar la diferència de les freqüències amb el que vam preveure. Tenint en compte que el que preveiem és en aquest cas la independència entre les dues variables, el valor χ^2, 8,334, és la mesura de la manca d'independència o associació entre les dues variables. Ara comparem això amb el valor crític de la distribució χ^2 amb un grau de llibertat, 3,84, i concloem que les nostres dades mostren una associació significativa entre sexe i posició envers la pena de mort.

En altres paraules, les diferències entre homes i dones són més altes de les que hauríem esperat si el model d'independència fos veritat. Per tant, concloem que el model d'independència no és correcte i que hi ha diferències.

Una taula encreuada més gran

Podem aplicar el mètode anterior a una taula encreuada que tingui qualsevol nombre de files i columnes. Per exemple, en una enquesta de 312 lectors d'un diari determinat, es classificava els lectors segons el grau de minuciositat de la seva lectura: "ullada ràpida" al diari, "bastant minuciosament" o "molt minuciosament". També tenim dades sobre l'educació de cada enquestat: "estudis primaris complets", "primer grau d'ensenyament secundari", "ensenyament secundari complet", "estudis universitaris parcials" i "estudis universitaris complets". Així tenim una altra vegada dues variables

discretes amb 3 i 5 categories cadascuna. A continuació podeu veure la taula encreuada dels 312 lectors:

	Ullada	Bastant minuciosament	Molt minuciosament	Total
Primaris	5	7	2	14
Secundaris I	18	46	20	84
Secundaris	19	29	39	87
Universitaris I	12	40	49	101
Universitaris	3	7	16	26
Total	**57**	**129**	**126**	**312**

Comencem formulant el model per a la independència de la taula, que consisteix a dir que no hi ha associació entre el grup educatiu i el nivell de lectura. Si considerem la taula com un conjunt de files, llavors la independència significarà que no hi ha cap diferència entre els grups educatius en funció dels seus percentatges de grups de lectors diferents, és a dir, cada fila de la taula hauria de ser proporcionada amb els valors globals de 57/312 = 0,183, 129/312 = 0,413 i 126/312 = 0,404 calculats a partir dels totals de columna. Per exemple, en l'última fila de la taula que correspon al grup amb estudis universitaris complets hi ha 26 lectors, i això donaria 0,183 · 26 = 4,76, 0,413 · 26 = 10,74 i 0,404 · 26 = 10,50 a les tres columnes d'aquesta fila.

Si fem la llista completa de les freqüències, tindrem 15 freqüències esperades i 15 d'observades, però no les mostrem totes, només les tres últimes corresponents a les tres freqüències esperades que acabem de calcular:

Categoria	E_i	O_i	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
...	...	...	...	...
Universitaris, ullada	4,76	3	-1,76	0,65
Universitaris, minuciosament	10,74	7	-3,74	1,30
Universitaris, molt minuciosament	10,50	16	5,50	2,88
Total	**312,00**	**312**	**0,00**	**26,00**

Per a calcular...

... automàticament les freqüències esperades podem usar la fórmula

$$E_{ij} = \frac{R_i \cdot C_j}{n},$$

per exemple per a "Universitari, ullada":
(26 · 57) / 312 = 4,76.

L'estadística χ^2 calculada sumant els 15 valors és igual a 26,0. Ara comparem aquest valor amb el valor crític d'una distribució χ^2 amb (5 2 1) · (3 2 1) = 8 graus de llibertat, que és 15,51. El valor és més gran que el valor crític, així es fa patent l'associació entre les dues variables, i podem concloure que hi ha diferències entre els grups educatius.

Veiem com podríem fer els càlculs utilitzant el MacAnova. Primer llegim les dades en un vector i a continuació el convertim en una matriu amb 5 files i 3 columnes, usant la funció `matrix()`. El valor 5 de l'ordre `matrix` és el nombre de files de la taula:

```
Cmd> data <- vector(5,18,19,12,3,7,46,29,40,7,2,20,39,49,16)

Cmd> table <- matrix(data,5)

Cmd> table
(1,1)            5            7            2
(2,1)           18           46           20
(3,1)           19           29           39
(4,1)           12           40           49
(5,1)            3            7           16
```

Podeu usar el MacAnova per a calcular els totals de files i columnes, però aquí simplement els llegim en dos vectors, `rsum` i `csum`:

```
Cmd> rsum <- vector(14,84,87,101,26)

Cmd> csum <- vector(57,129,126)
```

La instrucció més complicada és per a calcular els valors esperats:

```
Cmd> esp <- rsum %*% t(csum) / 312

Cmd> esp

(1,1)       2.5577       5.7885       5.6538
(2,1)       15.346       34.731       33.923
(3,1)       15.894       35.971       35.135
(4,1)       18.452        41.76       40.788
(5,1)         4.75        10.75         10.5
```

Comentari

Quan compareu l'estadística χ^2 amb el punt crític de la distribució χ^2, hi ha una presumpció inherent que cap de les freqüències esperades no sigui gaire petita. Per nosaltres, molt petites vol dir 'aproximadament 5'. Si mireu les freqüències esperades anteriors (en la taula `esp`), veureu que dues freqüències en la primera columna –2,56 i 4,75– són inferiors a 5. Això no és seriós en aquest cas, tenint en compte que només hi ha dos valors de 15 que violen la presumpció, i un valor, en tot cas, bastant a prop de 5. Hauríeu de tenir cura per a no fer càlculs χ^2 quan hi ha molts valors previstos inferiors a 5.

Aquí hem fet una multiplicació del vector de sumes de files, que té 5 files i 1 columna, per la transposició del vector de sumes de columnes, que és la suma de columna com a vector de fila amb 3 columnes. Per a efectuar aquesta multiplicació, hem d'usar l'ordre de multiplicació de matrius `FONT` de MacAnova, que és l'operador de multiplicació més corrent entre dos signes de percentatge `%*%`.

La funció `t()` fa la transposició del vector, és a dir, converteix el vector de columna en un vector de fila. Finalment, dividim el total de la taula per 312.

Calcular diferències entre les freqüències observades en `table` i les freqüències esperades en `esp` resulta ara bastant senzill:

```
Cmd> chisq <- (table - esp)^2 / esp

Cmd> chisq

(1,1)       2.3321      0.25358       2.3613
(2,1)      0.45894       3.6566       5.7145
(3,1)      0.60687        1.351      0.42526
(4,1)        2.256     0.074145       1.6531
(5,1)      0.64474       1.3081        2.881
```

Compareu l'última línia amb els valors que hem calculat abans –hi ha diferències petites perquè el MacAnova està fent els càlculs amb més precisió del que ho fèiem abans.

L'estadística χ^2 és la suma de tots els valors de `chisq`. Si apliqueu la funció `sum()` a aquesta matriu, obteniu les sumes de columnes:

```
Cmd> sum(chisq)
(1,1)       6.2987       6.6434       13.035
```

Per a sumar tots els elements en `chisq`, primer hauríeu de convertir la matriu de nou en un vector i aleshores fer la suma:

```
Cmd> chisq <- vector(chisq)
Cmd> chisq
(1)  2.3321   0.45894  0.60687     2.256  0.64474
(6) 0.25358    3.6566    1.351   0.074145   1.3081
(11) 2.3613    5.7145  0.42526    1.6531    2.881

Cmd> sum(chisq)
(1)      25.977
```

Observeu com hem utilitzat el nom `chisq` dues vegades en la primera afirmació: teníem `chisq` com a matriu, llavors l'hem transformat en un vector i l'hem tornat a guardar com a `chisq`.

Abans hem mostrat cada pas detalladament. Podem reduir el nombre d'afirmacions a les dues línies següents (suposant que tenim les dades en una matriu anomenada `table`, les sumes de files i columnes en `rsum` i `csum`, i el total de la matriu en `n`):

```
Cmd> esp <- rsum %*% t(csum) / n

Cmd> sum(vector((table-esp)^2/esp))
(1)       25.977
```

Què s'ha de fer si hi ha valors previstos tan petits? Hauríeu de combinar algunes categories per a eliminar el problema.

Per exemple, en el càlcul dels valors esperats, per a eliminar el valor previst petit de 2,56 en la primera fila podríeu canviar la taula original, per a tenir 4 files en comptes de 5 on les primeres dues files van juntes. És a dir, per calcular el χ^2 per a mesurar l'associació no distingiu entre els primers dos grups educatius.

Estadística χ^2 quan les variables tenen només dos grups

En l'apartat 5 hem vist el coeficient de correlació entre dues variables discretes quan cadascuna té només dues categories. Hi ha una relació entre el que hem fet abans i el que estem fent en aquest apartat. En el nostre primer exemple en aquest apartat, sobre les posicions sobre la pena de mort, hem calculat una estadística χ^2 per a l'associació entre sexe i posició de 8,334, basada en una mostra a escala $n = 200$.

Podem obtenir el coeficient de correlació de l'estadística χ^2 de la manera següent: dividiu l'estadística χ^2 per n i apliqueu-hi l'arrel quadrada –aquesta és la correlació.

D'aquesta manera la correlació entre sexe i posició en el nostre exemple és:

$$\sqrt{\frac{8,334}{200}} = 0,204.$$

Observeu que no distingirem entre una correlació positiva i una de negativa. En aquest cas, només es pot mesurar la correlació positiva ja que no hi ha classificació en les categories de sexe o posició.

Tornem a resumir el que entenem per correlació entre les dues variables de sexe i posició.

Suposem que codifiquem les dues variables de la manera següent:

Sexe	Posició
Home = 0	A favor = 1
Dona = 0	En contra = 0

Els 200 enquestats tenen 4 tipus diferents de respostes: 70 són homes a favor, és a dir amb valors 1 i 1; 30 són homes en contra, és a dir amb valors 1 i 0; 50 són dones a favor, és a dir amb valors 0 i 1, i 50 són dones en contra, és a dir amb valors 0 i 0. Podríem definir tots aquests valors en dos vectors utilitzant el MacAnova de la manera següent:

```
Cmd> sex <- vector(rep(1,100),rep(0,100))

Cmd> att <- vector(rep(1,70),rep(0,30),rep(1,50),rep(0,50))

Cmd> cor(sex,att)
(1,1)            1      0.20412
(2,1)      0.20412            1
```

- La primera afirmació estableix els 200 valors de sexe: 100 uns per als homes i 100 zeros per a les dones –la funció del MacAnova `rep(x,n)` és un vector amb valors `x` repetits `n` vegades.
- La segona afirmació estableix els 200 valors per a la posició: 70 uns per als homes a favor i 30 zeros per als homes en contra, i 50 uns i 50 zeros per a les dones a favor i en contra.
- La tercera afirmació calcula la correlació, i veiem el valor 0,204, que està d'acord amb el nostre càlcul basat en l'estadística χ^2.

L'elecció del sistema de codificació 0/1 per a les variables no afecta gens el càlcul de la correlació, excepte el signe de la correlació (podeu provar-ho intentant donar qualsevol valor que no sigui 0 ni 1 i veureu que obtindreu sempre el mateix coeficient de correlació, possiblement amb un signe negatiu que hauríeu d'ignorar).

Activitats

1. La taula següent mostra la taula encreuada de 88 persones segons dues variables, la música que prefereixen i el seu grup d'edat:

Grup d'edat	Preferència musical				Total
	Jazz	Pop/Rock	Llatina	Clàssica	
21-35 anys	5	16	0	5	26
36-49 anys	16	4	10	2	32
50 anys o més	3	3	8	16	30
Total	**24**	**23**	**18**	**23**	**88**

Hi ha alguna prova d'associació entre edat i preferència musical?

17. Relacions entre variables: observació, experimentació i causalitat

Els dos mètodes principals de recollida de dades són per mitjà de l'observació i l'experimentació. La majoria de dades es recullen per mitjà de l'observació d'un fenomen tal com s'esdevé, sense cap interferència externa. En aquests casos sovint detectem associacions entre certes variables, les quals ens permeten, per exemple, predir el valor d'una variable a partir del d'una altra. Això no vol dir, però, que el valor d'una variable influeixi directament sobre el d'una altra.

En circumstàncies molt especials podem dirigir un experiment en el qual controlem la unitat experimental. Un experiment així és la manera definitiva de provar una relació causal entre dues variables.

En aquest apartat sobre relacions entre variables aprendreu:

- com es formula un projecte de recerca en gestió ambiental en què l'estadística té un paper capdavanter;
- les diferències entre l'observació i l'experimentació;
- com es duu a terme un experiment comparatiu aleatoritzat;
- la qüestió de la causalitat.

Les dades observables

Les lliçons estadístiques que cal aprendre són universals: el paper interdisciplinari de l'estadística, el problema de l'alta variabilitat en les dades i la importància del modelatge estadístic.

La biometria...

... és una disciplina de la biologia que estudia els fenòmens quantitatius en els éssers vius amb mètodes estadístics. Els treballs de Galton, Pearson, Fisher i altres van establir les bases d'aquesta ciència.

L'experimentació

L'experimentació solament és possible en situacions especials en què es poden controlar certes variables crítiques, les quals són el centre de l'estudi. Si ens fixem en l'experiment en el qual es vol veure si el fet de prendre aspirina regularment redueix el risc d'un atac de cor, s'han de comprovar tots els tractaments mèdics nous per mitjà d'una experimentació acurada i precisa.

L'experiment més simple implica dos grups de subjectes, l'un rebrà el tractament i l'altre rebrà placebo. S'han de seleccionar els dos grups a l'atzar, i no s'hi haurien de barrejar eleccions personals. Això assegura la validesa de l'experiment, de manera que les diferències observades entre els dos grups seran degudes al tractament i no a cap altra raó.

Si els dos grups són diferents en algun aspecte, abans no comencem l'experiment, hi haurà un problema quant a separar l'efecte del tractament d'aquesta diferència; això s'anomena **indistingibilitat**. Per exemple, en l'estudi sobre l'aspirina, suposem que el grup que rep tractament tingués un conjunt de subjectes lleugerament més joves. Com que els atacs de cor tenen relació amb l'edat, qualsevol risc menor en el grup que rep tractament es podria explicar per l'aspirina o per les edats inferiors –seríem incapaèos de dir quina és l'explicació; d'aquí sorgeix el terme *indistingibilitat*.

Però el fet que es faci servir l'aleatorització per a decidir quins subjectes van a cada grup assegura que els grups són comparables en totes les variables. I si l'única diferència entre els dos grups és si prenen aspirina o no en prenen, s'elimina la confusió, i qualsevol diferència que hi hagi entre els grups serà deguda a l'aspirina.

L'experimentació en les humanitats

L'experimentació acurada, implicant-hi l'aleatorització per a assignar subjectes a diferents grups experimentals, és un camí segur per a detectar els efectes veritables. Sovint resulta impossible dur a terme experiments de debò, però els principis subjacents a l'experimentació es poden tenir en compte a l'hora de fer investigació.

18. Repàs: de les estimacions puntuals als intervals de confiança

En aquest apartat revisem tot el material que hem donat fins ara:

1) El nostre plantejament ha estat començar amb descripcions simples de la distribució d'un conjunt de dades observades per a una variable numèrica, resumint el centre i la dispersió de la distribució.

2) També hem tractat d'una mesura de correlació entre dues variables numèriques.

3) Hem distingit entre la població de totes les observacions possibles per a la variable i la mostra de valors que observem seleccionant unitats a l'atzar a partir d'una població.

4) Hem vist una distribució important esperada, la distribució normal, la qual sovint es fa servir com a distribució poblacional d'una variable.

5) Després ens hem centrat en els estadístics resum, en particular la mitjana, i hem vist que quan la distribució de la població és normal, una mitjana calculada sobre una mostra aleatòria d'aquesta població també està normalment distribuïda amb la mateixa mitjana que té la població.

6) És més, hem vist que la mitjana mostral té una distribució amb una dispersió menor que la distribució poblacional, i que aquesta esdevé cada cop més petita a mesura que la grandària mostral augmenta. Fins i tot si la distribució poblacional no és normal, el teorema central del límit ens diu que la mitjana d'una mostra té una distribució aproximadament normal, i aquesta aproximació esdevé cada cop més exacta a mesura que la grandària mostral s'incrementa.

7) Hem vist com es pot fer servir aquest coneixement de la distribució de la mitjana per a establir estimacions d'intervals en què confiem que hi ha la veritable mitjana poblacional.

8) Finalment, hem presentat l'estadístic χ^2, del qual fem ús quan comparem observacions discretes amb valors esperats de la població o segons el supòsit d'independència entre dues variables.

En aquest apartat de repàs revisem els conceptes clau quant a:

- la descripció d'una distribució;
- la distribució normal;

- la població i la mostra;
- la distribució mostral de la mitjana;
- els intervals de confiança per a una mitjana i una proporció;
- la utilització del χ^2 per a comparar freqüències observades i esperades.

Descripció d'una distribució

El material bàsic amb el qual treballem en estadística és un conjunt d'observacions sobre una o més variables. En el **cas de variables categòriques**, podem resumir el conjunt d'observacions d'una manera molt simple: comptant el nombre d'observacions de cada categoria i presentant els resultats amb la forma d'un diagrama de barres. En el **cas de les variables numèriques**, primerament podem agrupar les observacions en intervals de la mateixa llargada i dibuixar un histograma de la distribució. Després també podem resumir la distribució calculant diversos valors que indiquen el centre de la distribució i la seva dispersió a partir d'aquest centre.

Repasseu els apartats de l'1 al 4.

De l'apartat 1 al 4 hem tractat d'aquest tema, hem vist com es resumeixen distribucions de qualsevol forma –simètrica o asimètrica– fent ús de la mediana, els quartils i els valors extrems. El primer quartil, la mediana (o segon quartil) i el tercer quartil separen tots els valors observats en quarts i són fàcils d'interpretar, per exemple: un quart de totes les observacions cau sota el primer quartil. Quan la distribució és aproximadament simètrica, aleshores ens refiem solament de dos valors per a resumir-ne el centre i la forma respectivament, la mitjana aritmètica i la desviació estàndard.

Activitats

1. L'arxiu `SAVING` conté un conjunt de dades, en milions de pessetes, dels balanços de les llibretes d'estalvi de 100 clients d'un banc. Dibuixeu un histograma de la distribució. Calculeu els cinc nombres resum per a aquestes dades, i la mitjana i la desviació estàndard, i comenteu la diferència entre la mitjana i la mediana com a mesures del centre de la distribució.

(Una pista: per als cinc nombres resum necessitareu ordenar les dades en ordre ascendent; hi ha una funció del MacAnova anomenada `sort()` que us ho farà. Si introduïu les dades en el vector *x*, aleshores l'ordre `sort(x)` us en farà l'ordenació. Per a obtenir informació sobre aquesta funció, introduïu l'ordre següent dins el MacAnova: `help(sort)`.)

La població i la mostra

Repasseu els apartats 7 i 8.

La població és el conjunt total d'unitats d'interès en un estudi determinat. Solament en casos molt rars farem un estudi exhaustiu, o **cens**, de cada unitat de la població.

En la majoria de casos estudiarem un nombre més petit d'unitats, o **mostra**, amb el propòsit d'arribar a una conclusió sobre la població.

És important que la mostra sigui representativa de la població, i sovint subdividim la població en grups o estrats, i mostregem dins cada grup per assegurar una representació igual. S'ha de seleccionar la mostra aleatòriament, de manera que cada unitat de la població tingui la mateixa oportunitat de ser dins la mostra.

La distribució normal

En moltes situacions les observacions que estudiem, o els estadístics resum que calculem a partir d'aquestes observacions (per exemple: la mitjana), tenen una distribució que és simètrica i amb forma de campana. La distribució ideal per a usar en aquest cas és la distribució normal.

Repasseu els apartats 9 i 10.

La corba de densitat normal queda completament definida per la mitjana i la desviació estàndard. Per tant, convé remetre's sempre a una distribució normal determinada, la distribució normal estàndard, la qual té mitjana 0 i desviació estàndard 1, i després transformar totes les altres distribucions normals en aquesta. Qualsevol observació x extreta d'una distribució normal amb mitjana μ i desviació estàndard σ es pot transformar en una observació normal estàndard restant la mitjana i dividint per la desviació estàndard:

$$\frac{(x-\mu)}{\sigma}.$$

Això s'anomena **estandardització de les dades**.

A partir de les taules, o fent servir un programa estadístic com ara el MacAnova, podeu determinar l'àrea sota la corba normal estàndard entre dos valors, posem per cas z_1 i z_2, i, així mateix, entre dos valors qualssevol d'una corba normal. Aquesta àrea representa la probabilitat que una observació extreta aleatòriament a partir d'aquesta distribució es trobi entre aquests dos valors.

Activitats

2. Suposeu que la distribució de la durada de les trucades telefòniques fetes el diumenge a Barcelona és normal amb una mitjana de 157 segons i una desviació estàndard de 52 segons. Feu servir les taules o el MacAnova per a calcular la probabilitat que una trucada tingui una durada entre 3 i 4 minuts.

La distribució mostral de la mitjana

La mitjana és el valor més important que fem servir per a caracteritzar un conjunt d'observacions. Quan aquestes observacions són una mostra aleatòria d'una distribució normal amb una mitjana μ i una desviació estàndard σ, la mitjana també és normalment distribuïda amb la mateixa mitjana μ i una desviació estàndard més petita, igual a $\sigma / \sqrt{n}$, on n és la grandària mostral. Fins i tot quan la distribució de la població no és normal, el teorema central del límit mostra que la mitjana mostral té una distribució normal aproximada amb

Repasseu els apartats 11 i 12.

una mitjana igual a la mitjana poblacional i una desviació estàndard igual a la desviació estàndard poblacional dividida per $\sqrt{n}$. L'aproximació esdevé més exacta a mesura que la grandària mostral augmenta. La desviació estàndard de la mitjana s'anomena **error estàndard**.

També podem aplicar tots els nostres resultats a la distribució mostral d'una mitjana de dades 0/1, o binàries. Aquesta mitjana és la proporció d'èxits (uns) en la mostra. La desviació estàndard d'una variable binària amb una mitjana poblacional π (és a dir, la proporció d'èxits en la població) és igual a $\sqrt{\pi(1-\pi)}$. Per tant, l'error estàndard de la proporció p d'èxits en una mostra de grandària n és:

$$\sqrt{\frac{\pi(1-\pi)}{n}}.$$

Per a mostres grans, per mitjà del teorema central del límit, la proporció observada p és aproximadament normalment distribuïda.

Activitats

3. Com en l'activitat 2, ara treballem amb una distribució normal amb una mitjana de 157 segons i una desviació estàndard de 52 segons. Quina és la distribució de la mitjana de 1.000 trucades telefòniques preses aleatòriament?

4. En un casino de joc una màquina d'apostes determinada dóna al jugador una probabilitat de victòria de 0,4. El resultat d'una jugada no té cap connexió amb el resultat de la següent. Un jugador juga 200 vegades amb aquesta màquina. Quina és la probabilitat que el jugador guanyi 100 vegades o més?

Els intervals de confiança per a una mitjana i una proporció

Repasseu els apartats 13 i 14.

En l'exemple anterior coneixem la mitjana de la població i mirem la distribució de la mitjana d'una mostra de la població. Si no sabéssim la mitjana poblacional μ, la mitjana mostral seria una bona estimació d'aquella mitjana desconeguda. Ara es posa l'atenció a fer una argumentació més específica sobre què és aquesta μ desconeguda. Avaluem no solament la millor estimació puntual, sinó un interval que conté μ amb un nivell alt de certitud (almenys del 90%; normalment és del 95%). Un interval de confiança és sempre en la forma següent:

$$\text{mitjana} \pm \text{punt crític} \cdot \text{error estàndard de la mitjana},$$

és a dir:

$$\text{mitjana} \pm \text{punt crític} \cdot \frac{\text{desviació estàndard}}{\sqrt{n}},$$

on la mitjana és la mitjana mostral basada en una mostra aleatòria de n observacions, la desviació estàndard és la de la població de la qual hem mostrejat les dades, i el punt crític és el valor apropiat de distribució normal que

inclou (habitualment) el 95% de la probabilitat entre ± el seu valor (en altres paraules, el punt crític talla al 2,5% de la probabilitat de cada cua). Fem servir la notació $z_{\alpha/2}$ per a indicar el punt crític de la distribució normal que inclou el 100 $(1 - \alpha)$% de la probabilitat (per exemple $\alpha = 0{,}05$ per a un interval de confiança del 95%, $z_{0{,}025} = 1{,}96$).

Podem distingir dos casos que hem estudiat per a obtenir intervals de confiança per a la mitjana a un nivell de confiança del 00 $(1 - \alpha)$%:

1) La mostra és de qualsevol grandària i se suposa que prové d'una distribució normal amb una mitjana desconeguda μ però una desviació estàndard coneguda σ (no cal que se suposi la forma de la distribució si la mostra és molt gran). La mitjana mostral dóna un interval de confiança per a μ de:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

2) La mostra consisteix en un conjunt gran d'observacions binàries, codificades com a 0 (fracàs) i 1 (èxit), d'una distribució que té una proporció desconeguda π d'èxits. La mitjana de la mostra és la proporció p d'èxits observats i dóna un interval de confiança per a π de la forma:

$$p \pm z_{\alpha/2} = \sqrt{\frac{p(1-p)}{n}}.$$

Activitats

5. A què fa referència el 95% en l'expressió *interval de confiança del 95%*? Per què no podem tenir intervals de confiança del 100%?

6. Preguntem a una mostra aleatòria de 50 famílies quin pressupost mensual tenen per a despeses mèdiques. La mitjana de les respostes és 8.340 pessetes, i la desviació estàndard de la mostra es calcula i dóna 2.570 pessetes. Calculeu un interval de confiança del 95% per al pressupost mitjà mensual per a les famílies de la població, suposant que els pressupostos són normalment distribuïts.

7. A partir d'una mostra aleatòria de 1.492 adults, es va veure que el 35% estaven a favor d'incrementar el preu de la benzina per a subvencionar les autopistes. Calculeu l'interval de confiança del 95% per al veritable percentatge d'adults de la població que tinguin aquesta opinió.

Freqüències observades i esperades

L'estadística χ^2 és un dels valors calculats i tractats més freqüentment en la recerca de les ciències humanes. Hem introduït el concepte mitjançant una mostra de població que hem dividit en grups: grups d'edats, grups socioeconòmics i grups educatius. A partir del nostre coneixement de la població hem pogut especificar com esperàvem que es distribuís la mostra per aquests grups. La mostra, naturalment, exceptuant que s'hagués seleccionat específicament per a satisfer aquesta distribució esperada, tenia percentatges diferents dels grups. A continuació, s'ha calculat l'estadística χ^2 per a resumir la diferència

global entre les freqüències observades en la mostra i les freqüències esperades. Es calcula de la manera següent:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}.$$

L'estadística χ^2 es compara llavors amb el valor crític de la distribució χ^2, amb graus de llibertat iguals a un menys que el nombre de grups definits en la població i la mostra. Si l'estadística χ^2 és superior a aquest valor, podem deduir llavors que la mostra difereix significativament dels percentatges de població i que no es pot considerar una mostra aleatòria representativa de la població esmentada.

A continuació, hem aplicat el mateix concepte a una situació en què es comparen dues variables discretes mitjançant una taula encreuada. En aquest cas, les freqüències esperades s'han obtingut suposant que no hi ha associació entre les dues variables, és a dir, que són independents. Les freqüències esperades es calculen llavors a partir dels totals de files i columnes de la taula de la manera següent:

$$\text{freqüència esperada: } E_{ij} = \frac{R_i \cdot C_j}{n}$$

i el χ^2 és el mateix d'abans, però aplicat a totes les freqüències de la taula encreuada:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Aquí tornem a comparar l'estadística amb el valor crític de la distribució χ^2, però aquí els graus de llibertat són el producte (nombre de files $-$ 1) $\times$ (nombre de columnes $-$ 1). Si l'estadística χ^2 és superior a aquest valor, llavors és improbable que les dues variables siguin independents –es consideren associades o correlatives.

També hem vist que el coeficient de correlació es podia calcular en una taula encreuada simple que constava de dues files i dues columnes, on s'havia assignat qualsevol valor a les categories de fila i columna, i 0 i 1 són els valors més comuns. El coeficient de correlació en aquest cas especial es relaciona amb l'estadística χ^2 calculada en la taula de la manera següent:

$$\text{correlació} = \sqrt{\frac{\chi^2}{n}}.$$

Activitats

8. Comparem estudiants que van a la universitat amb estudiants que estudien regularment a la UOC (universitat a distància) per saber si estan d'acord que la universitat a distància pot reemplaçar l'ensenyament a la universitat. Els resultats són els següents, amb respostes que poden ser "d'acord que la universitat a distància pot reemplaçar l'ensenyament a la universitat", "indecisos" o "en desacord":

	D'acord	Indecisos	En desacord
UOC	18	4	5
Altres	6	10	2

Podeu concloure que els dos grups d'estudiants tenen opinions significativament diferents?

Solucionari

Activitats

Apartat 1

No hi ha activitats

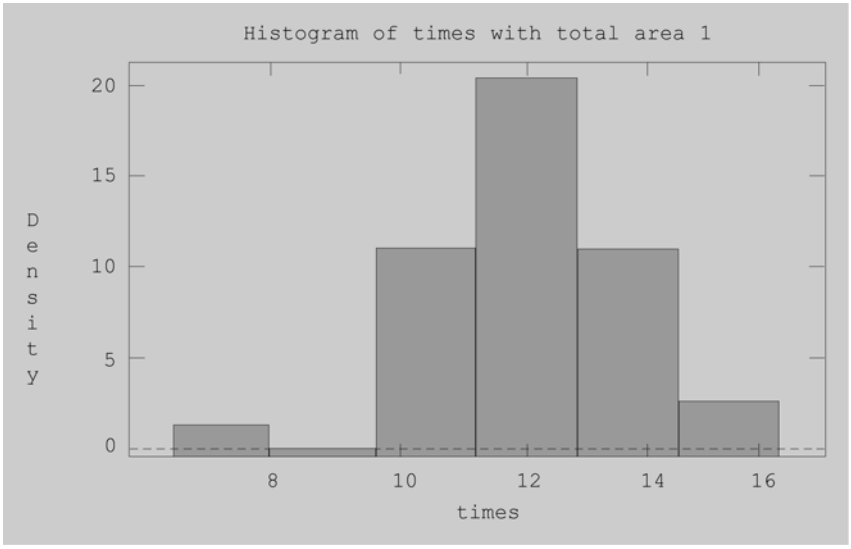
Apartat 2

1.

```
3|
3| 899
4| 0001222222333444
4| 5556677788889
5| 0012223344
5| 57
6| 1223
6| 5
7| 1
7|
```

El centre de la distribució és entre 4 i 5. La distribució és asimètrica i té una cua molt llarga a la dreta (envers els valors més alts).

2. Tenim 45 observacions. Fem un histograma amb unes 6 categories, igualment espaiades, amb l'ajuda d'un ordinador, el qual mostrem més avall:



L'histograma mostra que el centre de la distribució és més o menys 12, i que és simètrica a part d'una observació a l'esquerra separada de la resta de dades (segurament és un error i el valor ha de ser 16,8 i no 6,8).

Apartat 3

1. La mediana dels sous és 4,7 milions. La mitjana aritmètica és 4,83 milions. El fet que la mitjana sigui més gran que la mediana indica que la distribució és una mica asimètrica cap als ingressos més elevats. Això concorda amb les conclusions que hem fet sobre el diagrama de tiges i fulles en l'activitat 1 del segon apartat.

Apartat 4

1. Els cinc nombres resum per als índexs d'atur són 2,2, 5,3, 8,4, 13,5 i 23,5.

2. La mitjana de vida és de 70,44 minuts i la desviació estàndard és de 7,10 minuts.

Apartat 5

1. La correlació és 0,739.

Per a comprovar els vostres resultats:

	Matemàtiques	Estadística
Mitjana	6,74	7,00
Variància	0,38	1,556
Covariància entre matemàtiques i estadística	0,5667	

Apartat 6

1. Les ordres del MacAnova són:

a)

```
Cmd> sqrt(1.77^2 + 3.59^2)
(1)       4.0026
```

b)

```
Cmd> exp(-0.556/2)
(1)       0.7573
```

c)

```
Cmd> log(9.86)
(1)      2.2885
```

d)

```
Cmd> setoptions(angles:"degrees")
Cmd> cos(12.4/7.3)
(1)      0.99956
```

e)

```
Cmd> atan(0.7)
(1)      34.992
```

2.

```
Cmd> notes_stand <- (notes-centr)/sqrt(3.2206)

Cmd> notes_stand
(1)  -0.98802   -2.6318    0.82297   -0.54224   0.73938
(6)   -1.2666  -0.73727    0.54435    0.18216   -1.8239
(11)  0.73938   -0.9323   -0.98802    -1.2945   0.37719
(16) -0.18004 -0.096456     1.1016     1.2687   0.73938
(21)  0.37719 -0.096456    0.26574    0.90655    2.1882
(26) 0.098573   0.71152   -0.73727    -0.9323  -0.62582
(31)   1.1016    1.1852   -0.82085    0.82297    1.1016
(36)   1.4638  -0.65368    0.37719   -0.45865    1.3523
(41) -0.45865 -0.012872   -0.18004    -0.5701   -1.5452
(46)  -1.7124   -1.2666    0.26574    0.82297  -0.45865
(51) -0.29149   0.62794   0.098573     1.4638  0.098573
(56) -0.90443  -0.45865    -2.3254    0.99014   -2.1025
(61) -0.18004   0.73938  -0.096456     1.6309   0.73938
(66) -0.18004    1.4638   0.014989   -0.18004   0.82297
(71)  0.99014   -1.6567   -0.34721   0.014989    1.4638
(76) -0.54224    0.18216    0.37719

Cmd> describe(notes_stand)
component: n
(1)           78
component: min
(1)      -2.6318
```

```
component: q1
(1)     -0.62582
component: median
(1)     0.014989
component: q3
(1)      0.73938
component: max
(1)       2.1882
component: mean
(1)  -1.2859e-05
component: var
(1)      0.99999
```

Apartat 7

1. Els tres nombres aleatoris següents, els quals es troben entre 1 i 2.150, obtinguts a partir dels nombres aleatoris de l'annex 2, són 192, 454 i 9.

2.

```
Cmd> sample <- ceiling(runi(15)*92)
NOTE: random number seeds set to 1059509445 and 1907259950

Cmd> sample
 (1)       68       26       77       51       36
 (6)       20       26        1       64        8
(11)       63       16        8       40       45
```

(Observeu un cop més que el conjunt de nombres aleatoris que obteniu seran diferents dels de més amunt, ja que són aleatoris.)

3.

```
Cmd> rand <- runi(10)
Cmd> rand
(1)   0.92516   0.33314   0.56191   0.26256   0.16652
(6)   0.72102   0.59771   0.07504   0.89391   0.15179

Cmd> ind <- ceiling(rand*60)

Cmd> ind
(1)       56       20       34       16       10
(6)       44       36        5       54       10

Cmd> iq <- vecread("IQ")

Cmd> iq
 (1)      120      101      118      116      108
 (6)       96      110      102      115      103
(11)       91       88      107       94      104
(16)       97       95      101      103      105
(21)      100       94      124       90      106
(26)      107      106       98       96      100
(31)       87      112       95      106      103
(36)       89      119       96       90      104
(41)      105      125      110       98      102
(46)      108       98      131       85      104
(51)       93       93       94       87       97
(56)      100       92       89      100       96

Cmd> sample <- iq[ind]

Cmd> sample
(1)      100      105      106       97      103
(6)       98       89      108       87      103
```

Apartat 8

1. Això dependrà de les coses que l'estudi vulgui mesurar.

Si no podem distingir certs grups dins la població –com ara casat/solter, amb feina/sense feina, home/dona– que puguin respondre d'una manera diferent a les preguntes de l'estudi, aleshores es pot aplicar un esquema de mostreig aleatori simple a una llista de tots els estudiants. Es podria usar un disseny de mostreig probabilístic sistemàtic, ja que la llista dels estudiants tindria algun tipus d'ordre que no és pertinent quant a l'estudi, per exemple l'ordre alfabètic.

En el cas del curs estadístic, el fet d'aprovar o suspendre el curs, i també potser el sexe, pot ser que afecti el nivell de satisfacció. Si aquesta informació és a l'abast, podríem estratificar la població, prenent una mostra que amb seguretat sigui representativa de la població en aquests dos aspectes.

Apartat 9

(No hi ha activitats)

Apartat 10

1. Valors estandarditzats:
$-3{,}44\ -2{,}89\ -2{,}33\ -1{,}78\ -1{,}22\ -0{,}67\ -0{,}11\ \ 0{,}44\ \ 1{,}00\ \ 1{,}56\ \ 2{,}11$

a) 0,0019.
b) $1 - 0{,}9826 = 0{,}0174$.
c) $0{,}4562 - 0{,}0375 = 0{,}4187$.

(Observeu que la primera àrea és la que arriba i inclou 6 dies, i la segona àrea arriba i inclou 3 dies: veureu que aquí hi ha un problema perquè fem servir una distribució contínua per a descriure dades que són únicament enters.)

2.

```
Cmd> cumnor(1.0)-cumnor(-1.0)
(1)      0.68269

Cmd> cumnor(2.0)-cumnor(-2.0)
(1)       0.9545

Cmd> cumnor(3.0)-cumnor(-3.0)
(1)       0.9973
```

3.

```
Cmd> invnor(0.1)
(1)      -1.2816

Cmd> invnor(0.01)
(1)      -2.3263

Cmd> invnor(0.001)
(1)      -3.0902
```

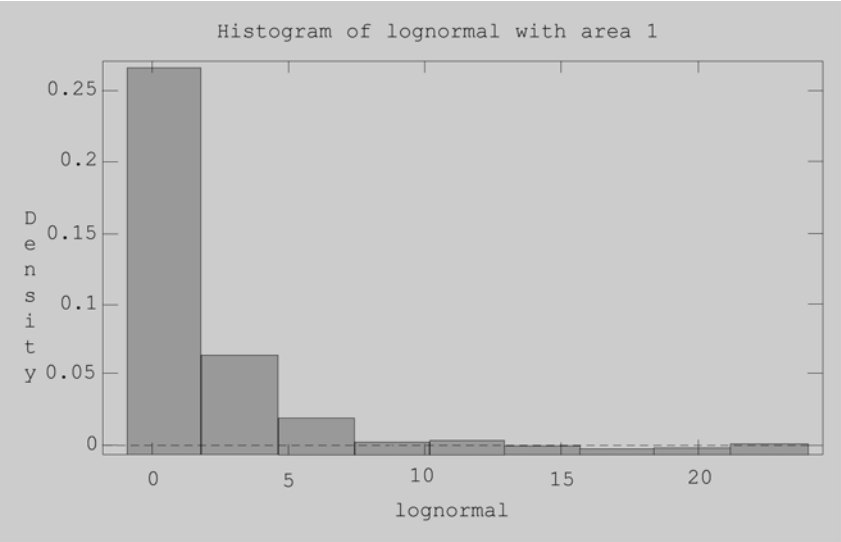
Apartat 11

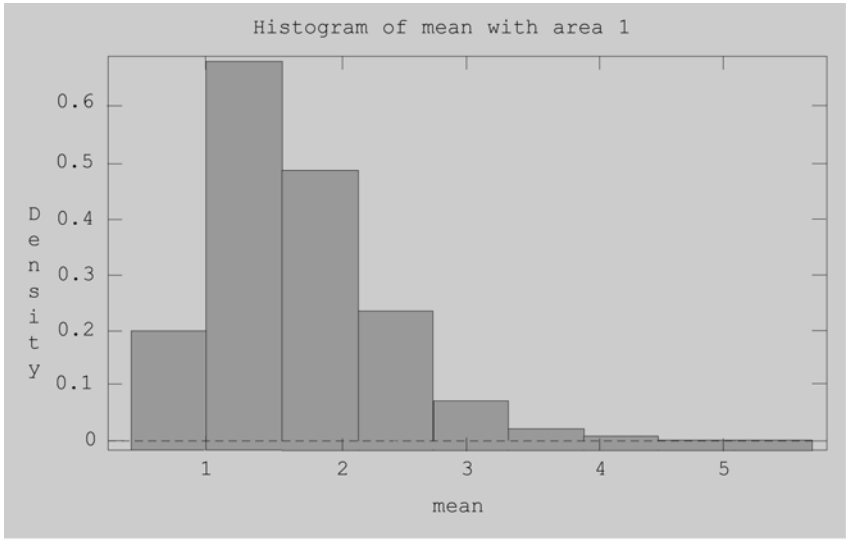
1. (Observeu un cop més que els resultats següents seran diferents dels que obtindreu, ja que es basen en un conjunt de nombres aleatoris diferents.)

```
Cmd> normal <- rnorm(400)
NOTE: random number seeds set to 1169637035 and 377753051

Cmd> lognormal <- exp(normal)

Cmd> hist(lognormal)

Cmd> batch("CLT10.MAC")
CLT10.MAC> sample<-exp(rnorm(10)); mean<-sum(sample)/10
CLT10.MAC> for(i,run(399)){
sample<-exp(rnorm(10));
  mean<-cat(mean,sum(sample)/10);;
}
```

```
CLT10.MAC> (end of file on CLT10.MAC)

Cmd> hist(mean)
```

```
Cmd> describe(mean)
component: n
(1)          400
component: min
(1)       0.6439
component: q1
(1)        1.208
component: median
(1)       1.5417
component: q3
(1)       2.0365
component: max
(1)        5.444
component: mean
(1)       1.6751

component: var
(1)      0.45526

Cmd> sqrt(0.45526)
(1)      0.67473

Cmd> batch("CLT50.MAC")
CLT50.MAC> sample<-exp(rnorm(50)); mean<-sum(sample)/50
CLT50.MAC> for(i,run(399)){
  sample<-exp(rnorm(50));
  mean<-cat(mean,sum(sample)/50);;
}
clt50.mac> (end of file on CLT50.MAC)

Cmd> hist(mean)
```






```
Cmd> describe(mean)
component: n
(1)          400

component: min
(1)        1.024
component: q1
(1)       1.4673
component: median
(1)        1.635
component: q3
(1)       1.8346
component: max
(1)        2.706
component: mean
(1)       1.6698
component: var
(1)      0.08116

Cmd> sqrt(0.08116)
(1)      0.28489
```

La desviació estàndard de la mitjana aritmètica de les mostres de grandària 50 és molt més petita. Incrementar la grandària mostral per un factor 5 redueix l'error estàndard per un factor $1/\sqrt{5}$; això és al voltant de 0,45. Per a les simulacions aquest factor és 0,28/0,67, o al voltant de 0,42, no gaire lluny del valor teòric (un cop més, la vostra pròpia execució d'aquestes ordres donarà un resultat diferent, però les vostres desviacions estàndard s'haurien d'acostar a les que hem calculat més amunt, i la vostra raó hauria de ser a prop de 0,45).

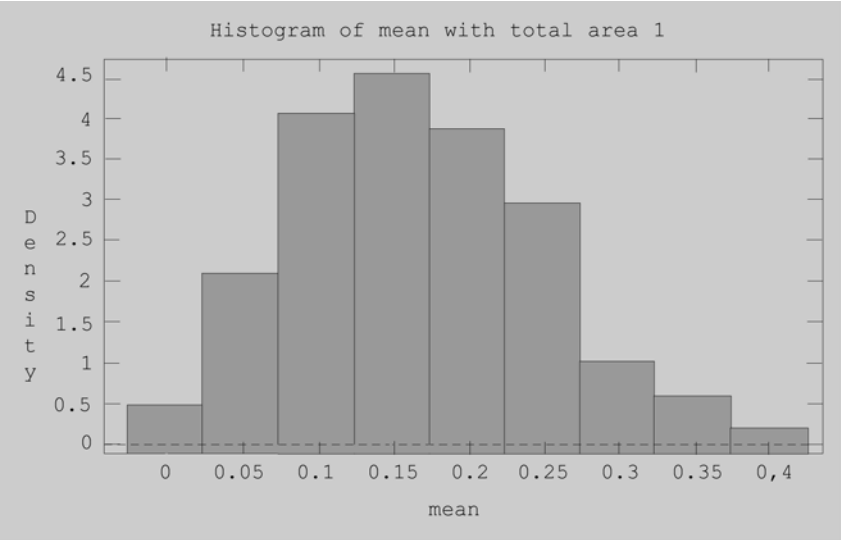
Apartat 12

1. Aquesta variable binària pren el valor 0 amb la probabilitat 0,5 i el valor 1 amb la probabilitat 0,5. La mitjana aritmètica és $(0 \cdot 0{,}5) + (1 \cdot 0{,}5) = 0{,}5$. La desviació estàndard és:

$$\sqrt{0{,}5 \cdot (1 - 0{,}5)} = 0{,}5.$$

2.

```
Cmd> n <- 20 ; pi <- 1/6

Cmd> batch("PROPN.MAC",echo:F)
Mean of 400 proportions
(1)      0.16738
Standard deviation of 400 proportions
(1)     0.083872

Cmd> hist(mean)
```



La distribució és lleugerament asimètrica, amb el centre al voltant de l'esperat 1/6.

3. El nombre de naixements masculins: $0,48 \cdot 50 = 24$.

La desviació estàndard de la proporció mostral és: $\sqrt{0,48 \cdot (1-0,48)/50} = 0,071$. Això correspon a una desviació estàndard de 3,5 expressada com a nombre de naixements sobre 50 = 3,5.

4.

```
Cmd> n <- 50 ; pi <- 0.48

Cmd> batch("PROPN.MAC",echo:F)
Mean of 400 proportions
(1)      0.48365
Standard deviation of 400 proportions
(1)     0.068611
```

La mitjana calculada i la desviació estàndard són a prop dels valors teòrics de 0,48 i 0,071 respectivament.

Apartat 13

1. S'espera que trobareu una xifra al voltant de 5, ja que l'interval de confiança és del 95%.

2. L'error estàndard de la mitjana és: $0,34 / \sqrt{50} = 0,048$.
a) L'interval de confiança del 95% és:
$1,89 \pm z_{0,025} \cdot 0,048 = 1,89 \pm 1,96 \cdot 0,048 = 1,89 \pm 0,094 = [1,796; 1,984]$.
b) L'interval de confiança del 99% és:
$1,89 \pm z_{0,005} \cdot 0,048 = 1,89 \pm 2,576 \cdot 0,048 = 1,89 \pm 0,124 = [1,766; 2,014]$.

3. L'error estàndard de la mitjana és $13/\sqrt{4} = 6,5$ unitats. L'interval de confiança del 95% per a la mitjana veritable és $46 \pm 2 \cdot 6,5 = 46 \pm 13 = [33, 59]$.

4. El marge d'error màxim és 5. L'error estàndard màxim permès, doncs, és 5/1,96 = 2,55.
Com que l'error estàndard (*ES*) = la desviació estàndard (*DS*) / $\sqrt{n}$, obtenim:

$$\sqrt{n} = \frac{DS}{ES}.$$

D'aquesta manera en aquest cas: $n = (13/2,55)^2 = 26$; per tant, cal prendre 26 mesures independents.

5. Una precisió de 0,5 segons és el mateix que dir que hi ha un marge d'error de 0,5. Això vol dir que l'error estàndard és 0,5/1,96 = 0,255.
Com que $\sqrt{n} = DS / ES$, obtenim $\sqrt{n} = 2/0,255 = 7,84$, per tant $n = 61,5$.
Per a obtenir la precisió que cal necessitem mostres d'almenys 62 unitats.
L'error estàndard de la mitjana mostral 12,4 és: $DS = ES / \sqrt{n} = 2 / \sqrt{10} = 0,632$.
Un interval de confiança del 95% per a la mitjana veritable és: $12,4 \pm 1,96 \cdot 0,632 = =$ [11,16; 13,63].

Els 10 segons de temps de resposta es troben fora d'aquest interval de confiança. Això se suposa que passa solament en un 5% de les vegades; per tant, és poc probable. Hem de concloure que és més probable que el temps de resposta hagi augmentat.

Apartat 14

1.
a) L'error estàndard és $\sqrt{0,1 \cdot 0,9/125} = 0,0268$. Per tant, l'interval de confiança del 95% és $0,1 \pm 1,96 \cdot 0,0268 = 0,1 \pm 0,0525 = [0,047; 0,153]$.

b) L'error estàndard és $\sqrt{0,1 \cdot 0,9 / 500} = 0,0134$. Per tant, l'interval de confiança del 95% és $0,1 \pm 1,96 \cdot 0,0134 = [0,074; 0,126]$.

2. L'error estàndard de les respostes "sí": $\sqrt{0,315 \cdot 0,685 / 1.500} = 0,0120$.

L'error estàndard de les respostes "no": $\sqrt{0,645 \cdot 0,355 / 1.500} = 0,0124$.
L'interval de confiança:
"sí": $0,315 \pm 0,024 = [0,291; 0,339]$;
"no": $0,645 \pm 0,024 = [0,621; 0,669]$.

3. La proporció de votants de CiU: $615/1.000 = 0,615$.

L'error estàndard: $\sqrt{0,615 \cdot 0,385 / 1.000} = 0,0124$.
L'interval de confiança del 95%: $0,615 \pm 0,030 = [0,585; 0,645]$.
Com que el 50% és de bon tros fora d'aquest interval de confiança, conclouríem que CiU pot donar la victòria per segura en un 95%.

4. $n = z^2_{\alpha/2} \frac{p(1-p)}{(\text{marge d'error})^2}$; $n = (1,64)^2 \frac{0,5 \cdot 0,5}{(0,05)^2} = 269$

5. $n = (1,96)^2 \frac{0,5 \cdot 0,5}{(0,01)^2} = 9.604$.

6. $n = (1,96)^2 \frac{0,5 \cdot 0,5}{(0,025)^2} = 1.537$.

Apartat 15

1. Dins els 100 experiments d'aquest tipus efectuats, n'hi va haver 98 on el nombre de cares era de 40 a 60 i només 2 quedaven fora d'aquests límits (un amb 61 cares i un altre amb 39 cares). Recordeu que en el vostre cas els resultats podrien ser diferents, però hi havia molt pocs experiments on el nombre de cares superés aquests límits.

2. Els valors previstos serien $(350/860) \cdot 56 = 22,79$ i $(510/860) \cdot 56 = 33,21$ respectivament. Aleshores, els càlculs són els següents:

Categoria	E_i	O_i	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
Home	22,79	25	-2,21	0,2143
Dona	33,21	31	2,21	0,1471
Total	**56,00**	**56**	**0,00**	**0,3614**

L'estadística χ^2 de 0,3614 és molt inferior al valor crític de 3,84 per a una distribució χ^2 amb un grau de llibertat, per això concloem que la mostra és representativa dels sexes.

3. La nostra presumpció seria que cada nombre del 0 al 9 té la mateixa possibilitat d'aparèixer al final d'un nombre guanyador, de manera que la previsió seria que una dècima part dels nombres guanyadors podrien acabar en 7. D'una mostra de 36 nombres guanyadors, el nombre previst de nombres acabats en 7 seria 3,6, i el nombre previst de nombres acabats en un altre nombre diferent de 7 seria 32,4. Els càlculs es fan de la manera següent:

Categoria	E_i	O_i	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
Acabats en 7	3,6	10	6,4	11,38
Diferents de 7	32,4	26	-6,4	1,26
Total	**36,0**	**36**	**0,0**	**12,64**

L'estadística χ^2 de 12,64 és molt superior al valor crític de 3,84, per això concloem que la proporció de 7 apareguts per casualitat és més alta de la prevista. Naturalment, això podria ser degut al fet que més persones estiguessin comprant nombres acabats en 7. Anteriorment, hem suposat com a hipòtesi que els nombres apareixien amb la mateixa probabilitat.

4. Per exemple, si el 16,5% de la població està en el grup d'edat fins a 18 anys, llavors 0,165 · · 1.045 = 172,4 de la mostra pertanyen a aquesta categoria, etcètera.

Categoria	E_i	O_i	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
Fins a 18 anys	172,4	207	34,6	6,944
18-35 anys	190,2	259	68,8	24,887
36-49 anys	313,5	305	-8,5	0,230
49-69 anys	266,5	188	-78,5	23,123
70 o més anys	102,4	86	-16,4	2,627
Total	**1.045,0**	**1.045**	**0,0**	**57,81**

La bondat de l'ajustament és molt superior al valor crític 9,488 de la distribució χ^2 amb 4 graus de llibertat. Aleshores la mostra no és representativa de la població respecte dels grups d'edats.

Apartat 16

1. Demostrem la solució mitjançant el MacAnova:

```
Cmd> table <- vector(5,16,3,16,4,3,0,10,8,5,2,16)

Cmd> table <- matrix(table,3)

Cmd> table
(1,1)          5         16          0          5
(2,1)         16          4         10          2
(3,1)          3          3          8         16

Cmd> rsum <- vector(26,32,30)

Cmd> csum <- vector(24,23,18,23)

Cmd> n <- 88

Cmd> esp <- rsum %*% t(csum) / n

Cmd> esp
(1,1)     7.0909     6.7955     5.3182     6.7955
(2,1)     8.7273     8.3636     6.5455     8.3636
(3,1)     8.1818     7.8409     6.1364     7.8409

Cmd> (table-esp)^2/esp
(1,1)    0.61655     12.468     5.3182    0.47438
(2,1)     6.0606     2.2767     1.8232     4.8419
(3,1)     3.2818     2.9887    0.56599     8.4902

Cmd> sum(vector((table-esp)^2/esp))
(1)       49.206
```

L'estadística χ^2 de 49,206 es compara ara amb el valor crític de la distribució χ^2 amb $(3 - 1) \cdot \cdot (4 - 1) = 6$ graus de llibertat, que a la taula C de l'annex 3 és igual a 12,59. És molt més alt que aquest valor, així que concloem que hi ha proves evidents d'una associació entre edat i preferència musical.

Apartat 17

(No hi ha activitats)

Apartat 18

1.

```
Cmd> nor<-vecread("SAVINGS")

Cmd> hist(nor)

Cmd> nor_sort<-sort(nor)

Cmd> nor_sort
 (1)     0.04     0.13    0.212    0.225    0.233
 (6)    0.269    0.272    0.273    0.287    0.296
(11)    0.297    0.309     0.31     0.33    0.336
(16)    0.347    0.354    0.369    0.379    0.379
(21)    0.383    0.388    0.394    0.416    0.419
(26)    0.431    0.436    0.455    0.473    0.475
(31)    0.475    0.494    0.506     0.51    0.513
(36)    0.516    0.546     0.55    0.561    0.561
(41)    0.568    0.575     0.58    0.581    0.594
(46)    0.608     0.63    0.632    0.647     0.65
(51)     0.65    0.667    0.688     0.69     0.71
(56)    0.718    0.723    0.769     0.79    0.794
(61)    0.797    0.824     0.87    0.877    0.894
(66)    0.894    0.907    0.912    0.948    0.965
(71)    0.994    1.012    1.023     1.05    1.058
(76)     1.06     1.12    1.136    1.143    1.186
(81)     1.21    1.233    1.269    1.308    1.322
(86)    1.383     1.39    1.449    1.634    1.637
(91)    1.703    1.858    2.236    2.318     2.35
(96)    2.475    2.567    2.991    3.521    3.684
```

A partir de la llista de valors confegida de més amunt, el màxim i el mínim són 0,04 i 3,684. La mediana és a mig camí entre els valors 50è i 51è, tant l'un com l'altre són 0,65; per tant, la mediana és 0,65. El primer quartil és a mig camí entre els valors 25è i 26è, és a dir (0,419 + 0,431) / 2 = 0,425. El tercer quartil és a mig camí entre els valors 75è i 76è, és a dir, 1,059. Els cinc nombres resum són, per tant, 0,040, 0,425, 0,650, 1,059 i 3,684.
Per a obtenir la mitjana també podeu fer servir la funció `describe(nor)`:

```
Cmd> sum(nor)/100
(1)      0.87621
```

La mitjana és més alta que la mediana, ja que la distribució és asimètrica amb una cua a la dreta.

2.

```
Cmd> li <- 180

Cmd> ls <- 240

Cmd> stli <- (ll-157)/52

Cmd> stli
(1)      0.44231

Cmd> stls <- (ul-157)/52

Cmd> stls
(1)       1.5962

Cmd> cumnor(stls) - cumnor(stli)
(1)      0.27391
```

3. La distribució de la mitjana també és normal, amb la mateixa mitjana (157 segons), però una desviació estàndard més petita: $52 / \sqrt{1.000} = 1,64$.

4. Segons el teorema central del límit, la proporció d'èxits segueix una distribució normal, amb una mitjana de 0,4 i un error estàndard: $\sqrt{0,4 \cdot 0,6 / 200} = 0,03464$. El valor estandarditzat és, doncs, ((0,5 − 0,4) / 0,03464 = 2,8868. També podeu fer els càlculs fent servir el MacAnova:

```
Cmd> se <- sqrt(0.4*0.6/200)

Cmd> se
(1)     0.034641

Cmd> x <- (0.5-0.4)/se

Cmd> x
(1)       2.8868
```

Finalment, la probabilitat d'èxit del 50% o més serà l'àrea sota la corba normal estàndard sobre 2,8868, la qual serà 0,00195, tal com us mostra el càlcul del MacAnova:

```
Cmd> 1 - cumnor(x)
(1)    0.0019462
```

Aquest exemple mostra que perdre en el joc és molt fàcil.

5. L'interval de confiança del 95% fa referència al fet que, si repetim el nostre experiment moltes vegades, esperem que la veritable mitjana aritmètica de la població es trobarà dins l'interval de confiança en el 95% de les vegades. Si volem un interval de confiança del 100%, voldrem que l'interval contingui sempre la mitjana poblacional. Solament podem estar segurs d'això si el nostre interval va de menys infinit a més infinit, però aquest interval no té cap utilitat.

6. L'error estàndard: 2.570 / $\sqrt{50}$ = 363,45.
El punt del 2,5% d'una distribució *t* amb 49 graus de llibertat és $t_{0,025;49} = -2,0096$ (fent servir el MacAnova `invstu(0.025,49)`).
Per tant, el marge d'error és 2,0096 · 363,45 = 730 ptes.
I l'interval de confiança és 8340 ± 730: [7.610, 9.070] ptes.

7. L'error estàndard: $\sqrt{0,35 \cdot 0,65 / 1.492} = 0,0123$.
Per tant, l'interval de confiança del 95% és: 0,35 ± 1,96 · 0,0123 = [0,326; 0,374]

8. L'estadística χ^2 és igual a

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$= 3,53 + 0,06 + 3,35 + 3,30 + 0,05 + 3,15 = 13,44.$$

Aquest valor és molt superior al valor crític de graus de llibertat = 1 · 2 = 2, per això concloem que és evident que hi ha una diferència d'opinió.

Annexos

Annex 1

Taula A

Àrees sota la corba normal estàndard										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0009	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2-3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6803	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.9729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.8573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	09989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Annex 2

Taula B

Dígits aleatoris							
19223	95034	05756	28713	96409	12531	42544	82853
73676	47150	99400	01927	27754	42648	82425	36290
45467	71709	77558	00095	32863	29485	82226	90056
52711	38889	93074	60227	40011	85848	48767	52573
95592	94007	69971	91481	60779	53791	17297	59335
68417	35013	15529	72765	85089	57067	50211	47487
82739	57890	20807	47511	81676	55300	94383	14893
60940	72024	17868	24943	61790	90656	87964	18883
36009	19365	15412	39638	85453	46816	83485	41979
38448	48789	18338	24697	39364	42006	76688	08708
81486	69487	60513	09297	00412	71238	27649	39950
59636	88804	04634	71197	19352	73089	84898	45785
62568	70206	40325	03699	71080	22553	11486	11776
45149	32992	75730	66280	03819	56202	02938	70915
61041	77684	94322	24709	73698	14526	31893	32592
14459	26056	31424	80371	65103	62253	50490	61181
38167	98532	62183	70632	23417	26185	41448	75532
73190	32533	04470	29669	84407	90785	65956	86382
95857	07118	87664	92099	58806	66979	98624	84826
35476	55972	39421	65850	04266	35435	43742	11937
71487	09984	29077	14863	61683	47052	62224	51025
13873	81598	95052	90908	73592	75186	87136	95761
54580	81507	27102	56027	55892	33063	41842	81868
71035	09001	43367	49497	72719	96758	27611	91596
96746	12149	37823	71868	18442	35119	62103	39244
96927	19931	36809	74192	77567	88741	48409	41903
53909	99477	25330	64359	40085	16925	85117	36071
15689	14227	06565	14374	13352	49367	81982	87209
36759	58984	68288	22913	18638	54303	00795	08727
69051	64817	87174	09517	84534	06489	87201	97245
05007	16632	81194	14873	04197	85576	45195	96565
68732	55259	84292	08796	43165	93739	31685	97150
45740	41807	65561	33302	07051	93623	18132	09547
27816	78416	18329	21337	35213	37741	04312	68508
66925	55658	39100	78458	11206	19876	87151	31260
08421	44753	77377	28744	75592	08563	79140	92454
53645	66812	61421	47836	12609	15373	98481	14592
66831	68908	40772	21558	47781	33586	79177	06928
55588	99404	70708	41098	43563	56934	48394	51719
12975	13258	13048	45144	72321	81940	00360	02428
96767	35964	23822	96012	94591	65194	50842	53372
72829	50232	97892	63408	77919	44575	24870	04178
88565	42628	17797	49376	61762	16953	88604	12724
62964	88145	83083	69453	46109	59505	69680	00900
19687	12633	57857	95806	09931	02150	43163	58636
37609	59057	66967	83401	60705	02384	90597	93600
54973	86278	88737	74351	47500	84552	19909	67181
00694	05977	19664	65441	20903	62371	22725	53340
71546	05233	53946	68743	72460	27601	45403	88692
07511	88915	41267	16853	84569	79367	32337	03316

Annex 3

Taula C

Valors crítics de la distribució χ^2			
Graus de llibertat	**Punt crític**	**Graus de llibertat**	**Punt crític**
1	3,84	20	31,41
2	5,99	21	32,67
3	7,82	22	33,92
4	9,49	23	35,17
5	11,07	24	36,42
6	12,59	25	37,65
7	14,07	26	38,89
8	15,51	27	40,11
9	16,92	28	41,34
10	18,31	29	42,56
11	19,68	30	43,77
12	21,03	40	55,76
13	22,36	50	67,51
14	23,69	60	79,08
15	24,00	70	90,53
16	26,30	80	101,88
17	27,59	90	113,15
18	28,87	100	124,34
19	30,14		