Estadística descriptiva

Introducció a l'anàlisi de dades

Àngel J. Gil Estallo

P08/05057/02302

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Tipus de dades i la seva representació gràfica

En aquesta sessió introduirem els conceptes bàsics de l'estadística descriptiva.

A l'hora de començar a treballar amb estadística, el primer que cal tenir clar és quina és la població objecte del nostre estudi.

La **població** és el conjunt d'individus que es vol estudiar, on els **individus** poden ser tant objectes (iogurts, rodes...) com persones, nens, programadors, etc.

També hem de tenir clar si tenim dades sobre tot el conjunt de la població o només sobre una part d'aquesta població, que anomenarem *mostra*.

Mostra i població

Cal distingir entre població (la totalitat dels individus) i mostra (una part qualsevol de la població). En un altre mòdul s'estudiarà com extreure mostres d'una població.

Una **mostra** és un subconjunt qualsevol de la població.

Exemples de poblacions

Els fitxers emmagatzemats en el disc dur del nostre ordinador, tots els catalans que miren algun cop al dia la televisió, els ordinadors que estan en aquest mateix moment en el magatzem d'una distribuïdora determinada, el conjunt de persones al món que disposen de connexió a Internet, etc. són exemples de poblacions.

Tècniques de mostreig

Us han preguntat mai quins programes mireu? Teniu curiositat per saber com és la mostra sobre la qual es calculen les audiències? Les **tècniques de mostreig** estudien diverses alternatives per a seleccionar mostres a partir de poblacions.

Representativitat d'una mostra

Imaginem que volem esbrinar els programes de la televisió més vistos pels catalans durant un determinat mes. La població seria el conjunt de tots els catalans que han mirat la televisió algun cop durant aquest mes. Com podeu imaginar fàcilment, és impossible d'accedir a tots els individus de la població per a preguntar-los quins programes han vist: el que se sol fer és seleccionar una mostra i preguntar només als individus de la mostra. Els resultats que obtindrem d'estudiar la mostra seran més fiables com més representativa sigui la mostra, és a dir, com més reprodueixi, en petit, l'estructura i composició de la població.

Un cop s'ha decidit la població o mostra objecte d'estudi, triarem les variables que convé estudiar.

Una **variable** és una característica dels individus objecte del nostre estudi.

Exemples de variables estadístiques

Dels fitxers del nostre disc dur, ens pot interessar la grandària (en kbytes), la data en què els vam salvar el darrer cop, l'aplicació que el va crear, etc. De les audiències de la televisió, les variables seran quins programes es miren, durant quanta estona, quin agrada més, etc.

Després recollirem la informació per a obtenir una sèrie d'observacions del valor de la variable sobre els individus de la població o de la mostra. Observeu que tindrem una observació per a cada individu, encara que, evidentment, els valors es poden repetir.

En aquest mòdul donarem les pautes bàsiques del que cal fer quan ens trobem amb un conjunt de dades (valors de les variables) corresponents a una característica determinada d'un conjunt d'individus (població o mostra). De mo-

ment, ens limitarem a estudiar aïlladament cada característica dels individus, és a dir, analitzarem cada variable per separat, i per això direm que fem una anàlisi univariant.

En essència, caldrà explorar les dades per a arribar a trobar una descripció de la població (o la mostra) la més acurada possible. En aquest sentit, es tracta de fer les funcions següents:

a) Esbrinar la distribució de la variable: els valors que pren i "com" els pren (si alguns valors es repeteixen molt, si alguns valors són "estranys", etc.).

b) Calcular alguns resums numèrics que ajudin a entendre quin és el "centre" dels valors i com es distribueixen els valors al voltant d'aquest "centre".

c) Dibuixar alguns gràfics que ajudin a visualitzar els punts mencionats anteriorment.

La importància dels gràfics

Un bon gràfic sempre és de gran ajuda: representeu sempre les dades!

Aquestes operacions ens han de permetre d'elaborar una descripció global de les dades a partir de la qual hem de ser capaços d'aconseguir el següent:

- arribar a conclusions sobre les dades, recolzades a partir dels resums numèrics i dels gràfics (com, per exemple, "tal com es veu en aquest gràfic..." o "les mesures de centre indiquen que...").

- comparar dades referides a la mateixa característica sobre dos conjunts diferents d'individus (com, per exemple, "en el gràfic corresponent al primer col·lectiu es veu..., mentre que en el del segon col·lectiu es veu..." o bé "els resums numèrics del primer col·lectiu mostren com..., mentre que en el segon col·lectiu veiem que...").

- plantejar-nos preguntes més complexes (com, per exemple, sobre relacions entre variables –regressió– o sobre si hi ha diferències entre dos col·lectius).

- veure, en una mostra, quines preguntes ens podem fer respecte les característiques globals de la població –*inferència*.

Veureu la inferència estadística i la regressió en altres mòduls.

1. Tipus de variables

Quan estudiem una població o mostra determinada, seleccionem unes quantes variables rellevants. Aquestes variables poden ser de diferents tipus:

1) Variables qualitatives: són aquelles que no s'expressen numèricament, sinó com a categories o característiques dels individus. De vegades reben el nom de **variables categòriques**.

2) Variables quantitatives: són les que s'expressen numèricament. D'entre aquestes darreres, en podem distingir els tipus següents:

Variables qualitatives i quantitatives

El color dels ulls, la professió d'una persona, el tipus d'ordinador que utilitza són variables qualitatives.

El nombre de fills d'una persona (comptem fills), els punts obtinguts per un equip en un partit de bàsquet (comptem punts), etc., són variables quantitatives (numèriques) discretes.

El pes d'una persona, la cotització de l'euro respecte al dòlar, el temps d'accés a una base de dades, etc., són variables numèriques contínues.

a) Variables quantitatives discretes: només prenen valors enters. Generalment provenen de comptar unitats d'una classe determinada.

b) Variables quantitatives contínues: poden prendre qualsevol valor dins un interval. Acostumen a ser el resultat de mesurar algun fenomen.

En la pràctica, moltes variables qualitatives es codifiquen (s'assigna un nombre a cada categoria) per a facilitar-ne l'estudi.

Exemple de codificació de variables qualitatives

La variable "Tipus d'ordinador que utilitza normalment per a connectar-se a Internet" té un nombre finit de possibilitats diferents. Si s'assigna un nombre enter a cada tipus d'ordinador possible (per exemple, 1 = "PC", 2 = "Mac", 3 = "Altres") obtenim una codificació numèrica que descriu la variable que volem estudiar.

De la mateixa manera, una variable discreta que pren un nombre finit de valors es pot tractar com una variable qualitativa, en què tots els individus on la variable pren un valor determinat formen una categoria.

Exemple de categorització d'una variable discreta

La variable "Nombre de fills d'una parella" és quantitativa discreta (i a més pren pocs valors, potser 15 o 16 valors diferents com a màxim). D'altra banda, totes les parelles en què la variable val 1 constitueixen la categoria de les parelles amb un fill únic, totes les parelles on la variable val 2 constitueixen la categoria de les parelles amb dos fills, etc.

2. Variables qualitatives i variables numèriques discretes que prenen un nombre petit de valors diferents

Suposem que hem d'estudiar una variable categòrica que pot prendre k valors possibles, o bé una variable quantitativa discreta que pot prendre k valors diferents, on k és un nombre relativament petit. Aquests dos tipus de variables admeten un tractament similar, en què cal començar per esbrinar-ne les característiques següents:

1) El nombre total d'individus dels quals es disposen dades; designarem aquest nombre per N.

2) La **freqüència absoluta** de cada valor de la variable: és a dir, el nombre d'individus per als quals la variable pren aquest valor. Designarem per n_i la freqüència absoluta del valor i.

Freqüències absoluta i relativa d'una categoria

La freqüència absoluta d'una categoria és el nombre d'individus que pertanyen a la categoria.

La freqüència relativa d'una categoria és la proporció d'individus que hi pertanyen.

3) La **freqüència relativa** de cada valor de la variable: és a dir, la proporció d'individus en els quals la variable pren aquest valor. Designarem per f_i la freqüència relativa del valor i, i l'acostumarem a donar en percentatge.

Amb això tindrem la **distribució de freqüències de la variable**, que és el conjunt dels valors que la variable pren i la freqüència amb què els pren.

Utilitat de la freqüència relativa

La freqüència relativa ens serà molt útil si hem de comparar la mateixa variable en dues poblacions diferents que tenen diferent nombre d'individus.

Atès que la variable ha de prendre algun dels k valors possibles, és evident que $n_1 + n_2 + ... + n_k = N$. A més a més, f_i s'obté de dividir el nombre d'individus en què la variable pren el valor i pel total d'individus, és a dir:

$$f_i = \frac{n_i}{N}$$

És molt fàcil veure que $f_1 + f_2 + ... + f_k = 1$, és a dir, la suma de les freqüències relatives de tots els valors és sempre igual a 1.

En cas que tingui sentit ordenar els valors de la variable, per a cada valor x_j es poden definir la **freqüència absoluta acumulada**, que és la suma de les freqüències dels valors més petits o iguals que x_j, i la **freqüència relativa acumulada**, que és la suma de les freqüències relatives dels valors més petits o iguals que x_j. Més formalment, si tenim N observacions d'una variable que pren k valors diferents $x_1, x_2, x_3, \ldots, x_k$ de manera que $x_1 < x_2 < x_3 < \ldots < x_k$, definim:

- La **freqüència absoluta acumulada** del valor x_j com $N_j = n_1 + n_2 + n_3 + \ldots + n_j$.
- La **freqüència relativa acumulada** del valor x_j com $F_j = f_1 + f_2 + f_3 + \ldots + f_j$.

D'aquestes definicions resulta evident que $N_k = n_1 + n_2 + n_3 + \ldots + n_k = N$ i que $F_k = f_1 + f_2 + f_3 + \ldots + f_k = 1$.

Normalment, les freqüències es representen en forma de taula. Ho veurem de seguida, en l'exemple dels tipus d'ordinadors. Abans, però, tractarem de representar gràficament la distribució de la variable que s'estudia; us suggerim dues formes molt útils i ben simples.

2.1. El diagrama de barres

Procediment per a dibuixar un diagrama de barres.

Per a fer una representació en forma de diagrama de barres es disposa un eix horitzontal sobre el qual se situen tantes barres com categories o valors pren la variable, separades per petits espais, i amb un títol clar a la base de cada barra que indiqui a quina categoria ens referim.

En l'eix vertical del diagrama de barres marcarem una escala que permeti de llegir bé l'altura de cada barra. Aquesta altura serà o bé la freqüència absoluta o bé la freqüència relativa de la categoria corresponent a la barra. En l'eix vertical cal indicar quina freqüència es representa. En l'eix horitzontal també marcarem, si cal, una escala amb les unitats corresponents.

2.2. El diagrama de sectors

Procediment per a dibuixar un diagrama de sectors.

Un diagrama de sectors consisteix en un cercle que es reparteix en diferents sectors (o porcions) de manera que les superfícies dels sectors siguin proporcionals a les freqüències de cadascuna de les categories.

Si ho hem de fer a mà, per a obtenir l'angle que correspon a cada sector hem de multiplicar els 360° de la circumferència per la freqüència relativa de cada classe; després cal fer servir el transportador d'angles.

Exemple del tipus d'ordinadors

Hem fet una enquesta entre estudiants de la UOC en què es demana el tipus d'ordinador que fan servir habitualment a casa per a connectar-se al campus virtual. A continuació en donem les respostes abreujades:

PC, PC, MAC, PCP, PCP, PCP, MAC, A, A, PC
PC, PC, PC, MACP, A, MAC, A, PCP, PCP, PC

on:

- 1 = PC = "PC compatible de sobretaula"
- 2 = PCP = "PC portàtil"
- 3 = MAC = "MAC"
- 4 = MACP = "MAC portàtil"
- 5 = A = "Altres"

Amb aquestes dades podem calcular fàcilment la distribució de freqüències de la variable. Primer comptem les respostes (20), amb la qual cosa $N = 20$. A partir d'aquí fem una taula on s'indiquin les freqüències absolutes i relatives de cada categoria i amb una filera addicional per a les sumes totals, per a comprovar que no ens hem equivocat.

	Freqüència	**Freqüència relativa**
Tipus d'ordinador	n_i	f_i
1 = PC	$n_1 = 7$	$f_1 = 7/20 = 0{,}35 = 35\%$
2 = PCP	$n_2 = 5$	$f_2 = 5/20 = 0{,}25 = 25\%$
3 = MAC	$n_3 = 3$	$f_3 = 3/20 = 0{,}15 = 15\%$
4 = MACP	$n_4 = 1$	$f_4 = 1/20 = 0{,}05 = 5\%$
5 = Altres	$n_5 = 4$	$f_5 = 4/20 = 0{,}2 = 20\%$
Totals	$N = 20$	**100%**

Taula de distribució de freqüències

Aquesta taula mostra la distribució de freqüències de la variable.

Observeu com la suma de la columna ni és:

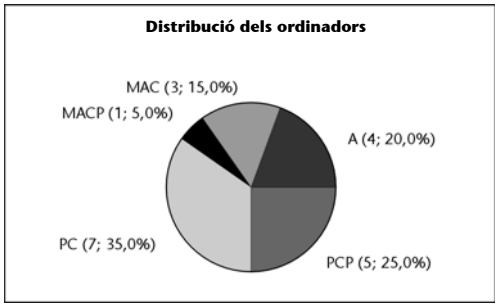
$$7 + 5 + 3 + 1 + 4 = 20 = N$$

i que la suma de les freqüències relatives és:

$$100\% = 1$$

Si fent les sumes no s'obtenen aquests valors segur que hi ha algun error (llevat de problemes d'arrodoniment).

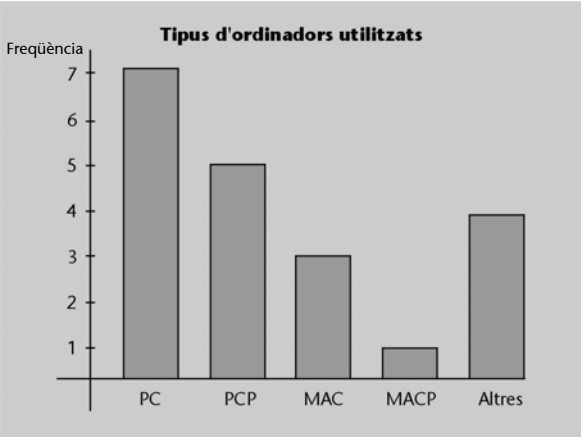
Representem en primer lloc un diagrama de sectors, on veiem que el "pastís" corresponent a la població està repartit en cinc categories i que les porcions més grosses corresponen a ordinadors personals (PC) i personals portàtils (PCP) que en total acaparen el 60% dels usuaris.



Com dibuixar les seccions circulars

Recordeu el transportador d'angles? La majoria de les vegades farem el gràfic amb ajuda d'algun programa informàtic, així que no ens n'haurem de preocupar.

A continuació dibuixem el diagrama de barres amb la freqüència absoluta de cada classe. Observem que les barres més altes corresponen als PC, bé siguin de sobretaula o bé portàtils. Amb aquest gràfic obtenim una visió global del repartiment d'individus per categories.



Exemple de les notes d'estadística

En el semestre anterior, les notes de l'assignatura d'estadística d'un grup d'alumnes van ser les següents:

4, 5, 5, 5, 6, 7, 8, 4, 9, 9, 10, 4, 7, 7 ,7 ,8, 6, 7, 8

A continuació calcularem la distribució de freqüències de la variable i hi afegirem les freqüències acumulades:

	Freqüència absoluta	**Freqüència relativa**	**Freqüència absoluta acumulada**	**Freqüència relativa acumulada**
Nota	n_i	f_i	N_i	F_i
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	3	0,15 = 15%	3	0,15 = 15%
5	3	0,15 = 15%	6	0,3 = 30%
6	2	0,1 = 10%	8	0,4 = 40%
7	6	0,3 = 30%	14	0,7 = 70%
8	3	0,15 = 15%	17	0,85 = 85%
9	2	0,1 = 10%	19	0,95 = 95%
10	1	0,05 = 5%	20	1 = 100%
Totals	**20**			

A partir de la taula es poden deduir els fets següents:

- Un 15% dels alumnes ha tret un 5 (freqüència relativa del valor 5).
- 17 alumnes han aprovat (suma de freqüències absolutes dels valors 5 a 10).
- El 85% dels alumnes ha aprovat (suma de freqüències absolutes dels valors 5 a 10).

3. Variables numèriques

Disposem ara d'un conjunt de dades corresponents a una variable numèrica. A part de les tècniques introduïdes per al cas de les variables discretes amb pocs

valors diferents i, especialment en cas que tinguem una variable discreta amb molts valors diferents o bé una variable contínua, el primer que cal esbrinar és:

- Quants individus apareixen a l'estudi (*N*).
- El **màxim** (*Màx*) i el **mínim** (*Mín*) dels valors que pren la variable.
- El **rang de la variable**, és a dir, la diferència entre el valor màxim i el mínim.

A continuació veurem diferents gràfics que es poden fer servir per a representar aquest tipus de variables.

3.1. El diagrama de punts

Procediment per a dibuixar un diagrama de punts

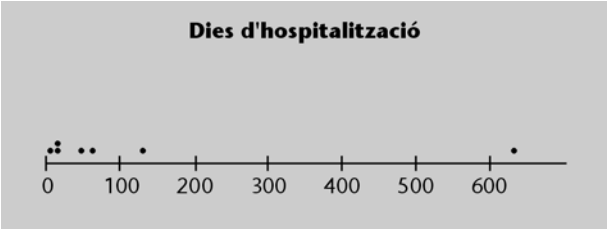
El diagrama de punts consta d'un únic eix horitzontal amb una escala fixada en què els individus es representen per punts dibuixats a sobre del valor que els correspon. En cas de valors repetits situem un punt a sobre de l'altre.

Exemple dels dies d'hospitalització

S'ha fet un estudi sobre els dies d'hospitalització d'un grup de pacients sotmesos a un mateix tractament i s'han obtingut les dades següents: 15, 15, 3, 46, 623, 126, 64 dies.

Tenim, doncs, $N = 7$ dades que van des de $Mín = 3$ fins a $Màx = 623$.

Dibuixem un eix que contingui aquests valors i representem cada pacient per un punt una mica per sobre dels dies que ha estat hospitalitzat, de manera que obtenim el gràfic següent on podem veure que hi ha una concentració de les dades entre els valors 0 i 70 i un punt molt allunyat dels altres (el 623):



Una dada allunyada de les altres o que està fora del patró previsible en el gràfic direm que és una **dada atípica** o **dada extrema**.

Nomenclatura

Utilitzarem indistintament els termes *anòmala, extrema, insòlita, atípica* o la seva expressió anglesa: *outlier*.

Si ens trobem una dada amb aquestes característiques, d'entrada cal comprovar que no es tracti d'un error en la introducció de les dades. Un cop comprovat que no ho és, convé esbrinar la causa d'aquest comportament anòmal.

3.2. El diagrama de tija i fulles

Procediment per a dibuixar un diagrama de tija i fulles

Per a fer un diagrama de tija i fulles cal dur a terme les operacions següents:

Nomenclatura

En anglès, el diagrama de tija i fulles es diu *steam and leaf diagram*.

1) Partim cadascuna de les observacions de la variable en dues parts: la primera conté tots els dígits menys el de més a la dreta (això serà la **tija**), i la segona conté l'últim dígit (això serà la **fulla**).

2) Situem les tiges una a sota de l'altra, ordenades de manera creixent.

3) Col·loquem al costat de cada tija les fulles que li corresponguin, també en ordre creixent.

Per a separar la tija de les fulles es pot dibuixar una línia vertical. Com veurem a continuació, aquest diagrama té un aspecte similar a un diagrama de barres col·locat en posició vertical. En aquest diagrama hi ha representades totes les observacions de la variable i, per tant, hem d'escriure tantes fulles repetides com calgui. Si el gràfic ens sembla poc descriptiu (perquè té pocs nivells, per exemple), podem optar per desdoblar algun nivell.

Exemple del jugador d'escacs

Considerem el nombre de minuts que un jugador d'escacs de gran nivell ha necessitat per a guanyar el programa d'ordinador *Deep Yellow* en 15 partides consecutives:

54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22

L'últim dígit és la fulla, de manera que, en aquest cas, a la tija hi haurà representades les desenes:

2	25
3	45
4	1166679
5	449
6	0

En aquest gràfic podem veure que les dades es concentren en el "nivell" dels 40 minuts i no hi ha dades extremes; sembla un jugador molt regular.

Si considerem que el gràfic és poc descriptiu, podem desdoblar cada desena en dos nivells: el primer contindrà les fulles del 0 al 4 i el segon, les fulles del 5 al 9, de manera que resultarà un diagrama com aquest:

2	2
2	5
3	4
3	5
4	11
4	66679
5	44
5	9
6	0

Aquí s'aprecia millor que l'acumulació de dades és en la zona que va de 45 a 49.

4. Variables contínues i histograma

En cas que la nostra variable numèrica tingui molts valors diferents o que hi hagi molts individus (més de 100, per exemple) el diagrama de tija i fulles pot ser massa carregat i difícil d'interpretar. També ens podem trobar amb taules de distribució de freqüències molt "avorrides" on totes les freqüències siguin iguals a 1 (aquest seria el cas si tots els valors de la variable fossin diferents). Per a simplificar aquestes situacions se sol agrupar les dades i representar-ne, no les freqüències de cada valor, sinó les de les diferents agrupacions.

Aquestes consideracions porten a la necessitat de definir allò que s'entén per **distribució de freqüències** d'una variable numèrica contínua o bé discreta amb molts valors diferents.

Per a obtenir una expressió de la **distribució de freqüències** farem el següent:

Procediment per a calcular la distribució de freqüències en el cas continu

1) Agrupem les observacions en intervals (generalment tots de la mateixa amplada) anomenats **classes.** Els intervals han de ser adjacents (no s'hi val, a deixar-hi forats!) i han de cobrir, com a mínim, tot el rang, des del mínim fins al màxim.

2) Calculem el punt mitjà de cada interval, anomenat **marca de classe**. La marca de classe serà el "representant" de totes les observacions que cauen dins l'interval.

3) Un cop tenim les classes, calculem la **freqüència absoluta** de cada classe (el nombre d'observacions que cauen dins l'interval) i la seva **freqüència relativa** (la proporció d'observacions que cauen dins l'interval).

4) Si cal, també podem calcular la **freqüència absoluta acumulada** de la classe (el nombre d'observacions que cauen en l'interval, més les que cauen en intervals anteriors) i la **freqüència relativa acumulada** de la classe (suma de les freqüències relatives de la classe més la de totes les classes anteriors).

En aquest cas, una manera molt eficient de representar la distribució de freqüències és mitjançant els anomenats *histogrames*.

Un **histograma** és més que un diagrama de barres: de fet, és un diagrama de rectangles. Cada rectangle del diagrama representarà una de les classes en què tenim distribuïts els valors de la variable, de manera que la proporció sobre l'àrea total de l'àrea de cada rectangle és precisament la freqüència relativa de la classe que representa. En el cas que la suma de l'àrea de tots els rectangles sigui 1, l'àrea de cada rectangle serà igual a la freqüència relativa de la classe que representa i direm que hem construït un **histograma de densitat**.

Per a construir un **histograma de densitat**, hem de seguir els passos següents:

Procediment per a dibuixar un histograma de densitat

1) Seleccionem una escala adequada en l'eix de les x.

2) Marquem en l'eix x totes les classes en què es distribueix la variable.

3) Seleccionem una escala adequada en l'eix de les y.

4) Per a cada classe, representem un rectangle que té com a base la classe mateixa, i com a altura, la freqüència relativa de la classe dividida per l'amplada d'aquesta, que, evidentment, és la longitud de la classe.

En aquest cas, és fàcil de veure que la suma de les àrees dels rectangles és 1.

Procediment per a dibuixar un histograma de freqüències relatives

En el cas bastant habitual que totes les classes tinguin la mateixa amplada, podem fer un altre tipus d'histograma que consisteix simplement a posar la freqüència relativa de la classe com a altura de cada rectangle. D'aquesta manera, la suma de totes les àrees dels rectangles és precisament l'amplada de les classes i si dividim l'àrea d'un dels rectangles per l'àrea total obtenim precisament la freqüència relativa de la classe. Aleshores diem que es tracta d'un **histograma de freqüències relatives**.

Recomanacions

Especifiqueu sempre les unitats en què mesureu l'escala de cadascun dels eixos. Indiqueu també els títols que convingui.

Si les classes tenen amplades diferents, és aconsellable de treballar amb histogrames de densitat.

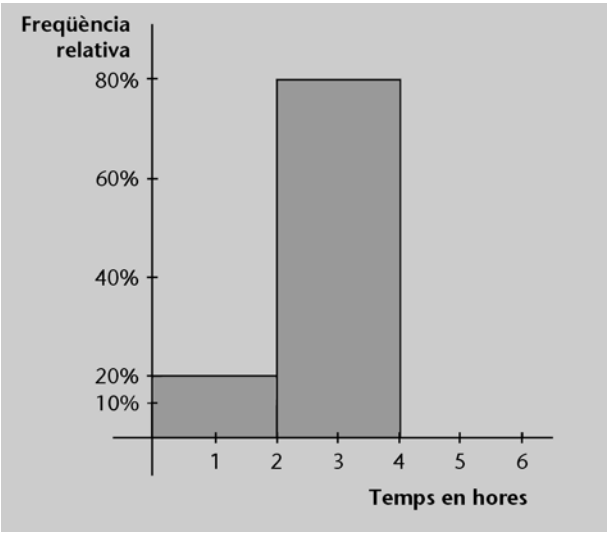
Normalment, i sempre que sigui possible, construirem histogrames de freqüències relatives, ja que són més fàcils d'interpretar i, de fet, tenen el mateix aspecte visual que un histograma de densitat (si les classes són de la mateixa amplada). En tot cas, sempre cal deixar clar quin tipus d'histograma es construeix i indicar si en l'eix de les y es representen freqüències relatives o densitat.

Exemple d'elaboració d'un diagrama de freqüències relatives

S'ha fet un estudi sobre el nombre d'hores diàries que els estudiants de la UOC miren la televisió i s'ha constatat que el 20% la miren menys de 2 hores i que el 80% restant, la miren 2 hores o més, però menys de 4 hores. Representem les dades en forma de taula:

Classes	f_i
[0,2)	20%
[2,4)	80%

Si dibuixem l'histograma corresponent (per exemple, de freqüències relatives, ja que l'amplada de totes les classes és igual a 2) obtenim el gràfic següent:



En aquest gràfic, la suma de les àrees de tots els rectangles és:

$$2 \times 20\% + 2 \times 80\% = 2;$$

L'àrea del primer rectangle és $2 \times 20\% = 0{,}4$, que és precisament el 20% de 2 (l'àrea total), i l'àrea del segon rectangle és $2 \times 80\% = 1{,}6$, que és precisament el 80% de l'àrea total (que és 2).

Exemple d'elaboració d'un diagrama de densitat

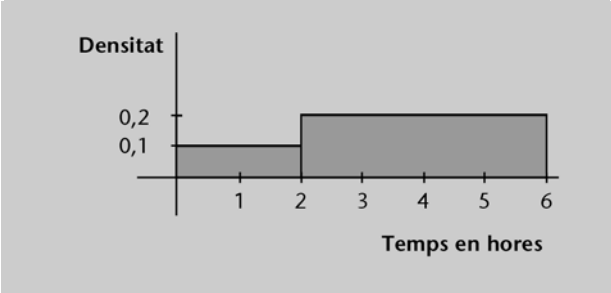
Un altre estudi sobre el nombre d'hores diàries que els estudiants de la UOC miren la televisió ha constatat que el 20% dels estudiants la miren menys de 2 hores i que el 80% restant, la miren 2 hores o més, però menys de 6 hores. Quan representem les dades en una taula, obtenim el següent:

Classes	f_i
[0,2)	20%
[2,6)	80%

En aquest cas hem de dibuixar un histograma de densitat, on l'altura de cada rectangle ve donada per les dades següents:

Classes	f_i	Altura del rectangle
[0,2)	20%	**20% / 2 = 0,1**
[2,6)	80%	**80% / 4 = 0,2**

L'histograma que en resulta és el que veiem a continuació (aquest histograma i l'anterior estan dibuixats amb la mateixa escala en l'eix de les *x* per a facilitar-ne la comparació):



Interpretació dels histogrames de densitat

El diagrama de densitat de l'exemple ens informa del fet que a cadascuna de les hores de la classe [0,2) els correspon un 0,1 = 10% de la població, mentre que a cadascuna de les hores que van de la 2 a la 6 els correspon un 80% / 4 = 20% de la població.

Ara, la superfície total és:

$$2 \times 0{,}1 + 4 \times 0{,}2 = 1.$$

L'àrea del primer rectangle és $2 \times 0{,}1 = 0{,}2 = 20\%$, que és precisament la freqüència relativa de la primera classe, i l'àrea del segon rectangle és $4 \times 0{,}2 = 0{,}8 = 80\%$, que és la freqüència relativa de la segona classe.

Els histogrames de densitat donen una idea de com es distribueixen les observacions dins de cadascuna de les classes.

4.1. Histograma de freqüències relatives acumulades

La funció de distribució acumulada

En parlar de probabilitat i de variables aleatòries farem servir la funció de distribució acumulada. En determinats casos, la representació d'aquesta funció és un histograma de freqüències relatives acumulades.

Per a representar de manera gràfica les freqüències relatives acumulades s'acostuma a utilitzar un histograma. En l'histograma de freqüències relatives acumulades, la. base del rectangle és la classe i l'altura, la freqüència relativa acumulada de la classe. En aquest histograma, l'altura dels rectangles creix fins a arribar a l'altura de l'última classe, que és evidentment 1.

4.2. Exemples de construcció d'histogrames

Com veieu, la descripció de com es calcula la distribució de freqüències per a aquestes variables i de com es fa un histograma són molt generals i depenen de quins són els intervals que seleccionem. No hi ha una norma universal que

expliqui quantes ni quines classes hem de considerar; cada cop ho haurem de decidir segons el tipus i la forma de les dades, sense perdre de vista l'objectiu final.

La finalitat de la construcció d'histogrames és obtenir una representació gràfica que resumeixi la distribució de les dades d'una manera entenedora, fàcil d'assimilar i que mostri els aspectes rellevants de les dades.

Moltes vegades haurem de fer diverses proves fins a aconseguir la representació que il·lustri més bé la forma de la distribució, o que argumenti més bé les nostres opinions.

Histogrames per ordinador

La majoria dels programes d'ordinador permeten d'escollir el nombre i la forma dels intervals: podem fer proves fins a trobar una "bona" representació.

Si hem de fer servir un histograma és convenient tenir en compte els aspectes següents:

a) Sempre que sigui possible, s'agafaran classes de la mateixa amplada.

b) El nombre de classes ha de ser aproximadament igual a $\sqrt{N}$, i millor si està entre 7 i 15 (N és, com sempre, el nombre d'observacions).

La importància del nombre de classes

Si el nombre de classes és massa petit tindrem histogrames amb poques barres i bastant altes; si hi ha massa classes, semblarà un conjunt de pals d'altures similars i amb forats.

c) Per a determinar on ha de començar i acabar cada classe, ens podem guiar per les consideracions següents:

- Ha de quedar molt clar a quina classe han de pertànyer els punts de canvi d'interval; generalment utilitzarem intervals de la forma [x, y), de manera que contenen l'extrem esquerre, però no el dret.

- Quan les dades prenen valors enters com, per exemple, 11, 13, 15, etc., és aconsellable de considerar classes de manera que el seu punt mitjà coincideixi amb els valors enters; en aquest cas, definiríem les classes [10,5-11,5), [11,5-12,5), etc.

Els resultats d'un test d'estadística

A continuació mostrem les notes d'un test d'estadística passat a trenta estudiants d'un curs de la UOC durant el semestre de tardor. La puntuació possible del test va de 0 a 20 punts.

11,5	7	5,25	19	8,5	11	10	6	15,5	11,75
9,75	15	16	16	7,5	8	11,5	0	10	11
13	10,25	10	8	14	6	14,5	10	5,6	10

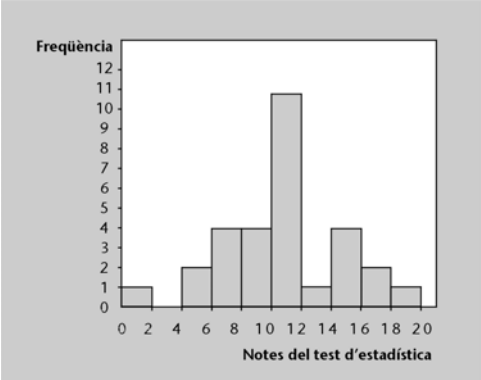
Primer en calculem el valor màxim (*Màx* = 19), mínim (*Mín* = 0) i el rang (*Màx* – *Mín* = 19). En el pas següent s'agrupen les dades en intervals de manera que cobreixin tot el rang. En aquest cas podríem agrupar les dades en intervals de longitud 1, començant pel 0 i acabant pel 20. Com que surten massa classes, optem per fer classes de longitud 2, de manera que obtenim una taula com aquesta, on també s'han calculat les freqüències acumulades.

		Freqüència absoluta	Freqüència relativa	Freqüència absoluta acumulada	Freqüència relativa acumulada
Interval	Marca de classe	n_i	f_i	N_i	F^i
[0,2)	1 = (0 + 2) / 2	1	1/30 = 0,03	1	0,03
[2,4)	3 = (2 + 4) / 2	0	0 = 0	1	0,03
[4,6)	5	2	2/30 = 0,07	3	0,10
[6,8)	7	4	4/30 = 0,13	7	0,23
[8,10)	9	4	4/30 = 0,13	11	0,37
[10,12)	11	11	11/30 = 0,37	22	0,73
[12,14)	13	1	1/30 = 0,03	23	0,77
[14,16)	15	4	4/30 = 0,13	27	0,90
[16,18)	17	2	2/30 = 0,07	29	0,97
[18,20)	19 = (18 + 20) / 2	1	1/30 = 0,03	30	1,00
Totals		**30**	**1**		

Amb aquestes dades podem dibuixar l'histograma següent:

Càlcul del valor mitjà

Per a calcular el punt mitjà d'un interval, se sumen els extrems i es divideix el resultat per 2.



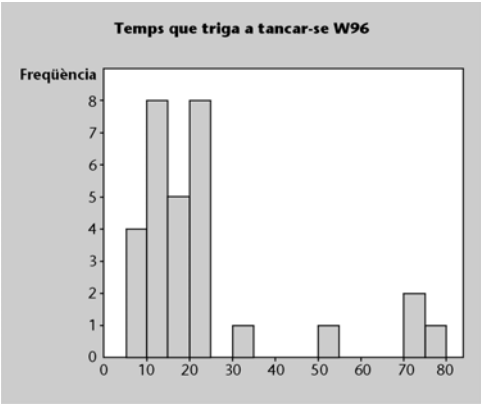
On s'aprecia un cim destacat que correspon a la classe [10,12) i una certa simetria al voltant d'aquesta classe. També s'hi observa una classe buida (la [2,4) té freqüència 0).

Temps que triga a tancar-se l'ordinador

En aquest segon exemple d'histograma utilitzarem les dades següents, que corresponen al temps que ha trigat a tancar-se el nostre ordinador (equipat amb Windows 96) des que hem donat la instrucció adient, les últimes 30 vegades que l'hem utilitzat:

16	32	10	24	23	12	15	21	16	10
24	20	21	71	12	50	76	15	19	8
8	10	12	23	9	71	13	14	20	6

Mitjançant un programa estàndard, hem obtingut l'histograma següent:



Com podeu veure, aquí les classes són [0,5), [5,10) ... fins a [75,80). Hi ha dos pics, corresponents als intervals [10,15) i [20,25), encara que el comportament és força irregular i gens simètric. També hi ha dades allunyades o extremes al voltant dels 70 segons i sembla que hi hagi tres agrupacions en les dades: de 0 a 30, de 50 a 60 i de 70 a 80 segons. Cadascuna d'aquestes agrupacions podria estar relacionada amb les circumstàncies en què l'apaguem i amb els programes que s'han executat.

4.3. Interpretació dels histogrames

En la interpretació dels histogrames és convenient destacar els aspectes següents:

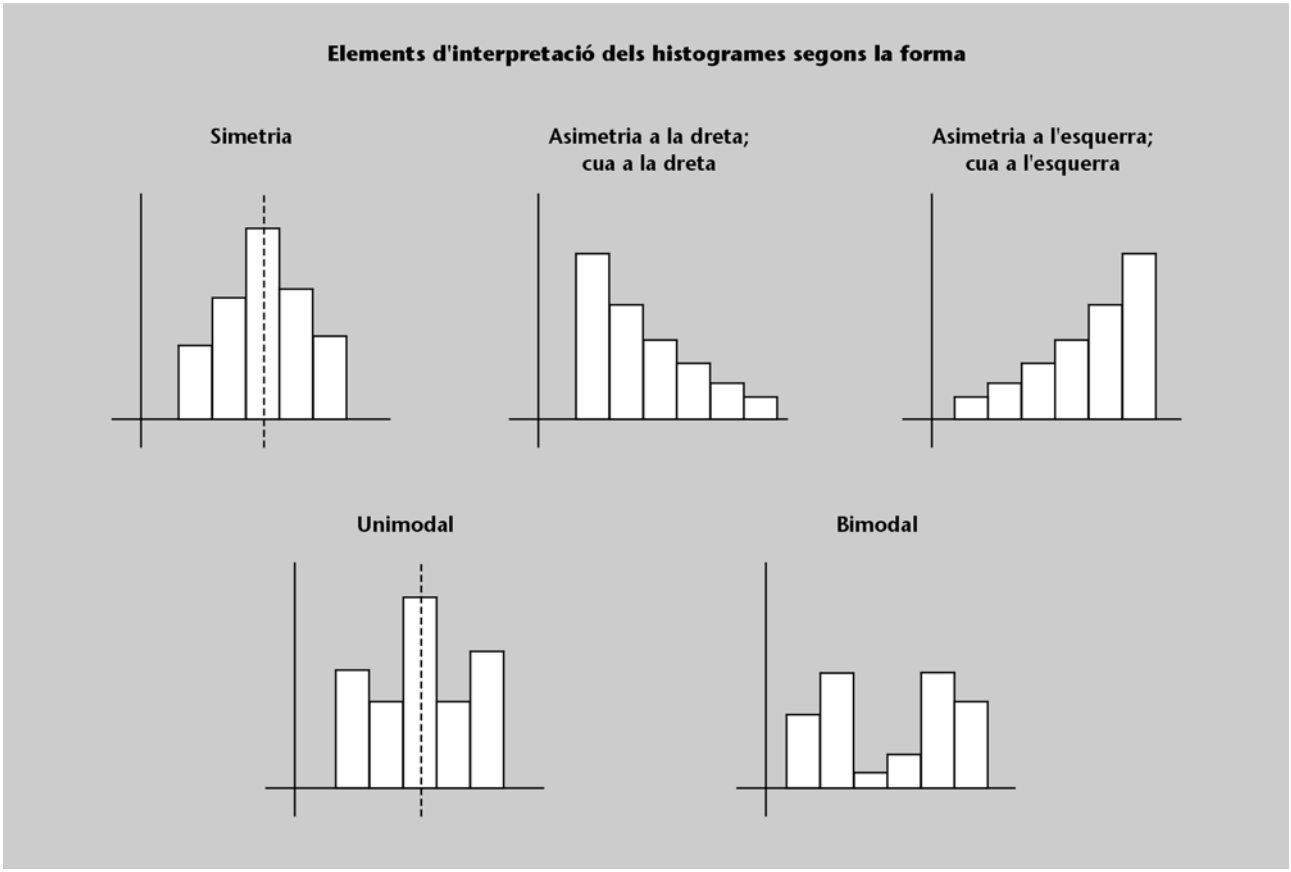
a) La **simetria**: l'histograma serà simètric si és possible traçar un eix vertical de manera que la part que hi ha a la dreta de l'eix és aproximadament igual a la imatge reflectida en un mirall de la part esquerra.

b) Els **pics**: corresponen a intervals on es tendeixen a concentrar els valors de la variable (són intervals amb freqüències grans). Si només hi ha un pic destacat direm que és un **cas unimodal**, si tenim dos pics d'altura similar, serà un **cas bimodal**.

c) Les **cues**: si no hi ha simetria, és possible que un dels costats de l'histograma s'estengui molt més lluny que l'altre. En aquest cas direm que hi ha una cua en el costat més extens (anomenat **cas de cua llarga**) o que hi ha asimetria cap aquest mateix costat.

d) Cal detectar les **dades extremes**, les **classes buides**, i si aquestes classes buides separen la població en grups.

A continuació representem gràficament alguns d'aquests casos, i indiquem l'eix de simetria en els que n'hi ha:



5. Resum

En aquesta sessió hem introduït el vocabulari bàsic de l'estadística (població/mostra, individu, variable, freqüència, distribució de la variable, etc.). Segons el tipus de variable considerada, s'han presentat diverses formes de representar-ne la distribució (diagrama de barres i de sectors per a variables qualitatives i numèriques discretes amb pocs valors, i diagrames de punts, de tija i fulles i histograma per a les quantitatives).

Exercicis

1. Considereu els quatre programes més vistos durant una setmana d'una cadena de TV determinada un dia concret. El nombre d'espectadors (en milers) que miren cadascun dels programes és el que es mostra en la taula següent:

Programa	Nombre d'espectadors
"Viu l'estadística"	875
"El germanot"	925
"Sang i fetge"	742
"Informatiu del dia"	682

a) Representeu gràficament les dades de la manera que cregueu més oportuna.

b) Si la setmana següent el nombre d'espectadors d'aquests mateixos programes és el que reflecteix la taula que hi ha a continuació, representeu les dades de manera que es pugui comparar el nombre d'espectadors en les dues setmanes.

Programa	Nombre d'espectadors
"Viu l'estadística"	100
"El germanot"	200
"Sang i fetge"	100
"Informatiu del dia"	682

2. Una empresa utilitza com a processador de textos el famós programa MacroHard Phrase. S'ha demanat als usuaris d'aquest programa que anotin quantes vegades se'ls ha "penjat" l'ordinador durant un mes mentre hi treballaven. S'han obtingut aquests resultats:

47	23	28	0	11	12	35	36	40	30	37	14

Representeu gràficament les dades de la manera que cregueu més adequada.

3. Atès que el comportament del programa MacroHard sembla força preocupant, hem decidit recollir dades sobre el nombre de vegades que s'ha "penjat" en un determinat mes en 50 ordinadors diferents:

Exemple: l'histograma de Macrohard

0	9	12	14	19
2	10	12	14	20
4	10	12	14	20
5	10	12	15	21
6	11	12	15	22
6	11	12	17	29
6	11	12	17	29
7	11	13	18	32
8	11	13	18	39
9	12	14	19	39

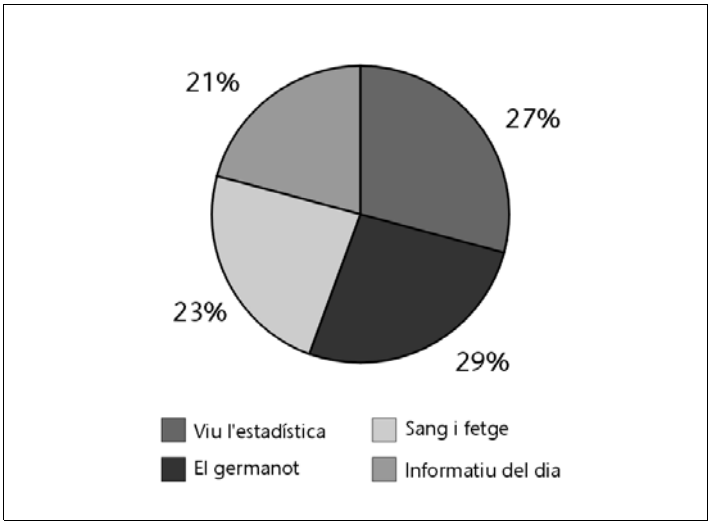
Trobeu la distribució de freqüències d'aquesta variable, representeu-la en forma d'histograma i comenteu-ne les característiques. Representeu l'histograma de freqüències relatives acumulades.

Solucionari

1.

a) Podem fer un gràfic de barres com aquest (per exemple, amb l'Excel):

També podem representar un diagrama de sectors amb les freqüències relatives:

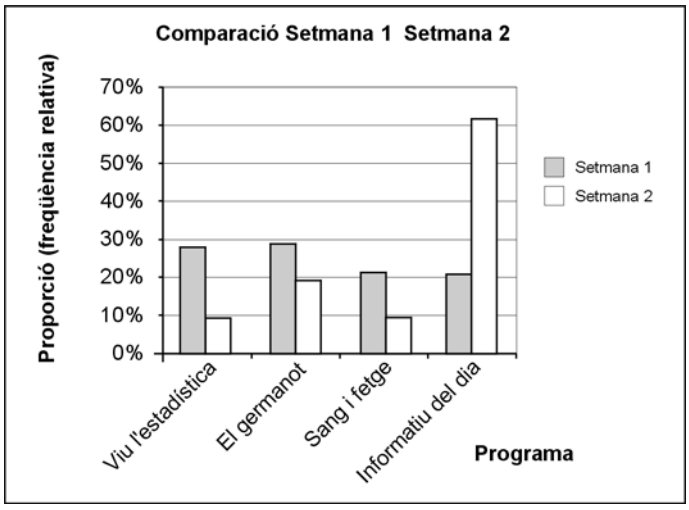


On podem veure que les audiències estan molt igualades.

b) Per a poder comparar les dues setmanes, és millor calcular primer les freqüències relatives de cada programa en cada setmana:

	Freqüència absoluta		Freqüència relativa	
Programa	**Setmana 1**	**Setmana 2**	**Setmana 1**	**Setmana 2**
Viu l'estadística	875	100	27,14%	9,24%
El germanot	925	200	28,69%	18,48%
Sang i fetge	742	100	23,01%	9,24%
Informatiu del dia	682	682	21,15%	63,03%
Total	**3.224**	**1.082**	**100% = 1**	**1**

A partir d'aquestes dades podem crear el gràfic que es mostra a continuació, on per a cada programa es representen dues barres, una per a cada setmana:



En el gràfic es veu clarament que *L'informatiu del dia* ha augmentat molt la seva quota de pantalla, i ha superat àmpliament els altres programes, encara que el nombre d'espectadors s'ha mantingut constant.

Terminolgia

En el cas específic de les audiències de televisió, la freqüència relativa s'anomena *quota de pantalla*.

2. Atès que tenim poques dades, podem optar per fer un diagrama de tija i fulles:

0	0
1	124
2	38
3	0567
4	07

Com podem veure, és un diagrama bastant simètric, distribuït al voltant dels valors entre 20 i 30, amb una certa tendència a desplaçar-se cap als valors grans (30 en endavant). Realment, el programa no sembla gaire fiable.

Començar en 0

És raonable començar en 0, ja que no pot ser que un ordinador es "pengi" menys de zero vegades.

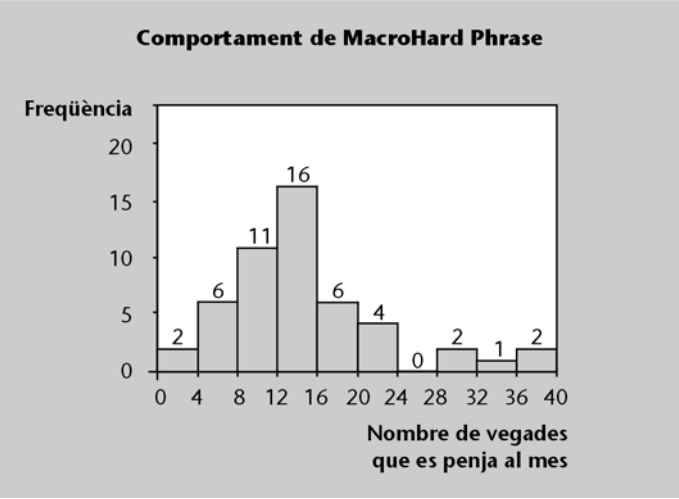
3. El valor mínim que pren la variable és 0, el màxim és 39 i hi ha 50 observacions. Com que el rang és aproximadament 40, podem fer 10 classes d'amplada 4, començant en 0 i acabant en 40. Provem aquesta opció. D'entrada, tabulem les dades per a obtenir les freqüències de cada classe:

		Freqüència absoluta	Freqüència relativa	Freqüència absoluta acumulada	Freqüència relativa acumulada
Interval	Marca de classe	n_i	f_i	N_i	F_i
[0,4)	2	2	4%	2	0,04
[4,8)	6	6	12%	8	0,16
[8,12)	10	11	22%	19	0,38
[12,16)	14	16	32%	35	0,7
[16,20)	18	6	12%	41	0,82
[20,24)	22	4	8%	45	0,9
[24,28)	26	0	0%	45	0,9
[28,32)	30	2	4%	47	0,94
[32,36)	34	1	2%	48	0,96
[36,40)	38	2	4%	50	1
Totals		**50**	**1**		

I després dibuixem l'histograma, en el qual, per comoditat, hem escrit les freqüències de cada classe.

Cas particular

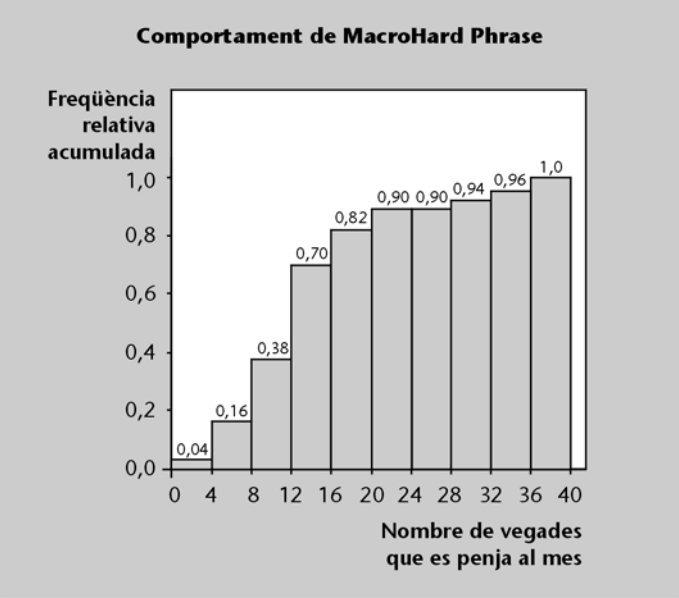
En cas que alguna observació fos exactament 40, podríem optar per considerar l'últim interval de la forma [36,40].



Interpretació

Dos ordinadors s'han penjat entre 0 i 4 vegades, 6 entre 16 i 20 vegades, etc.

L'histograma de freqüències relatives acumulades resulta de la manera següent:



Interpretació

El 70% dels ordinadors s'han penjat menys de 16 vegades; el 90% menys de 28 vegades, etc.

Mesures de centre i propietats

En aquesta sessió aprendrem a assignar, calcular i interpretar diferents indicadors numèrics que ens ajuden a descriure el lloc "central" de les dades. Evidentment, hi ha diverses maneres de descriure el "centre"; nosaltres ens concentrarem en tres mesures: la moda, la mediana i la mitjana, i n'analitzarem la informació que ens aporten i la manera com es poden emprar per a comparar els valors de la variable en mostres diferents.

1. La moda

La moda és el valor més repetit, és a dir, el valor que té una freqüència més gran.

El terme *moda*

Una cosa és moda quan és molt freqüent, és a dir, quan es porta molt!

Si apareixen diferents valors amb la màxima freqüència tindrem diverses modes i la distribució serà bimodal o trimodal.

Exemples de càlcul de la moda

En l'exemple dels tipus d'ordinadors, la moda és PC. En l'exemple dels dies d'hospitalització, la moda és 15. En l'exemple del test d'estadística, la moda és 10.

Si les dades són: 2, 2, 3, 4, 4, les modes són 2 i 4, ja que ambdós valors tenen la freqüència més gran: 2.

Vegeu l'exemple del tipus d'ordinadors en la secció 2.2 i el dels dies d'hospitalització en la 3.1 de la sessió "Tipus de dades i la seva representació gràfica".

2. La mediana

La **mediana** és un valor numèric que ocupa el lloc central un cop ordenades les observacions de més petites a més grans.

Atenció

Vigileu: per a calcular la mediana cal ordenar primer les dades.

La mediana ocupa justament el valor que hi ha al mig de totes les observacions. D'aquesta manera, almenys la meitat de les observacions han de ser més petites o iguals que la mediana i almenys la meitat de les observacions han de ser més grans o iguals que la mediana.

La mediana com a valor central

De vegades hi ha diferents valors que satisfan la definició de mediana; de seguida veureu quina és la pràctica habitual en aquests casos i donarem una regla per trobar la mediana segons el nombre d'individus.

Precisions en la definició de mediana

Pot ser que us pregunteu per què en la definició de mediana utilitzem les expressions *almenys*. A continuació en veiem un cas. Considereu les dades següents:

1	12	13	15	**15**	15	17	18	19

Aquí la mediana és 15 (el 15 que està destacat!) però hi ha 6 observacions amb valor menor o igual que 15 i 6 amb valor més gran o igual que 15, i només tenim 9 observacions!

Amb un parell d'exemples ho veurem ben clar:

a) Càlcul de la mediana quan el nombre de dades és imparell.

La mediana de les nou observacions 12, 13, 11, 16, 17, 19, 18, 15, 14 és 15, ja que si les ordenem ens queden d'aquesta manera:

11	12	13	14	15	16	17	18	19

On es veu clarament que el 15 és al lloc central i 5 observacions són més petites o iguals que 15, i 5 més grans o iguals que 15.

b) Càlcul de la mediana quan el nombre de dades és parell.

En el cas de les vuit observacions següents:

11	12	13	14	15	16	17	18

segurament dubtaríem entre els valors 14 o 15, ja que tots dos, sense ser el "centre" hi estan molt a prop (hi ha 4 observacions menors o iguals que 14 i 5 més grans o iguals, mentre que n'hi ha 5 de més petites o iguals que 15, i 4 de més grans o iguals). En aquest cas, que es dóna quan el nombre de dades és parell, la pràctica habitual és salomònica: s'opta per prendre la mediana com el valor mitjà entre els dos valors centrals. D'aquesta manera, la mediana és $(14 + 15) / 2 = 14{,}5$, que deixa quatre observacions per sota i quatre per sobre (la meitat per sota i la meitat per sobre).

Procediment per a calcular la mediana

Atès que en el càlcul de la mediana el factor important és la posició que les dades ocupen un cop ordenades, a continuació precisem una regla que ens serà molt útil en el cas de tenir un nombre de dades molt elevat. Per a obtenir la mediana d'un conjunt de N dades, hem de seguir els passos següents:

1) Ordenem les observacions de més petita a més gran.

2) Si N és imparell, la mediana és l'observació que ocupa el lloc $(N + 1) / 2$;

3) Si N és parell, calculem el valor $(N + 1) / 2$, i la mediana serà el valor mitjà de les dues observacions que ocupen els llocs més propers a $(N + 1) / 2$.

Error habitual

És un error freqüent dir que la mediana d'un conjunt de N dades és $(N + 1) / 2$. Recordeu que $(N + 1) / 2$ només ens ajuda a trobar la posició que ocupa la mediana; per a conèixer-ne el valor, hem de mirar les dades de què disposem.

D'aquest procediment, cal destacar-ne els dos aspectes següents:

a) $(N + 1) / 2$ és "l'adreça" de la mediana, no n'és el valor.

b) Si N és parell $(N + 1) / 2$ no és un nombre enter.

Exemple de càlcul de la mediana

Suposem que ens proporcionen un conjunt de dades en forma de diagrama de tija i fulles:

8	15
9	22
10	3456789
11	0
12	
13	09

I ens demanen de calcular-ne la mediana. En primer lloc, escrivim els valors de la variable ordenats (ajuntem cada fulla amb la tija corresponent):

81, 85, 92, 92, 103, 104, 105, 106, 107, 108, 109, 110, 130, 139

Atès que hi ha catorze observacions, $(N + 1) / 2 = 15/2 = 7{,}5$ i, per tant, la mediana s'obté com el valor mitjà de les observacions que ocupen els llocs 7 i 8; és a dir, la mediana és:

$$(105 + 106) / 2 = 105{,}5$$

3. La mitjana

La **mitjana** de les observacions d'una variable x (denotada per $\bar{x}$) s'obté dividint la suma de totes les observacions pel nombre d'individus; és, doncs, el valor mitjà de totes les observacions de la variable.

Càlcul de la mitjana

Per a calcular la mitjana:

1) Sumem totes les observacions.

2) Dividim el resultat pel nombre d'individus.

Amb una mica de notació podem escriure aquesta definició formalment. Si $x_1, x_2, \ldots, x_N$ són les observacions de la variable que considerem, aleshores la mitjana es defineix de la manera següent:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Exemple de càlcul de la mitjana

Suposem que els sous mensuals (en milers d'euros) dels companys d'un equip de bàsquet són 1, 4, 2, 4, 1, 6, 12. En aquest cas, la mitjana és:

$$(1 + 4 + 2 + 4 + 1 + 6 + 12) / 7 = 4{,}285 \text{ milers d'euros}$$

Observeu que hem repartit la suma total dels sous entre els 7 companys, de manera que si tots haguessin de cobrar el mateix –i la quantitat total es mantingués fixa– tocarien 4,285 milers d'euros a cadascun, la qual cosa no vol dir que efectivament els cobrin. Observeu, però, que només dues persones cobren més de 4 milers d'euros al mes.

La mitjana sortirà sovint al llarg de l'assignatura i, per tant, és convenient tenir-ne les propietats molt clares i, per què no, també les limitacions.

La mitjana és un repartiment de la suma total dels valors considerats entre totes les observacions. Matemàticament, si multipliquem la mitjana pel nombre total d'individus, obtenim la suma total de les observacions de la variable. En efecte, si en la definició de la mitjana s'aïlla la suma dels valors de les observacions obtenim:

$$N\bar{x} = x_1 + x_2 + \ldots + x_N = \sum_{i=1}^{N} x_i$$

Una propietat important

Si multipliquem la mitjana pel nombre d'individus obtenim la suma de totes les observacions de la variable.

3.1. La mitjana com a valor central

Per a veure en quin sentit la mitjana és el "centre" dels valors haurem de fer alguns càlculs basats en l'estudi de les desviacions respecte a la mitjana:

Les diferències entre els valors de les observacions i la seva mitjana $(x_i - \bar{x})$ s'anomenen **desviacions respecte la mitjana**.

Si es fan les operacions següents, és fàcil de veure que la suma de totes les desviacions dóna 0:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_N - \bar{x}) = (x_1 + x_2 + \ldots + x_N) - N\bar{x} = 0$$

Les desviacions banda i banda de la mitjana s'"equilibren".

On l'última igualtat es dedueix de la relació entre la mitjana i la suma dels valors de la variable.

Per tant, podem concloure que la suma de les desviacions respecte de la mitjana és 0.

De vegades es diu que la mitjana representa el centre de gravetat de la distribució.

Càlcul de les desviacions respecte de la mitjana

Considerem els valors 0, 1, 4 i 7, que tenen mitjana $\bar{x} = 3$. Si en calculem les desviacions:

x_i	$(x_i - \bar{x}) = (x_i - 3)$
0	−3
1	−2
4	1
7	4
Total	$\mathbf{-3 - 2 + 1 + 4 = 0}$

En fer la suma de les desviacions, les positives es compensen amb les negatives i el resultat és 0. Gràficament, el significat de les desviacions respecte de la mitjana es pot visualitzar de la manera següent:

3.2. Efecte dels valors allunyats

Els valors allunyats afecten molt la mitjana. De vegades es diu que la mitjana és poc resistent als valors extrems, o bé que és poc robusta. A més, si hi ha poques dades, l'efecte és encara més important.

La mitjana té tendència a desplaçar-se cap a les dades allunyades.

Exemples de l'efecte dels valors extrems en la mitjana

a) Suposem que un jugador de bàsquet ha marcat 1 punt en cadascun dels 5 partits que ha disputat; evidentment, la mitjana de punts d'aquest jugador és 1. Imaginem que juga un altre partit i marca 50 punts. Amb aquesta nova dada, la mitjana del jugador ha passat d'1 a 9,16. Dir que la mitjana de punts és 9,16, descriu suficientment el jugador? En aquest cas, més aviat no, ja que només una vegada ha marcat més d'un punt.

b) Una empresa fabrica biberons, fem un estudi sobre la seva capacitat i obtenim les dades (en mil·lilitres) següents:

270, 270, 271, 271, 271 i 272

Amb aquestes dades, la mitjana de la capacitat és 270,83 ml. Si considerem ara les dades següents:

270, 270, 271, 271, 271, 272 i 390

obtenim que la nova mitjana és 287,85. Com podeu veure, el valor afegit (que és una dada atípica) afecta molt el valor de la mitjana. Aquests exemples ens mostren dues coses: que les dades extremes afecten molt el valor de la mitjana i que si hi ha poques dades, aquest efecte és encara més important.

3.3. Efecte de les transformacions lineals

A continuació examinarem com afecten la mitjana algunes transformacions en les dades inicials. N'estudiarem tres casos, tots tres basats en una situació inicial que consta de N observacions d'una variable amb valors x_1, x_2, ..., x_N, la mitjana de les quals és $\bar{x}$.

1) Si totes les observacions tenen el mateix valor, és a dir, si $x_1 = x_2 = ... = x_N = \nu$, aleshores la mitjana té el mateix valor $x_1 = \nu$, tal com es demostra a continuació:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\overbrace{\nu + \nu + ... + \nu}^{N}}{N} = \frac{N\nu}{N} = \nu$$

2) Si sumem una mateixa quantitat K a totes les observacions, la mitjana dels nous valors (que són $y_1 = x_1 + K$, $y_2 = x_2 + K$, … , $y_N = x_N + K$), és igual a $\bar{x} + K$, tal com es veu si es calcula la mitjana de les y:

$$\bar{y} = \frac{y_1 + y_2 + ... + y_N}{N} = \frac{(x_1 + K) + (x_2 + K) + ... + (x_N + K)}{N}$$
$$= \frac{(x_1 + x_2 + ... + x_N) + NK}{N} = \frac{(x_1 + x_2 + ... + x_N)}{N} + \frac{NK}{N} = \bar{x} + K$$

3) Si multipliquem totes les observacions per una mateixa quantitat K, la mitjana dels valors nous (que ara són $y_1 = Kx_1$, $y_2 = Kx_2$, …, $y_N = Kx_N$) resulta ser igual a $K\bar{x}$, tal com es veu quan fem:

$$\bar{y} = \frac{y_1 + y_2 + ... + y_N}{N} = \frac{Kx_1 + Kx_2 + ... + Kx_N}{N} =$$
$$= \frac{K(x_1 + x_2 + ... + x_N)}{N} = K\bar{x}$$

En resum

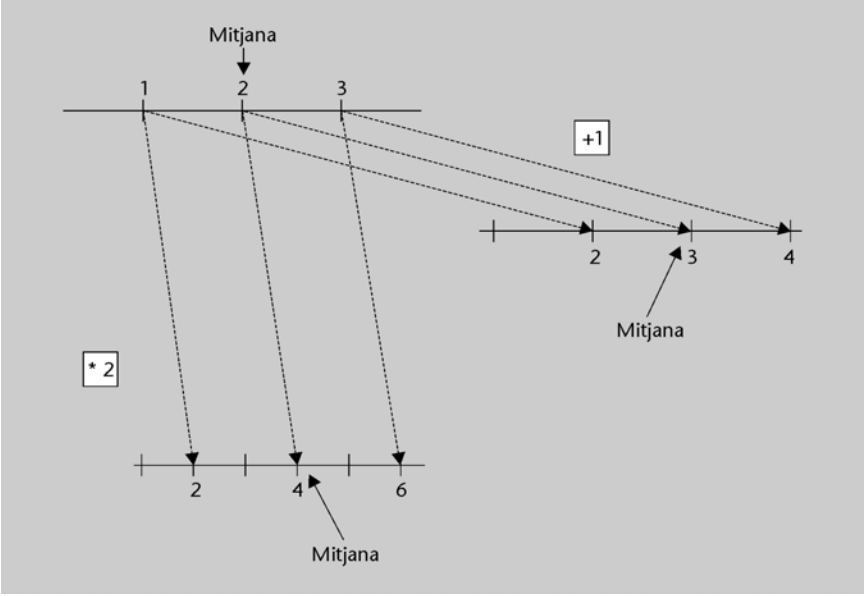
1. Si totes les observacions són iguals, la mitjana coincideix amb el valor de les observacions.

2. Si sumem un mateix nombre a totes les observacions, la mitjana augmenta en aquest mateix nombre.

3. Si multipliquem totes les observacions per un mateix nombre, la mitjana es multiplica per aquest nombre.

Exemples d'efecte d'algunes transformacions en la mitjana

El sou de tres companys és de 1, 2 i 3 milers d'euros respectivament; per tant, la mitjana dels sous és $\bar{x} = 2$. Si el sou de cadascun dels companys augmenta en 1.000 euros, la nova mitjana serà $\bar{y} = \bar{x} + 1 = 3$. Si els nostres amics multipliquen per 2 els seus sous, la nova mitjana serà $\bar{y} = 2\bar{x} = 2 \cdot 2 = 4$ milers d'euros. Podem veure les transformacions dels sous en el gràfic següent, on s'aprecia com la mitjana es desplaça de la mateixa manera que ho fan les observacions: se suma 1 en el primer cas, i es multiplica per 2, en el segon:



4. Comparació mitjana-mediana

De les dues mesures de centre, la mitjana i la mediana, quina és la millor? Malauradament, en el món de l'estadística els nombres són els que són, i no podem dir que siguin ni bons ni dolents. Unes vegades va millor la mitjana i d'altres, la mediana; en altres, depèn del que es pensi fer amb les dades.

Comparació mitjana-mediana

Suposen que la mitjana dels dies d'estada dels pacients amb un tractament concret a l'hospital és 113,1, i la mediana és 31. Si un pacient amb aquest tractament ens pregunta quant de temps és previsible que estigui ingressat, què haurem de respondre? Si responem la mitjana, l'espantarem massa, ja que dels vuit pacients només dos han estat ingressats més de 113 dies. En aquest cas podríem optar per informar-lo del valor de la mediana i del seu sentit: la meitat d'aquests pacients estan ingressats 31 dies o menys. En canvi, si sabem el cost diari d'hospitalitzar aquest tipus de pacient i en volem calcular el cost total, sembla que serà més útil la mitjana, ja que si multipliquem la mitjana de dies d'estada pel cost diari i pel nombre de pacients del tractament, obtindrem el cost total que comporta aquest tipus de pacient.

Un altre aspecte que cal destacar és que si la distribució de les dades presenta una forma simètrica, la mitjana i la mediana coincideixen i, si bé la mitjana té tendència a desplaçar-se cap a la cua llarga de la distribució o les dades allunyades, la mediana no se'n veu tan afectada.

Exemple d'efecte de les dades allunyades sobre la mitjana i la mediana

Si disposem de les dades següents, corresponents a la lectura dels tres termòmetres que tenim a casa (un al costat de l'altre): 20,9, 21 i 21,1, resulta que tant la mitjana com la mediana és 21. Si un altre dia les dades són 20,9, 21 i 36,1 la nova mitjana és 26, mentre que la mediana es manté en 21. (En aquest cas segur que ha passat alguna cosa rara: o no hem llegit bé un dels termòmetres o un d'aquests s'ha espatllat!).

5. Mesures de centre i dades tabulades

Moltes vegades accedim a informació que es presenta en forma de taula de freqüències (absolutes o relatives) dels valors d'una variable. En aquest cas podem calcular fàcilment la moda, la mediana i la mitjana de les observacions resumides a la taula; ho explicarem mitjançant diversos exemples.

Exemple dels virus atacants (I)

Considerem una variable x, les observacions de la qual corresponen al nombre de virus que han atacat cadascun dels 31 ordinadors de la nostra empresa durant l'any 2000. Després d'agrupar les dades, obtenim la taula següent:

x_i	n_i
0	10
5	15
9	6

Segons les dades de la taula, 10 ordinadors no han estat atacats per cap virus, 15 ordinadors han estat atacats (cadascun d'ells) per 5 virus i 6 ordinadors han estat atacats (cadascun d'ells) per 9 virus.

Quina és la moda del nombre de virus que ataquen un ordinador? Segons la definició, la moda és el valor de freqüència més gran, en aquest cas, el 5, que apareix 15 cops.

Si ho creieu convenient podeu escriure totes les observacions ordenades, amb la qual cosa tindríem 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 9, 9, 9, 9, 9, 9.

Quina és la mediana? Com que hi ha 31 ordinadors a l'empresa: $(N + 1) / 2 = 16$ i, per tant, la mediana és l'observació que ocupa el lloc 16, és a dir, el 5, de manera que descobrim que almenys la meitat dels ordinadors han estat atacats per 5 virus com a mínim.

Quina seria la mitjana d'aquestes observacions? Hem de calcular el nombre total d'atacs per virus que han sofert els ordinadors de l'empresa i dividir-lo pel nombre total d'ordinadors. Per a fer-ho, haurem de multiplicar cadascun dels valors de la variable pel nombre de vegades que hem observat aquest valor i dividir aquesta quantitat pel total d'ordinadors, és a dir:

Error greu

És un error greu calcular la mitjana com $(0 + 5 + 9) / 3$, ja que no es té en compte que hem observat 0 atacs en 10 ordinadors, 5 atacs en 15 ordinadors i 9 atacs en 6 ordinadors...

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

on k és el nombre de valors diferents que pren la variable x. Segons aquesta expressió, obtenim:

$$\bar{x} = \frac{10 \times 0 + 15 \times 5 + 6 \times 9}{31} = \frac{129}{31} = 4{,}16 \text{ virus}$$

És a dir, el valor mitjà del nombre de virus que han atacat cadascun dels ordinadors de l'empresa és 4,16.

Exemple dels virus atacants (II)

Estudiem els atacs per virus dels ordinadors d'una altra empresa i obtenim la taula següent:

x_i	f_i
0	10%
5	15%
9	75%

Aquestes dades ens informen que el 75% dels ordinadors han patit 9 atacs, el 15% n'han patit 5 i el 10% restant, cap. Apliquem la fórmula i obtenim:

$$\bar{x} = 10\% \times 0 + 15\% \times 5 + 75\% \times 9 = 7{,}5 \text{ atacs de virus}$$

Altres vegades la informació ens arriba en forma de taula de freqüències relatives. En aquest cas, no coneixem els valors de totes les observacions de la variable, sinó que només sabem el percentatge de vegades en què s'ha observat cada valor. Aleshores, a partir de la fórmula $\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N}$ de la mitjana, deduïm:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \sum_{i=1}^{k} \frac{n_i}{N} x_i = \sum_{i=1}^{k} f_i x_i$$

Exemple de la pàgina de Guillem Portes

És conegut que una de les pitjors pàgines personals de tot Internet és la de Guillem Portes, propietari d'una coneguda empresa de maquinari. S'ha fet un estudi del temps (en segons) que triga un internauta a sortir de la pàgina a partir del moment que surt la foto de Portes, i s'han obtingut els resultats següents:

Classe	m_i	f_i
[0,1)	0,5	80%
[1,5)	3	10%
[5,15)	10	10%

Segons aquestes dades, el 80% dels internautes triga menys d'1 segon a sortir de la pàgina web, el 10%, entre 1 i 5 segons, i el 10% restant, entre 5 i 15 segons (l'extrem superior no es considera dins els intervals). En aquest cas, la mitjana del temps que es triga a sortir és:

$$\bar{x} = 80\% \times 0{,}5 + 10\% \times 3 + 10\% \times 10 = 1{,}7 \text{ segons}$$

Encara podem considerar altres casos en què la informació ens arriba en forma de taula de freqüències relatives o absolutes de les classes en què hem distribuït els valors que pren una variable determinada; en aquest cas, procedirem com en els dos anteriors, però prendrem com a representant de cada classe la marca corresponent, és a dir:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i m_i}{N} = \sum_{i=1}^{K} f_i m_i$$

on m_i és la marca corresponent a cada classe i k és el nombre de classes en què hem distribuït la variable.

6. Resum

En aquesta sessió s'han presentat tres mesures estadístiques que permeten de descriure el centre de la distribució d'una variable segons diferents punts de vista. La moda, que és el valor més freqüent, la mediana, que és el valor central un cop les dades estan ordenades, i la mitjana, que fa que la suma de les desviacions respecte d'aquest valor sigui 0, s'han definit i estudiat. S'expliquen les propietats de la mitjana i es comparen les característiques de la mediana i de la mitjana. També s'han tractat els casos en què es disposa de dades en forma de taula de freqüències (absolutes o relatives).

Exercicis

1. Els sous (en euros) dels analistes contractats per una empresa informàtica són:

1.150	1.200	1.300	1.450	1.000	1.350	1.900	1.230	1.450	1.010	1.500	1.450

Calculeu el sou mitjà i la mediana dels sous.

Quant augmenten la mitjana i la mediana dels sous en les situacions següents?

a) Si es decideix un augment lineal (la mateixa quantitat per a tothom).

b) Si es decideix un augment de sou d'un 12% a tots els treballadors.

c) Si s'augmenta en 100 euros el sou del treballador que cobra més, si s'augmenta en 100 euros el sou del treballador que cobra menys i si s'augmenta en 100 euros el sou d'un treballador qualsevol.

2. Un estudi de la universitat ha recopilat el temps en minuts que els seus estudiants dediquen setmanalment a les tertúlies (*chats*), i s'ha obtingut la taula següent:

x_i	f_i
[0,10)	1%
[10,30)	15%
[30,200)	44%
[200,400)	40%

Calculeu la mitjana del temps que els estudiants dediquen a fer tertúlies i representeu gràficament la distribució de la variable.

Solucionari

1.

a) La mitjana és 1.332,5 euros. Les dades ordenades i les seves respectives posicions un cop ordenades són aquestes:

1	2	3	4	5	6	7	8	9	10	11	12
1.000	1.010	1.150	1.200	1.230	1.300	1.350	1.450	1.450	1.450	1.500	1.900

Com que hi ha 12 observacions, $N = 12$ i $(N + 1) / 2 = 6{,}5$. Per tant, la mediana és $(1.300 + 1.350) / 2 = 1.325$ euros, lleugerament inferior a la mitjana.

a) En general, si el sou augmenta en K euros, la mitjana i la mediana també augmenten en K euros.

b) Augmentar el sou en un 12% equival a multiplicar cada sou per 1,12. Per tant, la mitjana i la mediana també es multiplicaran per 1,12, és a dir, augmentaran en un 12%.

c) La mitjana augmenta en 100/12**€**, ja que si x_1, x_2, ..., x_n són els sous dels treballadors, la mitjana després de pujar 100 euros a qualsevol treballador serà:

$$\frac{x_1 + x_2 + ... + x_{12} + 100}{12} = \frac{x_1 + x_2 + ... + x_{12}}{12} + \frac{100}{12} = \bar{x} + \frac{100}{12}$$

on $\bar{x}$ és la mitjana dels sous originals. Si augmentem el sou al treballador que cobra més o que cobra menys, la mediana no canvia, ja que les observacions que permeten de calcular-la no estan afectades. La mediana només canvia en aquests tres casos:

– Si canviem el sou de 1.230 a 1.330, la nova mediana serà (1.330 + 1.350) / 2

– Si canviem el sou de 1.300 a 1.400, la nova mediana serà (1.350 + 1.400) / 2

– Si canviem el sou de 1.350 a 1.450, la nova mediana serà (1.300 + 1.450) / 2

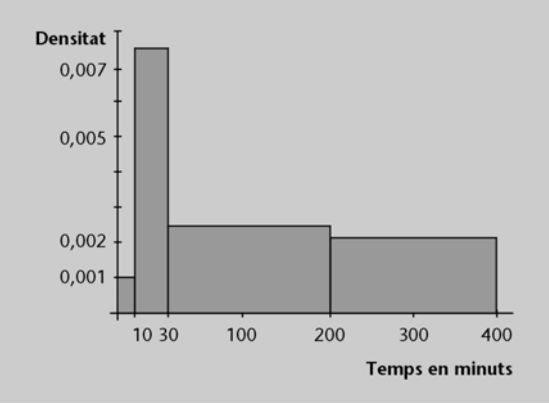
2. Afegim una columna a la taula:

x_i	f_i	**Altura del rectangle**
[0,10)	1%	1% / 10 = 0,001
[10,30)	15%	15% / 20 = 0,0075
[30,200)	44%	44% / 170 = 0,0025
[200,400)	40%	40% / 200 = 0,002

A partir de les marques de classe, la mitjana serà aquesta:

$\bar{x}$ = 1% × 5 + 15% × 20 + 44% × 115 + 40% × 300 = 173,65 minuts

Atès que les classes no tenen la mateixa amplada, representarem gràficament la distribució de la variable amb un histograma de densitat que elaborarem a partir de les dades de la taula:



Mesures de dispersió

En aquesta tercera sessió aprendrem a assignar, calcular i interpretar diferents indicadors numèrics que ens ajuden a descriure com es distribueixen les dades al voltant del seu lloc "central". Evidentment, la manera de distribuir les dades depèn de què entenem per "centre". Per això, estudiarem per separat la dispersió respecte de la mediana i la dispersió respecte de la mitjana. Comencem amb un exemple introductori:

Necessitat de les mesures de dispersió

Amb la mitjana no n'hi ha prou: cal cercar algun indicador que ens ajudi a esbrinar com es distribueixen les dades al voltant d'aquest valor, que ens parli de les possibles fluctuacions del valor de la variable. El mateix ens passarà amb la mediana.

Motivació per a la introducció de les mesures de dispersió

Les cotitzacions de les accions d'una important companyia de telecomunicacions al tancament i durant tres dies consecutius han estat de 9, 10 i 11 euros. Els mateixos dies, les cotitzacions de les accions d'una altra companyia de telecomunicacions han estat d'1, 10 i 19 euros. A simple vista es veu que les mitjanes de les cotitzacions han estat iguals per a les dues companyies, 10 euros, encara que el comportament de les cotitzacions és força diferent: les de la segona han experimentat uns canvis molt grans en la seva cotització, ja que han tingut més variabilitat o més dispersió.

1. Els quartils i la mediana

Començarem amb l'estudi del cas en què la mediana és un bon indicador del centre de les dades i introduirem eines per a descriure com es distribueixen les dades en cadascun dels dos grups en què la mediana divideix la població. La idea de *quartil* prové de l'interès a dividir les dades ordenades en quatre grups amb aproximadament el mateix nombre d'individus per a poder veure "l'espai que ocupa" cada grup en relació amb els altres. Per tant, definim els elements següents:

a) Primer quartil (**Q1**): és aquell valor numèric tal que almenys el 25% de les observacions són menors o iguals que aquell, i almenys el 75%, més grans o iguals.

b) Segon quartil (**Q2**): és la mediana.

c) Tercer quartil (**Q3**): és aquell valor numèric tal que almenys el 75% de les observacions són menors o iguals que aquell, i almenys el 25%, més grans o iguals.

A continuació presentem un procediment de càlcul dels quartils que aprofita les operacions fetes per a calcular la mediana:

Procediment de càlcul per a calcular els quartils Q1 i Q3.

1) Ordenem les observacions en sentit ascendent i calculem (N+1)/2 i la mediana.

2) Distribuïm les observacions en dos grups iguals:

- el grup de les que ocupen una posició estrictament més petita que $(N + 1) / 2$
- el grup de les que ocupen una posició estrictament més gran que $(N + 1) / 2$

3) Un cop fet això:

- el primer quartil (Q1) és la mediana del grup de les dades que ocupen una posició estrictament més petita que $(N + 1) / 2$.
- el tercer quartil (Q3) és la mediana de les dades que ocupen una posició estrictament més gran que $(N + 1) / 2$.
- el segon quartil (Q2) és la mediana.

Exemples de càlcul de quartils

a) Considerem les observacions següents: 17, 16, 12, 13, 15, 14. Aplicarem el procediment que acabem de descriure per a calcular els quartils:

1) Primer ordenem les dades:

12	13	14	15	16	17

També tenim que $N = 6$ i $(N + 1) / 2 = 3{,}5$. Per tant, la mediana és Q2 = 14,5.

2) Distribuïm les dades en dos grups:

Ocupen una posició estrictament més petita que 3,5	12	13	14			
Ocupen una posició estrictament més gran que 3,5				15	16	17

3) Calculem Q1 com la mediana del primer grup i Q3 com la mediana del segon grup; per tant, Q1 = 13 i Q3 = 16. Observem que hi ha dues observacions més petites o iguals que 13, i 5 més grans o iguals; com que el 25% de 6 és 1,5 i el 75% de 6 és 4,5, almenys el 25% de les dades són més petites o iguals que Q1 i almenys el 75% en són més grans o iguals.

b) Considerem les observacions d'una certa variable:

12	13	14	15	16	17	200

En aquest cas tenim $N = 7$ i $(N + 1) / 2 = 4$. La mediana és Q2 = 15. Com en l'exemple anterior, separem les dades en dos grups:

Ocupen una posició estrictament més petita que 4	12	13	14			
Ocupen una posició estrictament més gran que 4				16	17	200

Per tant, Q1 = 13 i Q3 = 17.

c) Considerem les observacions d'una certa variable:

12	13	14	15	15	15	200

Ara, $N = 7$ i $(N + 1) / 2 = 4$. La mediana és Q2 = 15. Si separem les dades en els dos grups d'abans:

Ocupen una posició estrictament més petita que 4	12	13	14			
Ocupen una posició estrictament més gran que 4				15	15	200

Per tant, Q1 = 13 i Q3 = 15. En aquest cas, Q2 = Q3 = 15.

Les definicions de mediana i quartils

La mediana i els quartils no tenen una definició única. En diferents llibres apareixen mètodes diferents de càlcul i els programes d'ordinador utilitzen diferents algorismes, que donen lloc a resultats diferents. En general, si hi ha moltes dades i pocs "forats" entre les dades, els resultats seran molt similars. En qualsevol cas, és convenient informar-se de com es calculen els valors esmentats en cada cas.

1.1. El rang interquartílic

El rang interquartílic és la diferència entre el tercer i el primer quartil, és a dir:

$$\text{Rang interquartílic} = Q3 - Q1$$

Atès que entre Q1 i Q3 es distribueixen aproximadament el 50% de les observacions centrals de la variable, el rang interquartílic és una mesura de la dispersió d'aquest col·lectiu. Així, un rang interquartílic petit significa que les dades centrals estan molt atapeïdes, mentre que un rang interquartílic gran n'indica una forta dispersió.

1.2. Els cinc nombres resum i el diagrama de caixa

Els cinc nombres resum de la distribució d'una variable són: el mínim, Q1, la mediana, Q3 i el màxim.

Hi ha una manera molt eficient de representar els quartils i els quatre grups en què els quartils divideixen les observacions: és el diagrama de caixa o *box-plot*.

Procediment per a dibuixar un diagrama de caixa.

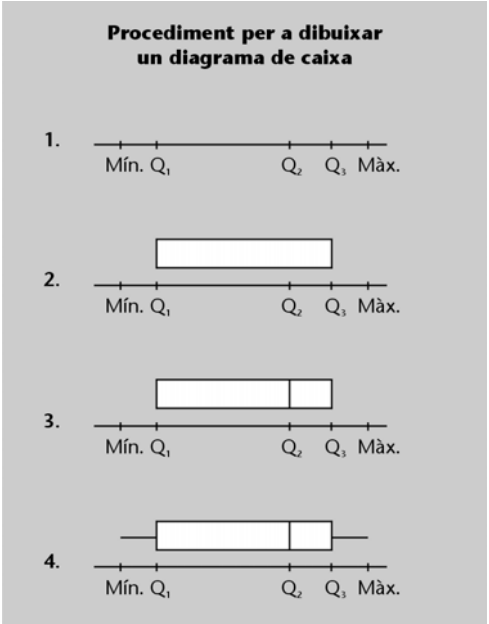
Un **diagrama de caixa** es construeix de la manera següent:

1) Es marca un eix (horitzontal o vertical) amb l'escala adequada de la variable, i es marquen els cinc nombres resum de la distribució.

2) Es dibuixa un rectangle (per sobre o al costat dret de l'eix) que tingui un costat en el valor corresponent a Q1 i un altre en el valor Q3.

3) Dins el rectangle es dibuixa una línia de costat a costat en el valor que correspon a la mediana.

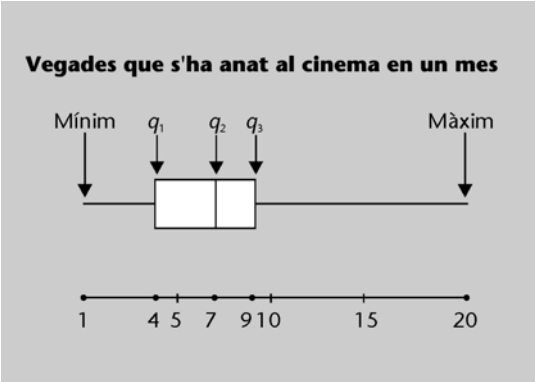
4) Es dibuixen dos braços que van des del punt mitjà dels costats construïts a Q1 i Q3 fins als valors mínim i màxim, respectivament.



Com sempre, és més difícil de descriure'n l'aspecte amb paraules que dibuixar-ne un.

Exemple d'elaboració d'un diagrama de caixa: vegades que es va al cinema

Preguntem a algunes persones quantes vegades han anat al cinema en els darrers trenta dies. Després de processar les dades, obtenim els següents cinc nombres resum: *Mínim* = 1, Q1 = 4, *Mediana* = 7, Q3 = 9, *Màxim* = 20. El diagrama de caixa corresponent és el següent:

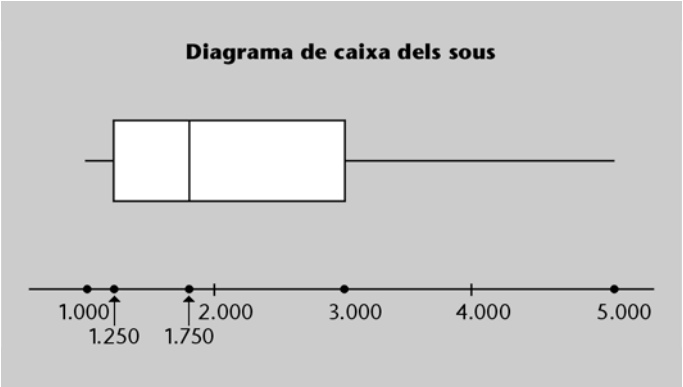


Observem que la població que estudiem es divideix en quatre grups amb aproximadament el mateix nombre de persones. El primer grup ha anat al cinema l'últim mes entre 1 i 4 vegades, el segon entre 4 i 7, el tercer entre 7 i 9, i l'últim grup, entre 9 i 20. El gràfic mostra que no hi ha simetria respecte de la mediana, que hi ha una cua cap a la dreta (el braç de Q3 al màxim és més llarg que del mínim a Q1), que les dades estan més atapeïdes en l'interval que va de Q2 a Q3 (recordeu que en els quatre grups hi ha aproximadament el mateix nombre d'individus i, per tant, aproximadament el 25% dels enquestats van anar entre 7 i 9 vegades al cinema) i que les dades estan més disperses en l'interval que va de Q3 al màxim (això vol dir que la quarta part de la població que va més al cinema –entre 9 i 20 vegades– ho fa amb freqüències bastant diferents). Respecte al 50% de la població amb valors més centrals, podem veure que el rang interquartílic és Q3 – Q1 = 5 i que, per tant, aquest col·lectiu distribueix el nombre de vegades que ha anat al cinema entre 6 possibilitats, de 4 a 9 vegades.

És un error freqüent pensar que els diagrames de caixa són tots iguals i que determinen quatre regions iguals: el diagrama de caixa distribueix les observacions en quatre grups que contenen aproximadament el mateix nombre d'observacions, però no tots els grups ocupen el mateix ''territori''.

Exemple de distribució de les observacions en un diagrama de caixa: els sous

Un estudi mostra que els cinc nombres resum corresponents a la distribució dels sous mensuals d'una empresa determinada són: *Mínim* = 1.000, Q1 = 1.250, *Mediana* = 1.750, Q3 = 3.000, *Màxim* = 5.000. Dibuixem el diagrama de caixa corresponent i obtenim:



Asimetria en l'economia

Sempre que s'estudien sous, despeses, ingressos... ens trobem amb situacions asimètriques cap a la dreta. Això es deu als valors extrems.

En aquest diagrama es veu que, a mesura que els sous creixen, també són més diferents entre si: mentre que la quarta part dels treballadors que cobren menys tenen sous molt similars (entre 1.000 i 1.250), la quarta part dels treballadors que cobren més tenen sous

molt diferents (van de 3.000 a 5.000). El rang interquartílic en aquest cas és 3.000 – 1.250 = 1.750; això vol dir que en l'interval del 50% dels sous més centrals hi ha una oscil·lació de 1.750 euros entre el més baix i el més alt.

1.3. Interpretació dels diagrames de caixa

Els diagrames de caixa donen molta informació sobre la simetria de les dades i sobre com es distribueixen les observacions en quatre grups que contenen aproximadament el mateix nombre d'observacions. També permeten de visualitzar el rang interquartílic. Entre els "defectes" que presenten, cal destacar que, tal com l'hem descrit aquí, l'aspecte d'aquests diagrames depèn bastant dels valors màxim i mínim (per tant, de dos valors individuals) i cal tenir present aquesta informació quan investiguem braços que van del mínim a Q1 i de Q3 al màxim.

Els diagrames de caixa són molt útils per a comparar la distribució d'una variable en diferents poblacions o mostres.

Vegeu la utilitat dels diagrames de caixes per a comparar mostres diferents en l'exercici 1 d'aquesta secció.

2. La desviació típica i la mitjana

Una manera de saber com es distribueixen les dades al voltant de la mitjana podria ser calcular el valor mitjà de les desviacions respecte a la pròpia mitjana. Malauradament, aquest valor mitjà és 0, perquè la suma de les desviacions és 0 (els valors positius es cancel·len amb els negatius). Per a evitar aquestes cancel·lacions, podem agafar el valor de les desviacions al quadrat (que són totes positives). Aquesta idea motiva la definició de variància.

Variància mostral i variància poblacional

Hi ha una "altra" variància, anomenada *variància mostral*, que s'obté com la poblacional però, dividint per $N - 1$. La variància mostral té molta importància quan s'estudia la inferència.

La **variància** (també anomenada **variància poblacional**) i que es denota per s^2 es calcula com la suma del quadrat de les desviacions respecte a la mitjana dividit per N, on N és el nombre de les observacions, és a dir:

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

És fàcil de comprovar que la variància es pot calcular també a partir de la fórmula següent:

Escrivim s_x^2 si volem destacar que calculem la variància de la variable x.

$$s_x^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - (\bar{x})^2$$

La **desviació típica** (denotada per s o $s_{x)}$ es defineix com l'arrel quadrada positiva de la variància, és a dir:

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

Desviació típica mostral i poblacional

Evidentment també disposem de "dues" desviacions típiques: la mostral i la poblacional. Moltes calculadores tenen una tecla per a cadascuna: σ_n per a la poblacional i σ_{n-1} per a la mostral.

Exemple de càlcul de la desviació típica

Considerem els valors 7, 4, 6, 5, 5. Per a calcular-ne la variància, organitzem els càlculs en una taula com aquesta:

Càlcul de la variància

Per a calcular la variància, cal seguir els passos següents:

1) Calcular la mitjana.

2) Calcular les desviacions respecte de la mitjana.

3) Calcular el quadrat de les desviacions respecte de la mitjana.

4) Sumar el quadrat de les desviacions i dividir pel nombre d'individus.

	x_i	$x_i - \bar{x} = x_i - 5{,}4$	$(x_i - \bar{x})^2 = (x_i - 5{,}4)^2$
	7	$7 - 5{,}4 = 1{,}6$	$(1{,}6)^2 = 2{,}56$
	4	$4 - 5{,}4 = -1{,}4$	$(-1{,}4)^2 = 1{,}96$
	6	$6 - 5{,}4 = 0{,}6$	$(0{,}6)^2 = 0{,}36$
	5	$5 - 5{,}4 = -0{,}4$	$(-0{,}4)^2 = 0{,}16$
	5	$5 - 5{,}4 = -0{,}4$	$(-0{,}4)^2 = 0{,}16$
Suma de la columna	**27**	**0**	$\mathbf{2{,}56 + 1{,}96 + 0{,}36 + 0{,}16 + 0{,}16 = 5{,}2}$

On la mitjana és:

$$\bar{x} = \frac{7+4+6+5+5}{5} = \frac{27}{5} = 5{,}4$$

Càlcul de la desviació típica

Per a calcular la desviació típica fem l'arrel quadrada positiva de la variància.

la variància és:

$$s_x^2 = \frac{5{,}2}{5} = 1{,}04$$

i la desviació típica és:

$$s = \sqrt{s^2} = \sqrt{1{,}04} = 1{,}02$$

2.1. Propietats de la variància i la desviació típica

Atès que en la definició de la variància i la desviació típica intervé la mitjana, i aquesta és sensible a les dades extremes, la variància i la desviació típica també es veuen afectades pels valors extrems.

Les transformacions lineals en les dades inicials afecten la variància i la desviació típica de la manera següent:

1) Si totes les observacions tenen el mateix valor, és a dir, si $x_1 = x_2 = \ldots = x_n$, la variància i la desviació típica valen 0:

La variància dels valors 3, 3, 3, 3 és 0 i la seva mitjana 3.

$$s^2 = \frac{(x_1 - \bar{x}) + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N} =$$

$$= \frac{(x_1 - x_1)^2 + (x_1 - x_1)^2 + \ldots + (x_1 - x_1)^2}{N} = 0$$

Observació

Fixeu-vos que la variància sempre és positiva.

ja que totes les observacions són iguals i iguals a la seva mitjana.

2) Si sumem una mateixa quantitat K a totes les observacions, la variància i la desviació típica no canvien, tal com es veu si es calcula la variància dels valors nous (que són $y_1 = x_1 + K$, $y_2 = x_2 + K$, … , $y_N = x_N + K$), i utilitzant que la mitjana dels nous valors és $\bar{y} = \bar{x} + K$:

$$s_y^2 = \frac{(y_1 - \bar{y})^2 + (y_1 - \bar{y})^2 + \ldots + (y_N - \bar{y})^2}{N} =$$

$$= \frac{((x_1 + K) - (\bar{x} + K))^2 + ((x_2 + K) - (\bar{x} + K))^2 + \ldots + ((x_N + K) - (\bar{x} + K))^2}{N} =$$

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N} = s_x^2$$

Observeu de quina forma *K* "desapareix".

3) Si multipliquem totes les observacions per una mateixa quantitat *K*, la variància dels valors nous (que ara són $y_1 = Kx_1$, $y_2 = Kx_2$, …, $y_N = Kx_N$) s'obté multiplicant la variància de les dades originals per K^2; la desviació típica dels nous valors s'obté multiplicant la desviació típica de les dades originals per $|K|$. Recordem que en aquest cas la mitjana dels nous valors és $\bar{y} = K\bar{x}$ i que, per tant:

$|K|$ és el valor absolut de *K*.

$$s_y^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_N - \bar{y})^2}{N} =$$

$$= \frac{(Kx_1 - K\bar{x})^2 + (Kx_2 - K\bar{x})^2 + \ldots + (Kx_N - K\bar{x})^2}{N} =$$

$$= \frac{K^2(x_1 - \bar{x})^2 + K^2(x_2 - \bar{x})^2 + \ldots + K^2(x_N - \bar{x})^2}{N} =$$

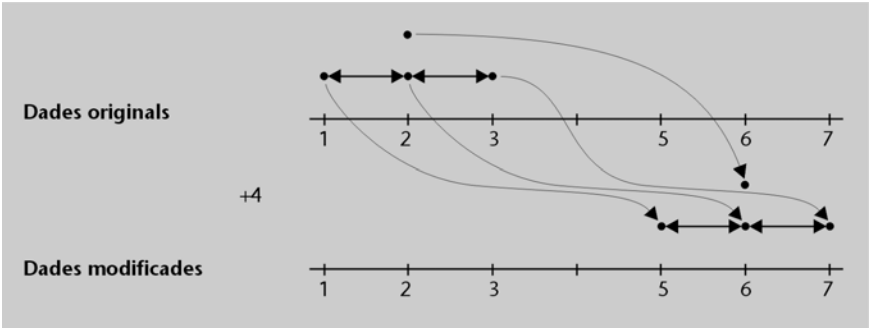
$$= K^2 s_x^2$$

Finalment: $s_y = \sqrt{s_y^2} = \sqrt{K^2 s_x^2} = |K| s_x$.

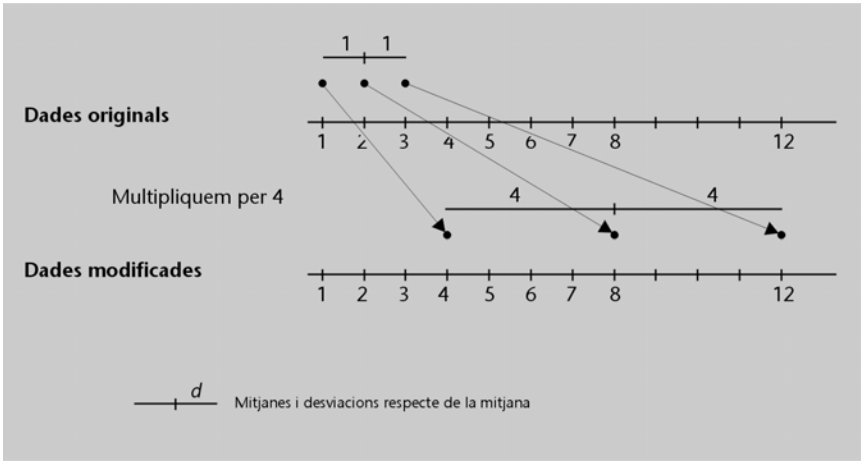
El valor absolut $|K|$ fa que S_y sigui positiva, encara que *K* sigui negativa.

Propietats de la variància i la desviació típica

Si partim dels valors 1, 2, 3 que tenen mitjana 2 i variància 2/3, i sumem 4 a tots, obtenim els valors 5, 6 i 7, que tenen mitjana 6 = 2 + 4 i variància 2/3, igual que la de les dades inicials. Observem en el gràfic que la mitjana es desplaça mentre que les desviacions (i, per tant, la variància i la desviació típica) es mantenen.



Si partim dels valors 1, 2 i 3, que tenen mitjana 2 i variància 2/3, i els multipliquem tots per 4, obtenim els valors 4, 8 i 12, que tenen una mitjana de $8 = 4 \cdot 2$, una variància de $10{,}66 = 4^2 \cdot 2/3$ i una desviació típica de $|4| \cdot \sqrt{2/3} = 3{,}26$.



2.2. La regla de Txebixev

La regla de Txebixev permet d'entendre més bé el paper que té la desviació típica com a mesura de dispersió de les dades al voltant de la mitjana. Començarem per enunciar la regla, després en donarem alguns casos particulars i n'estudiarem l'ús, però no en donarem la justificació aquí.

La regla de Txebixev afirma que, donat qualsevol conjunt de dades x_1, x_2, ..., x_n amb mitjana $\bar{x}$ i desviació típica s_x, si m és un nombre qualsevol, aleshores la proporció de dades que pertanyen a l'interval $(\bar{x} - ms_x, \bar{x} + ms_x)$ és, com a mínim:

$$1 - \frac{1}{m^2}$$

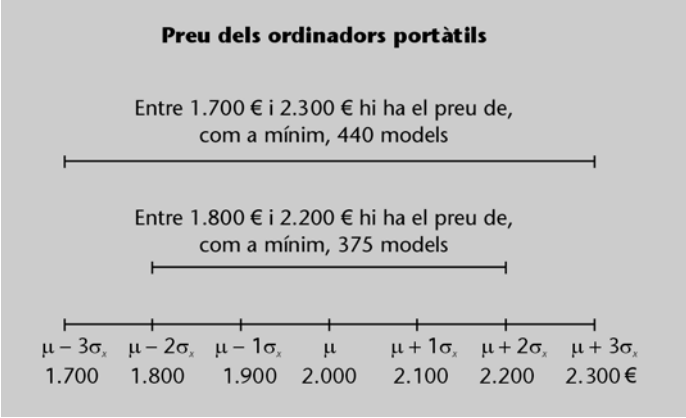
La taula següent registra alguns casos concrets:

m	Interval	Percentatge mínim d'observacions a dins de l'interval previst per la regla de Txebixev
1	$(\bar{x} - 1s_x,\ \bar{x} + 1s_x)$	$1 - \frac{1}{1^2} = 0$
2	$(\bar{x} - 2s_x, \bar{x} + 2s_x)$	$1 - \frac{1}{2^2} = 0{,}75 = 75\%$
3	$(\bar{x} - 3s_x,\ \bar{x} + 3s_x)$	$1 - \frac{1}{3^2} = 0{,}88 = 88\%$
4	$(\bar{x} - 4s_x,\ \bar{x} + 4s_x)$	$1 - \frac{1}{4^2} = 0{,}93 = 93\%$
...	...	...
m	$(\bar{x} - ms_x,\ \bar{x} + ms_x)$	$1 - \frac{1}{m^2}$

Per exemple, podem estar segurs que, siguin quines siguin les nostres dades, en l'interval $(\bar{x} - 4s_x, \bar{x} + 4s_x)$ sempre trobarem més del 93% de les dades, i així amb altres valors de m (excepte $m = 1$ sobre el qual, curiosament, la regla no diu res).

Exemple dels ordinadors portàtils

Recollim informació sobre el preu de tots els models d'ordinador portàtil que es poden trobar al mercat. Suposem que en total tenim 500 models diferents, que la mitjana dels preus és 2.000 euros i la desviació típica és 100 euros. En aquest cas, podem estar segurs que el preu de, com a mínim, 375 (= 75% de 500) portàtils està en l'interval $(2.000 - 2 \cdot 100, 2.000 + 2 \cdot 100) = (1.800, 2.200)$ i que el preu de com a mínim 440 ordinadors portàtils (88% de 500) està en l'interval $(2.000 - 3 \cdot 100, 2.000 + 3 \cdot 100) = (1.700, 2.300)$. Gràficament:



Com veurem en els exercicis, la regla de Txebixev és molt "pessimista": moltes vegades trobarem més dades de les que prediu; això es deu al fet que serveix per a tots els conjunts de dades, siguin quines siguin i, per tant, ha de ser poc restrictiva.

2.3. Dades estandarditzades

Un altre concepte que ajuda a interpretar les relacions entre la mitjana i la desviació típica és el de dades estandarditzades. Donat un conjunt de N observacions d'una variable amb valors $x_1, x_2, ..., x_N$, la mitjana dels quals és $\bar{x}$ amb desviació típica s_x, el valor estandarditzat de x es defineix de la manera següent:

$$z = \frac{x - \text{Mitjana de les } x}{\text{Desviació típica de les } x} = \frac{x - \bar{x}}{s_x}$$

Habitualment, considerarem un conjunt d'observacions amb la mitjana i la desviació típica corresponents i n'estandarditzarem totes les observacions, és a dir, en transformarem els valors originals als valors estandarditzats corresponents:

$$x_i \xrightarrow{\text{Estandarditzar}} z_i = \frac{x_i - \bar{x}}{s_x}$$

Donat un conjunt d'observacions qualsevol, els **valors estandarditzats** corresponents tenen mitjana 0 i desviació típica 1, ja que si les dades estandarditzades són $z_1, z_2, \ldots, z_N$ es verifiquen les propietats següents:

Nomenclatura

Els valors estandarditzats també s'anomenen *z-valors*.

a) $$\bar{z} = \frac{z_1 + z_2 + \ldots + z_N}{N} = \frac{\frac{x_i - \bar{x}}{s_x} + \ldots + \frac{x_N - \bar{x}}{s_x}}{N} =$$

$$= \frac{x_i + \ldots + x_N - N\bar{x}}{Ns_x} = \frac{0}{Ns_x} = 0$$

Recordeu que la suma dels valors de les observacions és el producte de la mitjana pel nombre d'individus.

b) $$s_z^2 = \frac{(z_1 - \bar{z})^2 + \ldots + (z_N - \bar{z})^2}{N} = \frac{(z_1 - 0)^2 + \ldots + (z_N + 0)^2}{N} =$$

$$= \frac{\left(\frac{x_1 - \bar{x}}{s_x}\right)^2 + \ldots + \left(\frac{x_N - \bar{x}}{s_x}\right)^2}{N} = \frac{(x_1 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{s_x^2\, N} =$$

$$= \frac{1}{s_x^2}\frac{(x_1 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N} = \frac{s_x^2}{s_x^2} = 1$$

c) Una altra propietat interessant és que el valor estandarditzat corresponent a una observació determinada ens diu quantes desviacions típiques dista de la mitjana. Això es veu molt fàcilment si s'aïlla la x en l'expressió del valor estandarditzat:

$$z = \frac{x - \bar{x}}{s_x} \xrightarrow{\text{implica}} x = \bar{x} + zs_x$$

d'on s'obté que el valor d'una observació determinada és igual a la mitjana més el producte del seu valor estandarditzat per la desviació típica.

Si la distribució de les dades és prou simètrica podem utilitzar la mitjana i la desviació típica per a descriure-la d'una manera ràpida. En aquest cas, l'ús dels valors estandarditzats ens permet d'entendre molt ràpidament on es troba una dada en relació amb la seva mitjana i comparar molt fàcilment dades que pertanyen a conjunts d'observacions diferents.

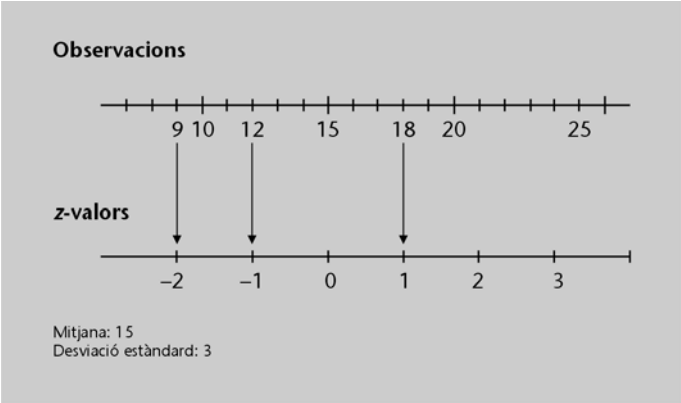
El concepte *estandardització* és clau en l'estudi de les variables aleatòries i del model normal.

Exemple d'estandardització de variables estadístiques

Suposem un conjunt de dades amb mitjana 15 i desviació típica 3. Si una observació té un valor estandarditzat de 3, aleshores:

$$z = \frac{x - \bar{x}}{s_x} = 3 \longrightarrow x = \bar{x} + 3s_x = 15 + 3 \cdot 3 = 24$$

Si el valor d'una observació és $x = 9$, aleshores el seu valor estandarditzat és –2, que indica que està a –2 desviacions típiques de la mitjana de la distribució. Observeu aquestes situacions en el gràfic següent:



Les notes d'àlgebra i estadística

Les notes dels exàmens d'àlgebra i estadística de cinc alumnes han estat les que es mostren a continuació:

Àlgebra	1	2	6	4	5
Estadística	5	7	6	8	9

Suposem que un alumne ha tret un 6 en les dues matèries; en quina ha tret millor nota? Evidentment un 6 és un 6, però ara ens plantegen com està situat aquest 6 en relació amb la resta de notes en cadascuna de les matèries. Si anomenem x_i les notes d'àlgebra i y_i, les notes d'estadística, la mitjana de les notes d'àlgebra és $\bar{x} = 3{,}6$ i la desviació típica és $s_x = 1{,}8547$, mentre que en el cas d'estadística la mitjana és $\bar{y} = 7$ i la desviació típica és $s_y = 1{,}4142$. Així, doncs, en relació amb les notes d'àlgebra, un 6 té un valor estandarditzat de $(6 - 3{,}6) / 1{,}8547 = 1{,}29$ mentre que en relació amb l'estadística, el mateix 6 té un valor estandarditzat de $(6 - 7) / 1{,}4142 = -0{,}7071$ (un valor estandarditzat negatiu indica que som a l'esquerra de la mitjana). Per tant, treure un 6 en àlgebra equival a ser 1,29 desviacions típiques per sobre de la mitjana de la nota d'àlgebra, mentre que treure un 6 en estadística equival a situar-se 0,7071 desviacions típiques per sota de la mitjana (el valor estandarditzat és negatiu). D'aquí podem concloure que en relació amb aquests dos conjunts de notes, destaca més un 6 en àlgebra que un 6 en estadística.

3. Usos de la mitjana i la desviació típica o de la mediana i els cinc nombres resum

Fins ara hem vist diferents maneres de descriure amb nombres i gràfics els diversos valors que pot prendre una variable (la distribució de la variable); ara ens podem preguntar quan convé fer servir cadascuna d'aquestes maneres. Com podeu suposar, no podem donar regles fixes. Atès que és ben difícil d'argumentar l'ús dels resums numèrics, veurem què opina sobre el tema un important autor:

En castellà *mitjana* és *media* i *mediana* és *mediana*.

"Una distribución asimétrica con unas pocas observaciones en la cola larga de la distribución tendrá una desviación típica grande. En tal caso, la desviación estándar no proporciona una información demasiado útil. Como en una distribución muy asimétrica la dispersión de las colas es muy distinta, es imposible describir bien la dispersión con un solo número. Los cinco números-resumen, con los dos cuartiles y los dos valores extremos, proporcionan una información mejor. Es preferible utilizar los cinco números-resumen en lugar de la media y la desviación típica para describir una distribución asimétrica. Utiliza la media y la desviación típica sólo para distribuciones razonablemente simétricas. [...] Recuerda que la mejor visión global de una distribución la da un gráfico. Las medidas numéricas de centro y de dispersión reflejan características concretas de una distribución, pero no describen completamente su

forma. Los resúmenes numéricos no detectan, por ejemplo, la presencia de múltiples picos ni de espacios vacíos. [...] REPRESENTA SIEMPRE TUS DATOS GRÁFICAMENTE."

D.S. Moore, "Estadística aplicada básica".

4. Variància i dades tabulades

De la mateixa manera que s'ha fet amb la mitjana, cal adaptar les fórmules de càlcul de la variància al cas en què les dades estiguin en forma de taula, ja sigui de freqüències relatives o d'absolutes. Així, farem servir les fórmules següents:

$$s^2 = \frac{\sum_{i=1}^{k} n_i(x_i - \bar{x})^2}{N} = \sum_{i=1}^{k} f_i(x_i - \bar{x})^2 \text{ o bé } s^2 = \frac{\sum_{i=1}^{k} n_i(m_i - \bar{x})^2}{N} = \sum_{i=1}^{k} f_i(m_i - \bar{x})^2$$

en cas que les dades estiguin agrupades en intervals amb marca de classe m_i.

Exemple de càlcul de la variància a partir de dades tabulades

En l'exemple dels virus atacants (I) en què es calcula el nombre de virus que han atacat cadascun dels diferents ordinadors de la nostra empresa durant l'any 2000, la mitjana val 4,16. Per a calcular la variància ens pot ajudar la taula següent:

x_i	n_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_i \cdot (x_i - \bar{x})^2$
0	10	–4,16	17,3056	173,056
5	15	0,84	0,7056	10,584
9	6	4,84	23,4256	140,5536
Totals	**31**			**324,1936**

I, per tant, la variància és 324,1936 / 31 = 10,45.

5. Resum

En aquesta sessió s'han presentat diverses mesures que permeten d'estudiar la distribució de les dades al voltant d'una mesura de centre.

Si descrivim el centre amb la mediana, podem complementar la informació que aporta amb els quartils i el rang interquartílic; en aquest cas resulta especialment interessant el diagrama de caixa.

La variància i la desviació típica ajuden a entendre com es distribueixen les dades al voltant de la mitjana; per això cal estudiar les propietats de la variància i la desviació típica, entre les quals destaca la regla de Txebixev.

Finalment, es defineix el concepte d'**estandardització de dades**, que dóna el nombre de desviacions típiques que un valor dista de la mitjana.

Exercicis

1. S'ha mesurat el temps en segons que triga a arrencar l'última versió del programa Macrohard Phrase en els ordinadors de la nostra empresa, segons el sistema operatiu amb què funciona. Els resultats han estat els següents:

- En els ordinadors equipats amb Doors95: 27, 25, 50, 33, 25, 86, 28, 31, 34, 36, 37, 44, 20, 59 i 85 segons.

- En els ordenadors equipats amb Doors98: 33, 7, 25, 14, 5, 31, 19, 10, 29 i 18 segons.

Calculeu els cinc nombres resum i la mitjana de la distribució corresponent al temps que el programa triga a arrencar i dibuixeu alguns gràfics que us semblin rellevants per a comparar el temps que el programa triga a arrencar segons el sistema operatiu. A partir d'aquests gràfics, compareu el comportament del programa segons el sistema operatiu i expliqueu si creieu que hi ha diferència entre fer servir Doors95 i Doors98.

2.

a) Construïu una llista de 10 nombres tals que *Mínim* = 2, *Màxim* = 20, *Primer quartil* = 5, *Mediana* = 10 i *Tercer quartil* = 19.

b) Igual que abans però amb la mitjana = 11.

c) En les condicions del punt **a**, la mitjana podria ser igual a 21? (Recordeu que la suma de les desviacions respecte de la mitjana ha de ser 0.)

3. Les dades següents corresponen al nombre de vegades que el programa MiniHard Word es va "penjar" durant un mes en cadascun dels 50 ordinadors de la nostra empresa. Comproveu que se satisfà la regla de Txebixev per a $m = 1$, $m = 2$ i $m = 3$.

0	9	12	14	19
2	10	12	14	20
4	10	12	14	20
5	10	12	15	21
6	11	12	15	22
6	11	12	17	29
6	11	12	17	29
7	11	13	18	32
8	11	13	18	39
9	12	14	19	39

Solucionari

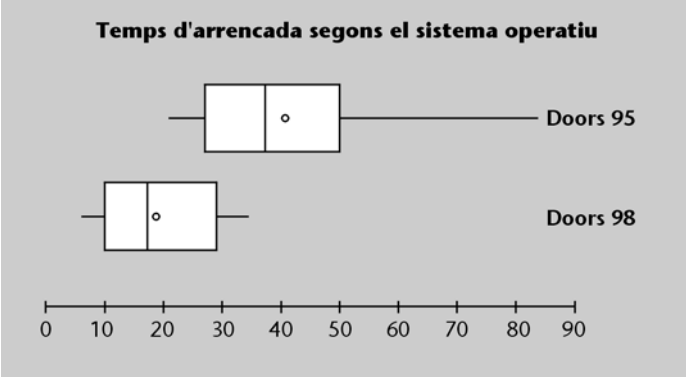
1. Les dades ordenades són:

- Doors95: 20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86
- Doors98: 5, 7, 10, 14, 18, 19, 25, 29, 31, 33

Aquesta taula registra els cinc nombres resum i la mitjana:

Sistema	Mitjana	Mínim	Q1	Q2	Q3	Màxim
Doors95	41,33	20	27	34	50	86
Doors98	19,1	5	10	18,5	29	33

Si representem simultàniament i amb la mateixa escala els diagrames de caixa corresponents a les dues distribucions, obtenim el gràfic següent, on també s'han marcat amb cercles les mitjanes corresponents:



A partir dels resums numèrics i dels gràfics podem concloure que Doors98 triga molt menys a arrencar que Doors95; cal destacar els aspectes següents:

- Tant la mitjana com la mediana de Doors98 són aproximadament la meitat que les de Doors95.
- La mediana de Doors98 és més petita que el mínim de Doors95; això vol dir que la meitat de les vegades Doors98 triga menys a arrencar que la vegada que més ràpid ha arrencat Doors95.
- El màxim de Doors98 és més petit que la mediana (és a dir, la meitat de les vegades que arrenquem Doors95 triga més que la vegada que més ha trigat amb Doors98).
- El diagrama de caixa de Doors98 és molt més simètric i no té cap cua cap a valors allunyats o més grans dels normals. Aquesta simetria suggereix que Doors98 té un comportament més regular i més estable que Doors95.

2.

a) Suposem que tenim una llista ordenada de 10 nombres. La mediana ocuparà la posició 10 + 1/2 = 5,5 i, per tant, es calcularà com a valor mitjà de la 5a. i 6a. observacions. En aquestes circumstàncies, el primer quartil ha de ser la 3a. observació i el tercer quartil la 8a. observació, tal com es veu en aquesta taula, on també escrivim el mínim i el màxim (M1 i M2 indiquen les observacions que es necessiten per a calcular la mediana).

Posició	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Mínim		Q1		M1	M2		Q3		Màxim
	2		5					19		20

Observem que en aquesta taula el màxim, el mínim i el primer i tercer quartils estan fixats. Podem acabar d'omplir la taula de diferents maneres, sempre que les observacions apareguin ordenades i que la mediana tingui el valor 10; per exemple:

Posició	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Mínim		Q1		M1	M2		Q3		Màxim
	2	3	5	9	10	10	18	19	19	20
	2	2	5	6	8	12	13	19	19,5	20
	2	3	5	5	5	15	18	19	20	20

b) Si la mitjana ha de ser 11, hem de tenir cura que la suma dels valors sigui igual a 10 × 11 = 110. Atès que les posicions 1, 3, 8 i 10 tenen el valor fixat i la suma dels valors 5è. i 6è. ha de ser 20, la suma dels valors que apareixen en les posicions 2, 4, 7 i 9 ha de ser 110 – 20 – 2 – 5 – 19 – 20 = 44. Així, per exemple, podem fer:

Posició	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	Mínim		Q1					Q3		Màxim
	2	3	5	5	8	12	17	19	19	20

c) La mitjana no pot superar l'observació més gran. Una possible raó és que si ho fos, totes les desviacions respecte de la mitjana serien estrictament negatives i, per tant, la suma de les desviacions no podria ser 0.

3. Si calculem la mitjana i la desviació típica obtenim $\bar{x} = 14{,}28$; $\sigma_x = 8{,}11$; amb la qual cosa els intervals interessants, les proporcions d'observacions en

cada interval i el mínim predit per la regla de Txebixev es poden veure en la taula següent:

Primer es calculen els intervals substituint els valors de *m*, $\bar{x}$ i s_x.

m	Interval	Observacions en l'interval	Percentatge d'observacions en l'interval	Percentatge mínim previst per la regla de Txebixev
1	$(\bar{x}-1s_x,\ \bar{x}+1s_x)=$ $=(6{,}18,\ 22{,}38)$	38	38/50 = 76%	0%
2	$(\bar{x}-2s_x,\ \bar{x}+2s_x)=$ $=(-1{,}92,\ 30{,}48)$	47	47/50 = 94%	75%
3	$(\bar{x}-3s_x,\ \bar{x}+3s_x)=$ $=(-10{,}03,\ 38{,}56)$	48	48/50 = 96%	88%

El percentatge d'observacions...

... en l'interval es calcula comptant quantes observacions cauen dins l'interval.

Veiem com en tots els casos tenim més observacions en els intervals assenyalats que les previstes per la regla. Recordeu que la regla diu quins són els percentatges més baixos que podem trobar en cada interval.