Estadística descriptiva univariant

Models estadístics per a la descripció de dades univariants

Alicia Vila, Ángel A. Juan i Patricia Carracedo

PID_00242445

Temps de lectura i comprensió: **4 hores**

UOC
Universitat Oberta de Catalunya

Índex

Introducció

Les societats modernes i les modernes administracions són riques en dades que requereixen ser processades i analitzades per a ser convertides en informació rellevant. En el treball de les administracions s'han de determinar les necessitats de la societat, avaluar les polítiques públiques que es duen a terme o el funcionament de les organitzacions responsables de dur-les a terme. Això converteix l'estadística en una ciència interessant i útil, ja que proporciona estratègies i eines que permeten obtenir informació a partir de les dades esmentades. A més, gràcies a l'evolució de la tecnologia (ordinadors i programari estadístic) avui en dia és possible automatitzar gran part dels càlculs matemàtics associats a l'ús de tècniques estadístiques, la qual cosa permet estendre'n l'ús a un gran rang de professionals en àmbits tan diversos com la biologia, les ciències empresarials, la sociologia o les ciències de la informació.

La pràctica de l'estadística requereix aprendre a obtenir i explorar les dades –tant numèricament com mitjançant gràfics–, a pensar sobre el context de les dades i el disseny de l'estudi que els ha generat, a considerar la possible influència d'observacions anòmales en els resultats obtinguts, a discutir la legitimitat dels supòsits requerits per cada tècnica i, finalment, a validar la fiabilitat de les conclusions derivades de l'anàlisi. L'estadística requereix tant de coneixements sobre els conceptes i tècniques emprats com de la suficient capacitat crítica que permeti avaluar la conveniència d'usar unes tècniques o unes altres segons el tipus de dades disponible i el tipus d'informació que es vol obtenir.

En aquest mòdul inicial de l'assignatura, s'examinen les dades procedents d'una única variable: en primer lloc s'explica com organitzar i resumir les dades esmentades, tant numèricament com gràficament (estadística descriptiva); en segon lloc, s'introdueixen els conceptes bàsics associats amb la idea de probabilitat; finalment, es presenten alguns models matemàtics que permeten analitzar el comportament d'algunes variables.

Objectius

Els objectius acadèmics que es plantegen en aquest mòdul són els següents:

1. Entendre la importància de l'estadística en la societat moderna i especialment en l'activitat de les administracions públiques.

2. Aprendre a organitzar i resumir un conjunt de dades procedents d'una variable mitjançant gràfics, taules de freqüències i estadístics descriptius.

3. Comprendre el concepte de probabilitat d'un esdeveniment i descobrir-ne les principals propietats i aplicacions.

4. Conèixer les principals distribucions estadístiques que s'usen per a modelitzar el comportament de variables discretes i contínues.

5. Saber calcular probabilitats associades a cada una de les distribucions introduïdes.

6. Aprendre a usar programari estadístic i d'anàlisi de dades com a instrument bàsic en l'aplicació pràctica dels conceptes i les tècniques estadístiques.

1. Introducció a l'estadística

L'estadística és la ciència que s'ocupa d'obtenir dades i processar-les per a transformar-les en informació. És, per tant, un llenguatge universal àmpliament utilitzat en les ciències socials, en les ciències experimentals, en les ciències de la salut i en les enginyeries. Les tecnologies de la informació i la comunicació (TIC) han incrementat notablement la producció, disseminació i tractament de la informació estadística. En particular, Internet és una font inesgotable de dades que poden oferir informació i, a partir d'aquesta, coneixement. D'altra banda, la constant evolució dels ordinadors personals i del **programari estadístic** i d'anàlisi de dades possibilita i facilita l'anàlisi de grans quantitats de dades mitjançant l'ús de tècniques estadístiques i de mineria de dades. En la societat de la informació es fa, doncs, imprescindible disposar d'un cert coneixement estadístic, fins i tot per a poder comprendre i interpretar correctament els indicadors econòmics (IPC, inflació, taxa de desocupació, euríbor, etc.), o els indicadors socials (esperança de vida, índex d'alfabetització, índex de pobresa, indicador social de desenvolupament sostenible, etc.), que són freqüentment referenciats en els mitjans de comunicació.

Nota

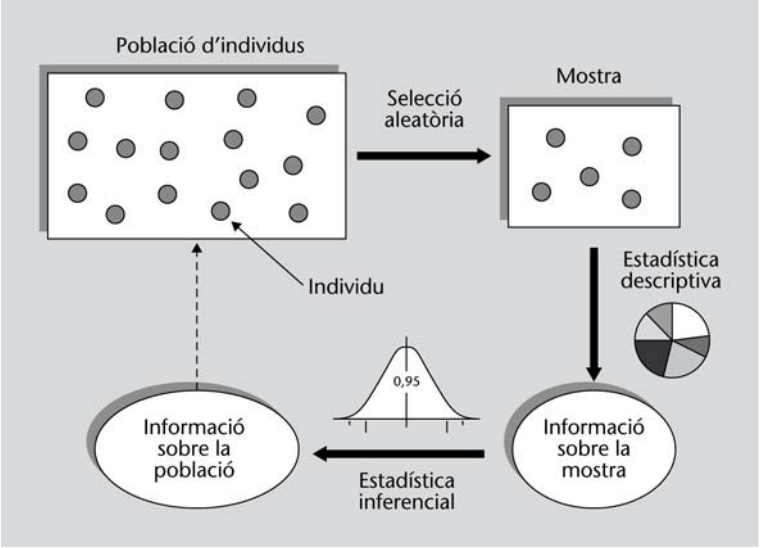
Les agències governamentals, com l'Institut Nacional d'Estadística (INE) o Eurostat proporcionen dades sobre gairebé qualsevol àmbit socioeconòmic.

Programari estadístic

Actualment hi ha excel·lents **programes informàtics** per a l'anàlisi estadística de dades. En són alguns exemples: MINITAB, SPSS, MS Excel, SAS, R, S-Plus, Statgraphics o Statistica.

El camp de l'estadística es pot dividir en dues grans àrees: l'estadística descriptiva i l'estadística inferencial (figura 1).

Figura 1. Estadística descriptiva i estadística inferencial



L'estadística descriptiva s'ocupa de l'obtenció, presentació i descripció de dades procedents d'una mostra o subconjunt d'una població d'individus. Per la seva part, l'estadística inferencial utilitza els resultats obtinguts, mitjançant l'aplicació de les tècniques descriptives a una mostra, per a inferir informació sobre el total de la població a la qual pertany la mostra esmentada.

Alguns termes bàsics

Al llarg d'aquest material s'usaran abundants termes estadístics, molts dels quals força coneguts. S'introdueixen i revisen a continuació alguns d'aquests termes bàsics que convé entendre bé:

- **Població:** col·lecció o conjunt d'elements (individus, objectes o esdeveniments) les propietats dels quals es volen analitzar. Exemples: (a) els estudiants universitaris d'un país; (b) el conjunt dels municipis d'una regió administrativa; (c) el conjunt dels treballadors d'un ajuntament, etc.

- **Mostra:** qualsevol subconjunt d'elements de la població. Exemples: (a) els estudiants d'una determinada universitat; (b) els municipis costaners de la regió administrativa; (c) els treballadors de l'Ajuntament amb coneixements d'anglès, etc.

- **Mostra aleatòria:** mostra els elements de la qual han estat escollits de forma aleatòria. Exemples: (a) un subconjunt de 200 estudiants escollits a l'atzar (mitjançant l'ús de nombres aleatoris) d'entre tots els matriculats en universitats d'un país; (b) un subconjunt de 50 municipis de la regió escollits a l'atzar; (c) un subconjunt de 15 treballadors de l'Ajuntament escollits a l'atzar, etc.

- **Marc del mostreig:** llista que conté els elements de la població candidats a ser seleccionats en la fase de mostreig. No necessàriament coincidirà amb tota la població d'interès, ja que a vegades no serà possible identificar tots els elements de la població. Exemples: (a) llista de tots els estudiants matriculats en universitats d'un país en un semestre concret; (b) municipis de la regió administrativa que disposen de correu electrònic; (c) tots els treballadors en el registre de nòmines de l'Ajuntament, etc.

- **Variable aleatòria:** característica d'interès associada a cada un dels elements de la població o mostra considerada. Exemples: (a) l'edat de cada estudiant; (b) el nombre de visites diàries que rep la pàgina web de l'Ajuntament; (c) valoració de la qualitat en el lloc de treball per part de cada treballador de l'Ajuntament, etc.

- **Dades o observacions:** conjunt de valors obtinguts per a la variable d'interès en cada un dels elements de la mostra. Exemples: (a) les edats registrades són {25, 23, 19, 28...}; (b) les visites diàries registrades són {1326, 1792, 578, 982 ...}; (c) les valoracions de la qualitat del lloc de treball són {7,5; 5,8; 9,3...}.

- **Experiment:** estudi en què l'investigador en controla o modifica expressament les condicions amb la finalitat d'analitzar els diferents patrons de resposta en les observacions. Exemples: (a) estudiar com varien les

qualificacions d'un grup d'estudiants segons disposin o no d'ordinadors amb accés a Internet a les aules; (b) estudiar com varia el nombre de visites a un diari en línia segons el tipus d'informació i d'enllaços que s'incloguin a la portada del web; (c) estudiar com canvia la valoració dels treballadors de la qualitat del lloc de treball quan es flexibilitza l'horari laboral , etc.

- **Inspecció/enquesta:** estudi en el qual l'investigador no pretén modificar les condicions de la mostra respecte a la variable d'interès, sinó simplement obtenir les dades corresponents a unes condicions estàndards. Exemples: (a) registrar les qualificacions dels estudiants d'un màster determinat; (b) realitzar una enquesta als visitants del lloc web de l'Ajuntament; (c) obtenir la valoració de la qualitat del lloc de treball dels treballadors de l'Ajuntament, etc.

- **Paràmetre:** valor numèric que sintetitza alguna propietat determinada de la població. Els paràmetres s'associen a tota la població i se solen representar per lletres de l'alfabet grec com μ (mu), σ (sigma), etc. Exemples: (a) l'edat mitjana de tots els estudiants universitaris d'un país; (b) el nombre màxim de visites diàries rebut per una web municipal; (c) el rang o diferència entre la màxima qualificació i la mínima qualificació del lloc de treball dels treballadors de l'Ajuntament, etc.

- **Estadístic:** valor numèric que sintetitza alguna propietat determinada d'una mostra. Els estadístics s'associen a una mostra i se solen representar per lletres de l'alfabet llatí com ara $\overline{x}$, s, etc. Exemples: (a) l'edat mitjana dels estudiants d'una mostra aleatòria; (b) el nombre màxim de visites diàries rebudes per una web municipal; (c) el rang o la diferència entre qualificacions superior i inferior del lloc de treball segons els treballadors de l'Ajuntament, etc.

- **Variable qualitativa o categòrica:** variable que categoritza o descriu qualitativament un element de la població. És una etiqueta que identifica els elements de la població que tenen una mateixa propietat. Encara que l'etiqueta sigui numèrica, no té sentit usar-la en operacions aritmètiques. Exemples: (a) el telèfon o l'adreça electrònica d'un estudiant; (b) el color dominant emprat a la pàgina principal del web municipal; (c) el barri de residència del treballador municipal, etc.

- **Variable quantitativa o numèrica:** variable que quantifica alguna propietat d'un element de la població. És possible realitzar-hi operacions aritmètiques. Exemples: (a) l'import de la beca que rep un estudiant; (b) el cost de manteniment del lloc web de l'Ajuntament; (c) el nombre de dones que treballen en un Ajuntament, etc.

- **Variable quantitativa discreta:** variable quantitativa que pot prendre un nombre finit o comptable de valors diferents. Exemples: (a) nombre d'as-

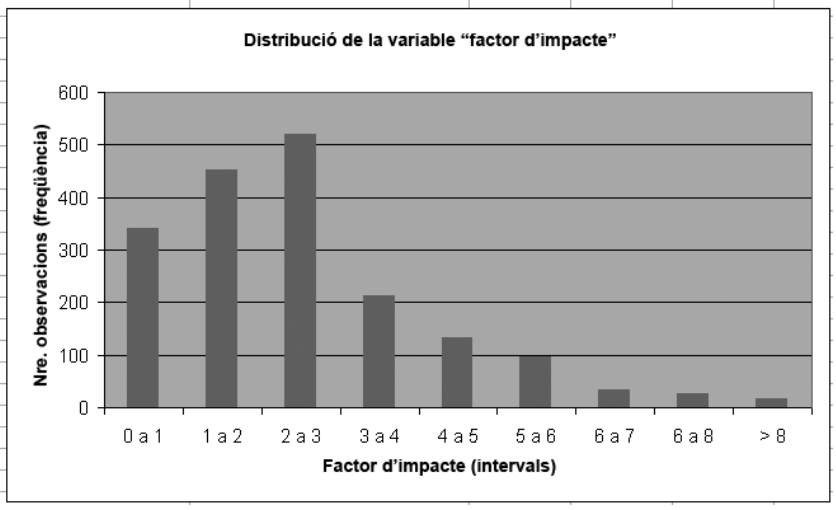
signatures a les que s'ha matriculat un estudiant; (b) nombre d'enllaços a altres fonts d'informació que ofereix el lloc web de cada Ajuntament; (c) nombre de fills de cada treballador de l'Ajuntament, etc.

- **Variable quantitativa contínua:** variable quantitativa que pot prendre un nombre infinit (no comptable) de valors diferents. Exemples: (a) alçada o pes d'un estudiant; (b) temps que transcorre entre la publicació d'una enquesta en línia i l'instant en què aquesta ha estat resposta per un centenar d'internautes; (c) factor d'impacte (sense arrodonir) d'una revista (que és la mesura de la importància de cada revista acadèmica en el seu camp de coneixement), etc.

- **Distribució d'una variable:** en sentit ampli, una distribució és una taula, gràfic o funció matemàtica que explica com es comporten o distribueixen els valors d'una variable, *i. e.*: quins valors pren la variable així com la freqüència d'aparició de cada un. Exemple: donada una mostra aleatòria de revistes acadèmiques, la distribució de la variable "factor d'impacte d'una revista" pot representar-se mitjançant una taula de freqüències o mitjançant una gràfica com s'aprecia a la figura 2. S'observa que 342 de les revistes considerades tenen un factor d'impacte entre 0 i 1,452 de les revistes tenen un factor d'impacte entre 1 i 2, etc.

Figura 2. Distribució d'una variable aleatòria

	A	B	C	D	E	F
1	**Factor d'impacte**	**Nre. observacions (freqüència)**				
2	0 a 1	342				
3	1 a 2	452				
4	2 a 3	521				
5	3 a 4	212				
6	4 a 5	132				
7	5 a 6	97				
8	6 a 7	35				
9	6 a 8	27				
10	> 8	16				



2. Descripció de dades mitjançant taules i gràfics

Quan es disposa d'un conjunt d'observacions procedents d'una mostra, convé realitzar-ne una primera anàlisi exploratòria mitjançant gràfics i taules que ajudin a interpretar les dades i a extreure'n informació. Hi ha diferents tipus de gràfics que es poden usar en aquesta fase exploratòria, i l'ús dels uns o els altres dependrà del tipus de dades de què es disposi (qualitatives o quantitatives), així com de la informació que es vulgui visualitzar. En aquest apartat es presenten alguns dels gràfics i taules més habituals per a la descripció de **dades univariants**.

Dades univariants

Les dades univariants són les que provenen d'una única variable. En alguns casos, les dades poden procedir de dues o més varibles, i aleshores s'usa l'expressió bivariant (si es tracta de dues variables) o multivariant (si se'n consideren més de dues).

Gràfics i taules per a dades qualitatives o categòriques

Si es disposa de dades qualitatives o categòriques, aquestes es poden sintetitzar mitjançant una taula que reculli, per a cada categoria: el nombre de vegades que hi apareix (freqüència absoluta), el percentatge d'aparicions sobre el total d'observacions (freqüència relativa), i els acumulats d'ambdós valors. La taula 1 mostra aquesta informació per a la variable *nombre de connexions Wi-Fi* (*hotspots*) *identificats en cada comunitat autònoma*.

Taula 1. Exemple de taula de freqüències per a una variable categòrica

CA	Connexions Wi-Fi per CA			
	Freqüència	Freqüència acumulada	Freqüència relativa	Freq. rel. acumulada
Andalusia	885	885	11,9%	11,9%
Aragó	177	1.062	2,4%	14,2%
Astúries	148	1.210	2,0%	16,2%
Cantàbria	164	1.374	2,2%	18,4%
Castella - la Manxa	144	1.518	1,9%	20,3%
Castella - Lleó	302	1.820	4,0%	24,4%
Catalunya	1.391	3.211	18,6%	43,0%
C. Valenciana	622	3.833	8,3%	51,3%
Extremadura	137	3.970	1,8%	53,2%
Galícia	516	4.486	6,9%	60,1%
I. Balears	183	4.669	2,5%	62,5%
I. Canàries	151	4.820	2,0%	64,6%
La Rioja	126	4.946	1,7%	66,3%
Madrid	1.776	6.722	23,8%	90,0%
Múrcia	160	6.882	2,1%	92,2%
Navarra	153	7.035	2,0%	94,2%
País Basc	430	7.465	5,8%	100,0%
Totals	**7.465**		**100,0%**	

Nota

Observeu que la **freqüència acumulada** s'obté simplement acumulant freqüències anteriors.

A més, mitjançant una taula de freqüències, sol ser habitual representar dades categòriques mitjançant l'ús de gràfics circulars (figura 3) o bé mitjançant diagrames de barres (figura 4).

Figura 3. Exemple de gràfic circular per a una variable categòrica

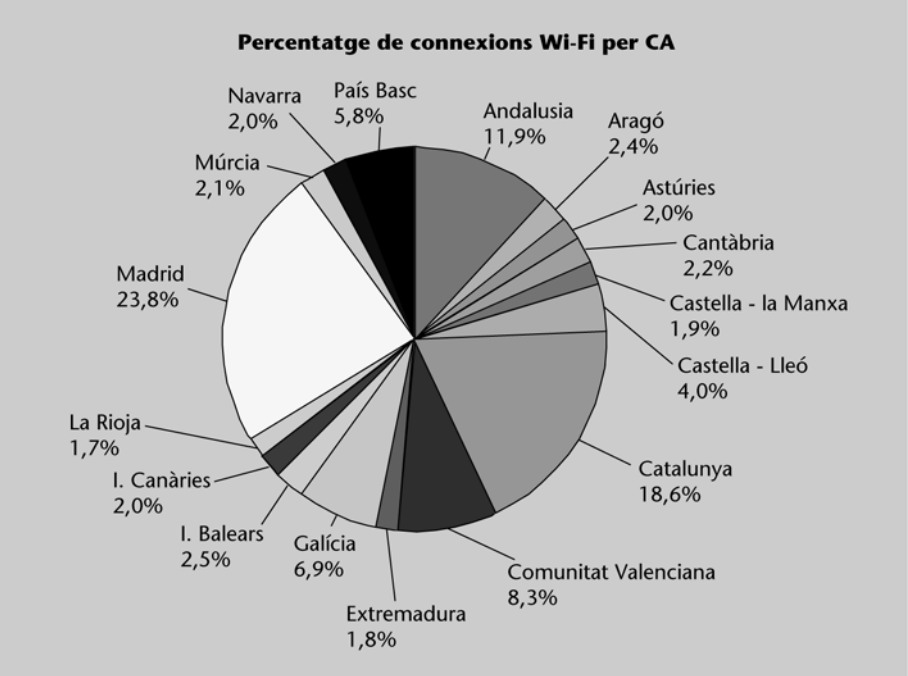
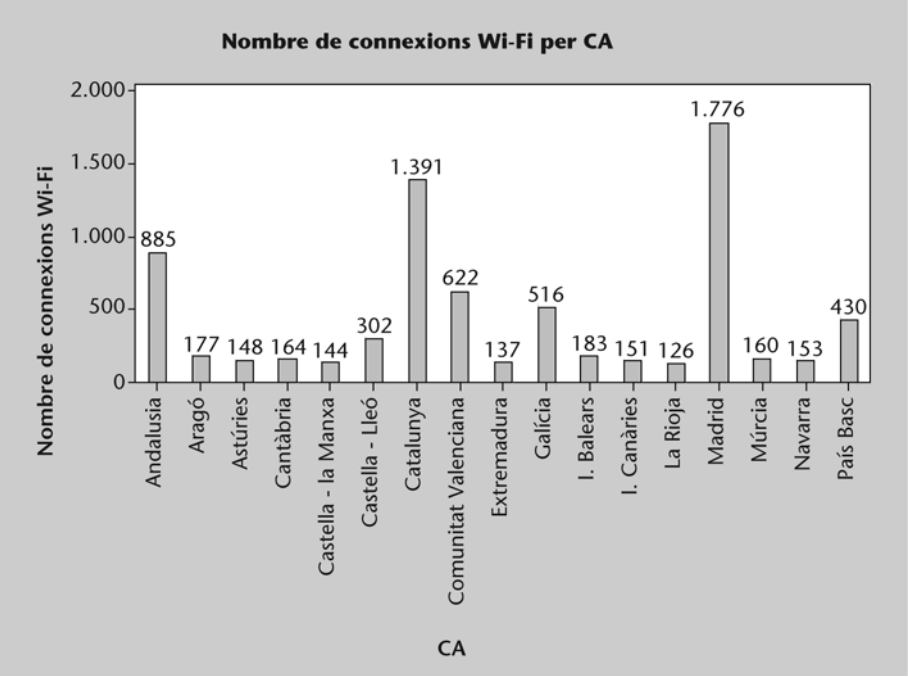
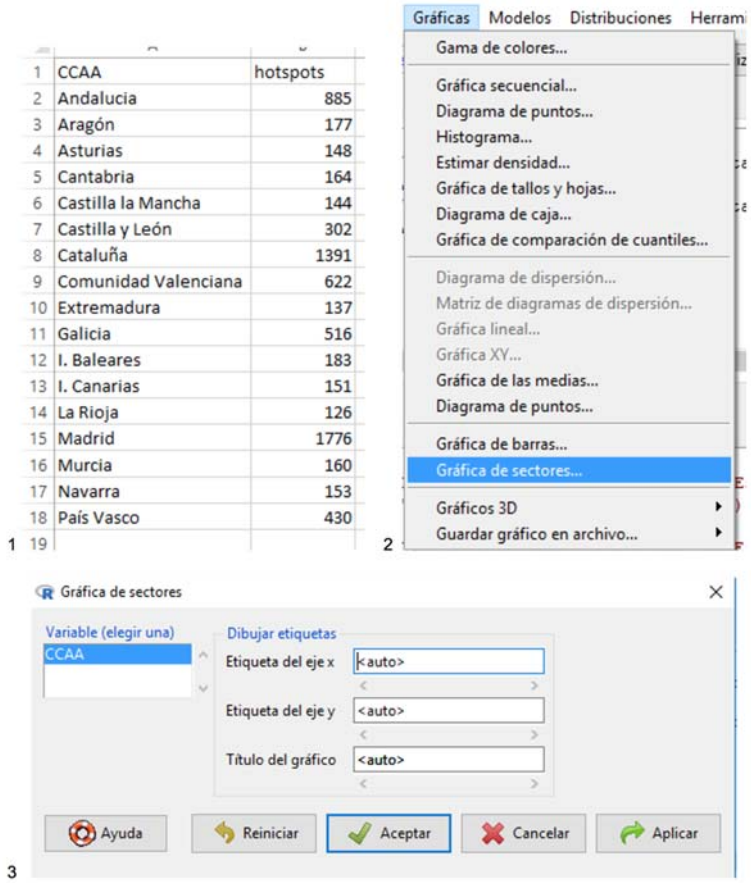


Figura 4. Exemple de diagrama de barres per a una variable categòrica



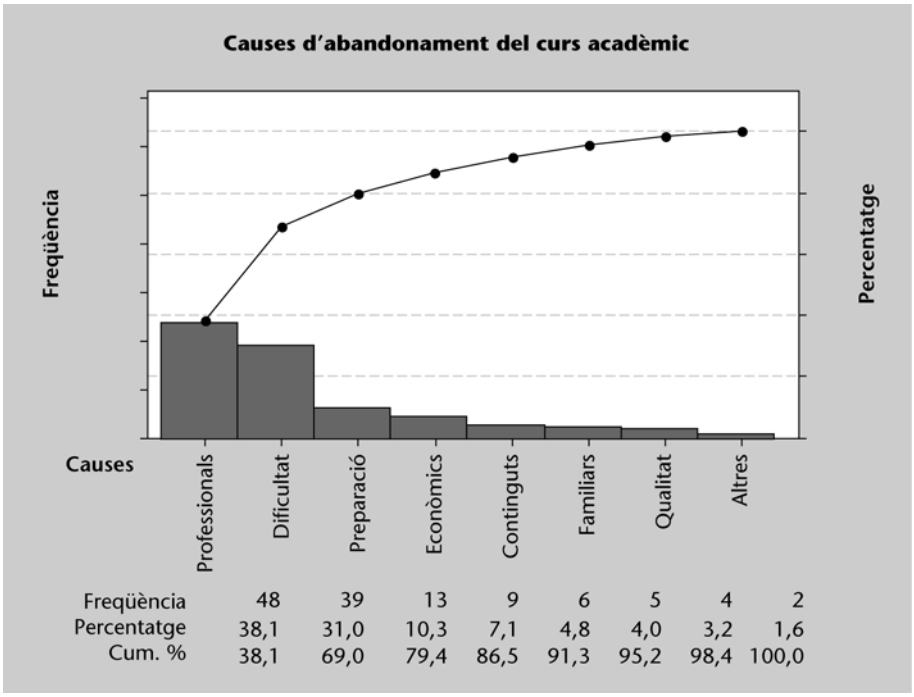
Aquest tipus de gràfics es poden obtenir fàcilment amb qualsevol programari estadístic o d'anàlisi de dades (per exemple: R, Minitab, MS Excel, SPSS, etc.). La figura 5 mostra els passos bàsics per a generar un gràfic circular (*pie chart*) mitjançant R Commander. La generació d'un diagrama de barres (*bar chart*) s'aconsegueix de manera similar, igual com ocorre amb la majoria dels gràfics que s'introdueixen en aquest apartat.

Figura 5. Passos a seguir per a la generació d'un gràfic circular amb R Commander



Un gràfic que sol ser també força utilitzat per a descriure dades qualitatives és l'anomenat diagrama de Pareto. Aquest gràfic és compost per: (a) un diagrama de barres en el qual les categories són ordenades de major a menor freqüència, i (b) una línia que representa la freqüència relativa acumulada (figura 6).

Figura 6. Diagrama de Pareto sobre les causes d'abandonament d'un curs



Passos a seguir

Una vegada introduïdes les dades en el programa **(1)**, se segueix la ruta *Gráficos* > *Gráfica de sectores* **(2)** i se seleccionen les variables a la finestra corresponent **(3)**.

Nota

Les captures de pantalla de R corresponen a la **versió 3.2.3** (2015-12-10) d'aquest programa. És possible que altres versions ofereixin lleugeres diferències en els menús i finestres, tot i que bàsicament el procés serà el mateix. Per a obtenir més detalls sobre les opcions disponibles, sempre és possible consultar l'ajuda en línia del programa o bé alguns dels nombrosos manuals d'ús que es poden trobar a Internet.

Diagrama de Pareto

Per a generar un diagrama de Pareto en R Commander es fa servir la *llibrería qcc* i la funció *pareto.chart*.

Els diagrames de Pareto són molt útils per a detectar quan un percentatge reduït de categories (per exemple, un 20% de les categories) acapara o representa un percentatge alt d'observacions (per exemple, un 80% de les dades). Aquests fenòmens d'excessiva representativitat per part d'unes quantes categories solen donar-se amb freqüència en contexts socioeconòmics (per exemple, un percentatge reduït dels ciutadans d'un país acapara un alt percentatge de la renda), educatius (per exemple, un percentatge reduït de causes generen la major part dels abandonaments del curs) o d'enginyeria de la qualitat (per exemple, un alt percentatge d'errors són deguts a un nombre molt reduït de causes). Identificar les poques categories que representen una gran part del percentatge total pot servir per corroborar certs desequilibris distributius –com una distribució poc equilibrada de les rendes en un país o dels sous en una empresa–, o per a proporcionar pistes sobre els principals factors de causa d'un problema –com l'alt nivell d'abandonament en un curs o un elevat nivell d'errors en un servei o producte–.

Gràfics i taules per a dades quantitatives

En el cas de dades quantitatives, la representació gràfica o les taules permeten apreciar la forma de la distribució estadística, com ara, la forma com es comporta la variable d'interès (quins són els valors mitjans o centrals, quins són els valors més habituals, com varia, com de dispersos són els valors, si mostra algun patró de comportament especial, etc.).

Un dels gràfics més senzills d'elaborar és l'anomenat gràfic de punts (*dotplot*). Es tracta d'un gràfic en el qual cada punt representa una o més observacions. Els punts s'apilen l'un sobre l'altre quan es repeteixen els valors observats (figura 7).

Figura 7. Gràfic de punts per a les qualificacions d'un curs

Un gràfic similar, encara que una mica més elaborat i amb una orientació traslladada dels eixos, és l'anomenat diagrama de tiges i fulles (*stem-and-leaf*). S'hi representen també els valors observats, però usant els valors numèrics en lloc de punts, la qual cosa proporciona un nivell de detall més alt. La figura 8 mostra un exemple de gràfic de tiges i fulles per a les mateixes dades emprades a la figura 7. Observeu que el gràfic s'ha construït a partir d'una mostra de 50 qualificacions i que s'ha usat una unitat de fulla (*leaf*) de 0,1. Això significa que la segona columna del gràfic representa la

part sencera de la qualificació, mentre que cada un dels nombres situats a la dreta representa la part decimal d'una observació amb l'esmentada part sencera. Així, es poden llegir les qualificacions següents per ordre de menor a major: 2.5, 3.7, 4.0, 5.0, 6.0, 6.0, 6.0, etc.

Figura 8. Gràfic de fulles i tiges per a les qualificacions d'un curs

```
> with(Dataset, stem.leaf(A, na.rm=TRUE))
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 15
   1     2 | 5
   2     3 | 7
   3     4 | 0
   4     5 | 0
   7     6 | 000
  (2)    7 | 05
   6     8 | 02
   4     9 | 000
   1    10 | 0
```

Atenció

Observeu que en un gràfic de tiges i fulles les dades s'apilen d'esquerra a dreta, en lloc de dalt a baix, com passava amb el gràfic de punts.

Quan les observacions generen un nombre elevat de valors diferents, és recomanable agrupar els esmentats valors en classes o intervals disjunts d'igual mida. D'aquesta manera, cada observació és classificada en una classe o interval segons el seu valor. La taula 2 mostra un exemple de taula de freqüències en què s'han agrupat les dades en intervals. La freqüència de cada interval és determinada pel nombre d'observacions els valors de les quals estan en l'interval esmentat. La marca de classe representa el valor mitjà de l'interval.

Taula 2. Exemple de taula de freqüències agrupades usant intervals

Interval	Marca de classe	Freqüència	Freqüència relativa
[0, 2)	1	12	8,1%
[2, 4)	3	23	15,5%
[4, 6)	5	67	45,3%
[6, 8)	7	31	20,9%
[8, 10)	9	15	10,1%
Totals		**148**	**100,0%**

Un gràfic que utilitza també intervals per a agrupar les dades a representar és l'histograma, el qual mostra la freqüència (absoluta o relativa) de cada classe, cosa que permet visualitzar de manera aproximada la distribució de les dades (figura 9). Tanmateix, cal tenir present que la forma final de l'histograma pot variar bastant segons el nombre d'intervals que es defineixin per a agrupar les dades, la qual cosa de vegades no permet apreciar correctament la forma exacta de la distribució estadística que segueixen les observacions.

Nota

Una regla habitual és definir $\sqrt{n}$ classes o intervals, en què n és el nombre d'observacions disponibles.

Figura 9. Histograma d'una distribució aproximadament normal

La figura 9 mostra un histograma amb forma de campana: és una forma bastant simètrica, que presenta més altura a la part central i disminueix gradualment a les cues o extrems. Aquesta forma és bastant habitual i sol caracteritzar el comportament de moltes variables (com ara, notes numèriques en un examen, pes o alçada d'individus, temperatures diàries, etc.). Tanmateix, també serà habitual trobar variables que mostren patrons de comportaments completament diferents. Per exemple, la figura 10 mostra un histograma en el qual s'aprecia una distribució més "uniforme" o homogènia de les dades, mentre que la figura 11 mostra un histograma en el qual se n'aprecia una distribució asimètrica o "esbiaixada".

Figura 10. Histograma d'una distribució aproximadament uniforme

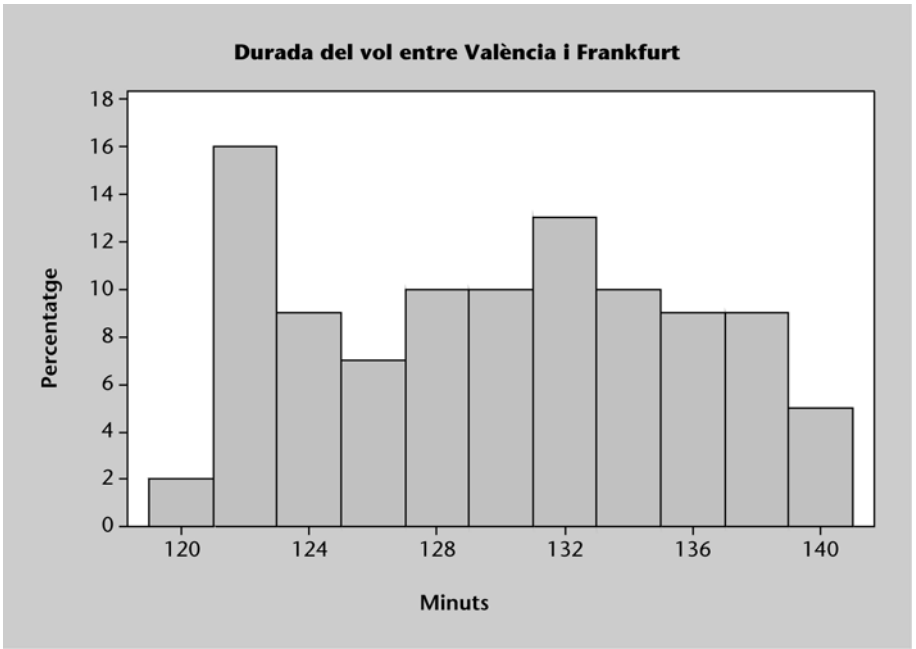
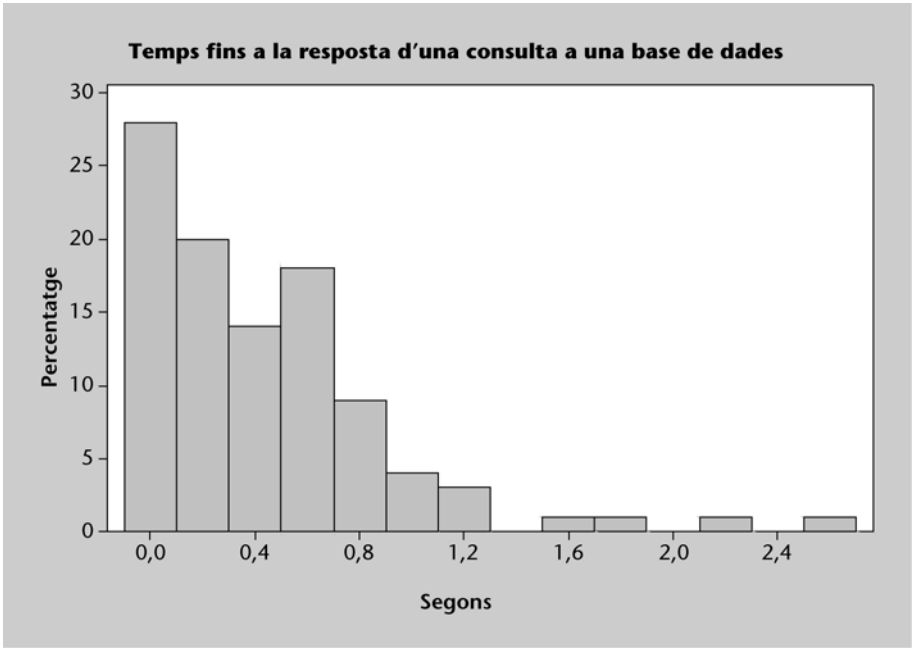


Figura 11. Histograma d'una distribució esbiaixada a la dreta



3. Descripció de dades mitjançant estadístics

Donat un conjunt de n dades o observacions, x_1, x_2 ..., x_n, associades a una variable d'interès, X, sol ser útil sintetitzar algunes de les propietats principals en uns quants valors numèrics. Els estadístics descriptius són, precisament, aquests valors numèrics capaços de proporcionar informació a partir del conjunt de les observacions. Els estadístics resulten molt útils a l'hora d'entendre el comportament de les dades, ja que un simple valor numèric és capaç de caracteritzar les propietats més rellevants d'una variable com, per exemple, el valor més representatiu del conjunt de dades, el valor màxim, el valor mínim, el valor que es repeteix amb més freqüència, un índex de dispersió o variabilitat, etc.

Web

Recordeu que el World Wide Web (per exemple, Wikipedia, etc.) és una font de consulta excel·lent per a ampliar les definicions i els conceptes estadístics que es proporcionen en aquest i altres mòduls. Un recurs especialment interessant, perquè ofereix una visió molt completa de tècniques i conceptes estadístics és el llibre en línia de StatSoft http://www.statsoft.com/textbook/.

Com ja es va comentar anteriorment, aquests estadístics fan referència a una mostra d'observacions i se solen representar mitjançant lletres de l'alfabet llatí ($\bar{x}$, s, etc.), la qual cosa permet distingir-los clarament dels seus paràmetres associats que sintetitzen propietats de tota la població i es representa mitjançant lletres gregues (μ, σ, etc.). Bàsicament, es poden distingir dos grups d'estadístics descriptius: (a) els de centralització, que proporcionen informació sobre quins són els valors centrals o més representatius del conjunt de dades (com ara, el valor mitjà de les dades) i (b) els de dispersió, que expliquen com se situen i varien les dades respecte als valors centrals ens informen de fins a quin punt els valors centrals són representatius (com ara, el rang o diferència entre el valor màxim i el valor mínim de les dades).

Nota

Recordeu que els símbols μ i σ es pronuncien com "mu" i "sigma", respectivament. La pronunciació d'altres símbols de l'alfabet grec es pot consultar, per exemple, a la Wikepedia.

Estadístics de centralització

A continuació es presenten els estadístics de centralització més comunament utilitzats:

- **Mitjana (*mean*):** la mitjana (també coneguda com a *valor mitjà* o *valor esperat*) d'un conjunt d'observacions mostrals es representa pel símbol $\bar{x}$. Intuïtivament, la mitjana simbolitza el "centre de masses" o "punt d'equilibri central" del conjunt de dades considerat. El paràmetre associat, la mitjana poblacional, és representada per μ. Per a calcular la mitjana d'un conjunt de dades s'usa l'expressió següent:

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

Mitjana mostral

Recordeu que la mitjana mostral és un **estadístic** que fa referència al "centre de masses" de les dades d'una mostra (subconjunt de la població), mentre que la mitjana poblacional és un **paràmetre** que representa el "centre de masses" de tota la població.

Exemple: la mitjana de les 5 dades següents {6, 3, 8, 6, 4} és

$$\bar{x} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5,4$$

- **Mediana (*median*):** la mediana d'un conjunt d'observacions mostrals sol ser representada pel símbol $\tilde{x}$. En el cas d'una població, el paràmetre me-

diana es denota per *M*. Una vegada s'ordenen tots les dades de menor a major, la mediana és el valor que deixa a l'esquerra la meitat de les observacions (és a dir, és el valor tal que el nombre d'observacions més petites que aquest coincideix amb el nombre d'observacions més grans que aquest). Els passos per a calcular la mediana són: (1) ordenar les dades de menor a major, (2) calcular la posició *i* que ocupa la mitjana en el conjunt ordenat de dades, $i = \frac{n+1}{2}$, i (3) seleccionar l'observació x_i (la qual ocupa la posició determinada en el pas anterior). Observar que si el nombre de dades *n* és senar (per exemple, $n = 5$), la posició *i* serà un valor sencer (per exemple, $i = 3$), que correspondrà amb un valor concret, x_i, del conjunt de dades. Tanmateix, si *n* és parell (per exemple, $n = 6$), la posició *i* serà un nombre no enter (per exemple, $i = 3{,}5$), i en aquest cas la mediana serà donada per la mitjana dels dos valors que ocupen les posicions enteres més properes a *i* (en aquest cas per la mitjana dels valors que ocupen les posicions 3 i 4).

Exemple: atès el conjunt de 8 dades {5, 11, 7, 8, 10, 9, 6, 9}, el primer que cal fer és ordenar-les de menor a major, amb la qual cosa s'obté la sèrie {5, 6, 7, 8, 9, 9, 10, 11}; ara, la posició de la mediana serà donada per $i = \frac{8+1}{2} = 4{,}5$, és a dir, la mediana serà entre els valors que ocupen les posicions 4 i 5, per la qual cosa es calcula la mitjana d'ambdós per donar el valor de la mediana, com ara: $\tilde{x} = \frac{8+9}{2} = 8{,}5$.

És important notar aquí que la mitjana és molt sensible a l'existència de valors extrems (*outliers*). La inclusió o no d'un valor que estigui molt allunyat de la resta de les dades pot canviar considerablement el valor resultant de la mitjana. Per contra, la mediana està molt menys afectada per la presència dels valors esmentats, la qual cosa significa que la mediana és un "centre" més estable o robust que la mitjana, en el sentit que està menys afectat per la presència de valors extrems en les dades.

- **Moda (*mode*):** la moda d'un conjunt de dades és el valor que més vegades es repeteix (el que apareix amb més freqüència).

  Exemple: la moda de la sèrie de dades {6, 3, 4, 8, 9, 6, 6, 3, 4} és 6, ja que és el valor que més vegades apareix en la sèrie.

Estadístics de dispersió

Es presenten ara els principals estadístics de dispersió que, com s'ha comentat anteriorment, proporcionen informació sobre la variabilitat del conjunt de dades o fins a quin punt són representatius dels valors de la variable els estadístics de centralització:

- **Rang (*range*):** el rang d'un conjunt de dades és la diferència entre el valor màxim i el mínim d'aquests.

  Exemple: atès el conjunt de dades {2, 3, 8, 3, 5, 1, –8}, el seu rang són 8 – (–8) =16

- **Variància mostral (*sample variance*):** la variància d'una mostra es representa pel símbol s^2. En el cas d'una població, el paràmetre variància es representa pel símbol σ^2. La variància mostral serà més gran com més grans siguin les diferències entre cada una de les observacions, x_i, i la mitjana de les dades, $\overline{x}$, en concret:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  Això significa que la variància és una mesura de la dispersió de les dades respecte a la seva mitjana. Per exemple, com més petita sigui la variància, molt més agrupades estaran les dades al voltant del seu valor mitjana; per contra, com més gran sigui la variància, molt més disperses estaran les dades.

  Exemple: la variància mostral de la sèrie de 5 dades {6, 3, 8, 5, 3} és:

$$s^2 = \frac{(6-5)^2 + (3-5)^2 + (8-5)^2 + (5-5)^2 + (3-5)^2}{5-1} = 4,5$$

- **Desviació estàndard (*standard deviation*):** la desviació estàndard (o típica) d'una mostra es representa pel símbol *s*, mentre que la desviació estàndard d'una població es representa per σ. La desviació estàndard és l'arrel quadrada positiva de la variància, això és: $s = \sqrt{s^2}$ (o, dit d'una altra manera, la variància és el quadrat de la desviació estàndard). L'avantatge de la desviació estàndard sobre la variança és que aquella es mesura amb les mateixes unitats de mesura que la variable, mentre que la variança es mesura amb les unitats de mesura al quadrat, cosa que en pot dificultar la comprensió.

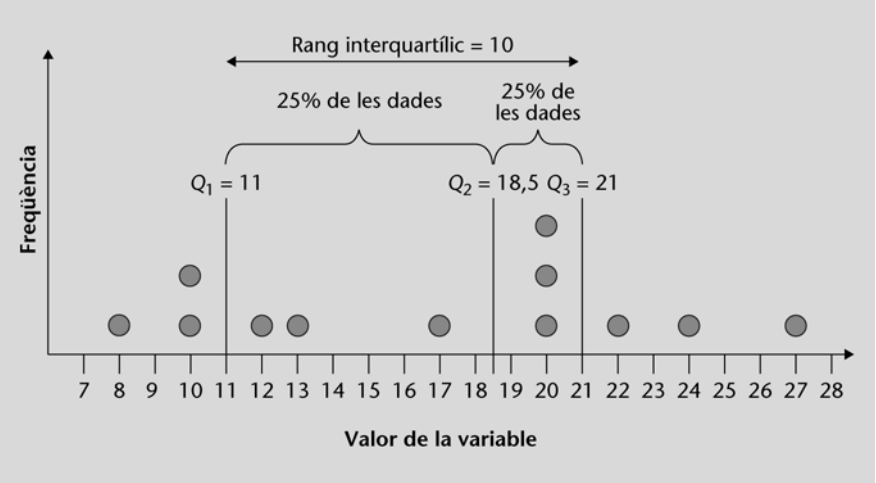
  Exemple: per a les dades de l'exemple anterior, $s = \sqrt{4,5} = 2,1$

  Igual com passava amb la variància, més desviació estàndard més dispersió en les dades i viceversa.

- **Quartils (*quartiles*):** en un conjunt de *n* observacions ordenades de menor a major valor, es poden considerar tres valors numèrics concrets anomenats quartils, que divideixen el conjunt en quatre parts, cada una contenint una quarta part de les observacions (figura 12). El primer

quartil, Q_1, és el valor que deixa la quarta part de les dades ordenades a la seva esquerra (un 25% de les dades mostren valors inferiors a aquest i un 75% de les dades mostren valors superiors a aquest). Per la seva part, el segon quartil, Q_2, és el valor que deixa la meitat de les dades ordenades a la seva esquerra (un 50% de les dades mostren valors inferiors a aquest i un 50% de les dades mostren valors superiors a aquest). Finalment, el tercer quartil, Q_3, és el valor que deixa tres quartes parts de les dades ordenades a la seva esquerra (un 75% de les dades mostren valors inferiors a aquest i un 25% de les dades mostren valors superiors a aquest).

Figura 12. Quartils d'un conjunt ordenat de dades



Observeu que, en realitat, el quartil segon o Q_2 coincideix amb el concepte de mediana presentat anteriorment. Els quartils són molt útils a l'hora de classificar una observació en una determinada franja del conjunt de dades. Per exemple, si l'observació és inferior a Q_1 significa que aquesta està situada entre el 25% de valors més baixos; si l'observació és superior a Q_3 significa que aquesta està situada entre el 25% de valors més alts, etc.

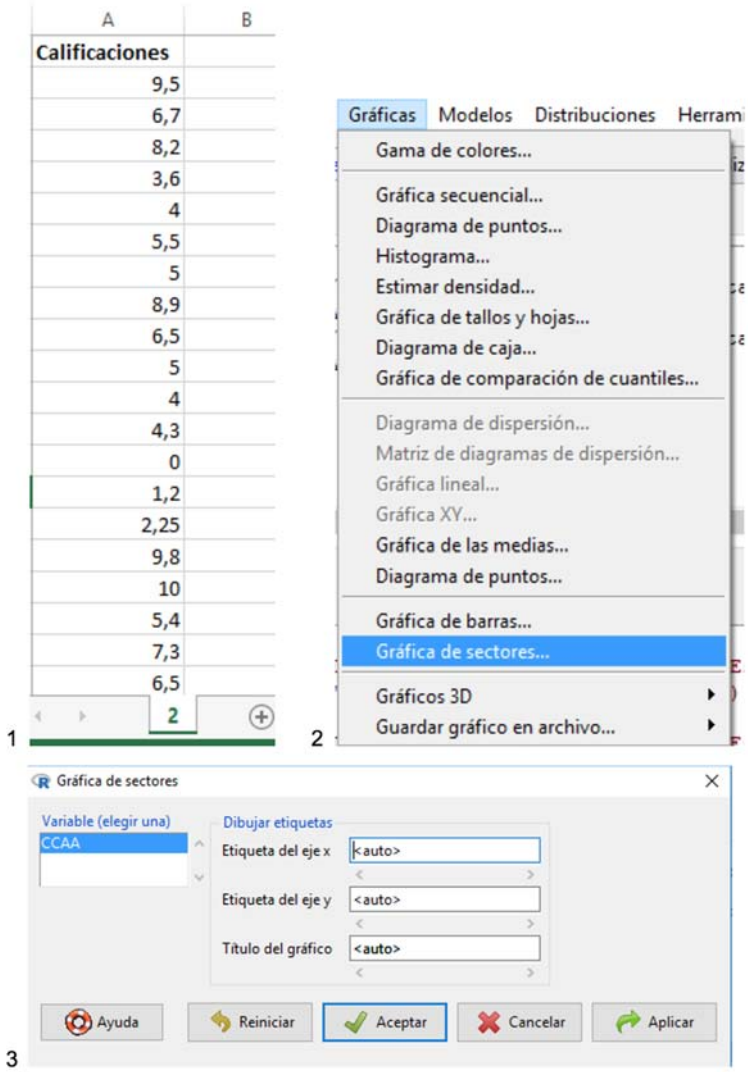
- **Rang interquartílic (*interquartilic range*):** aquest rang se sol representar per *IQR* i és simplement la diferència entre el tercer quartil i el primer quartil: $IQR = Q_3 - Q_1$. El rang interquartílic indica l'espai que ocupen el 50% de les observacions "centrals" (figura 12), per la qual cosa, de forma similar al que passava amb la variància, dóna una mesura de la dispersió de les dades (com més *IQR* més dispersió i viceversa).

Obtenció d'estadístics descriptius mitjançant programari

A la pràctica, és habitual utilitzar algun programari estadístic o d'anàlisi de dades per a calcular els estadístics anteriors i fins i tot alguns estadístics addicionals que proporcionin informació sobre el conjunt de dades. La figura

13 mostra els passos bàsics necessaris per a obtenir els principals estadístics descriptius amb R Commander. La sortida del programa, per a un exemple amb 50 observacions, es mostra a la figura 14. Per la seva part, la figura 15 mostra una sèrie d'estadístics descriptius generats amb MS Excel per al mateix conjunt de dades (en aquest cas els quartils s'han obtingut usant les fórmules integrades d'Excel).

Figura 13. Passos per calcular estadístics descriptius amb R Commander



Passos a seguir

Una vegada introduïdes les dades en el programa **(1)**, se segueix la ruta *Estadísticos* > *Resúmenes* > *Resúmenes numéricos*… **(2)** i se seleccionen les variables a la finestra corresponent **(3)**.

Figura 14. Estadístics descriptius obtinguts amb R Commander

```
> numSummary(Dataset[,"Calificaciones"], statistics=c("mean", "sd", "IQR
+   quantiles=c(0,.25,.5,.75,1))
    mean       sd  IQR  0% 25% 50%  75% 100%  n
 6.41625 2.465308 3.85 0.5 4.5 6.8 8.35   10 40
```

Diferències en els mètodes de càlculs

Noteu que hi ha lleugeres diferències entre els valors dels quartils calculats per Minitab i els corresponents valors d'Excel. Això es deu al fet que usen mètodes de càlcul diferents.

Podeu trobar una discussió interessant sobre els diferents mètodes que hi ha per a calcular els quartils a: http://mathforum.org/library/drmath/view/60969.html.

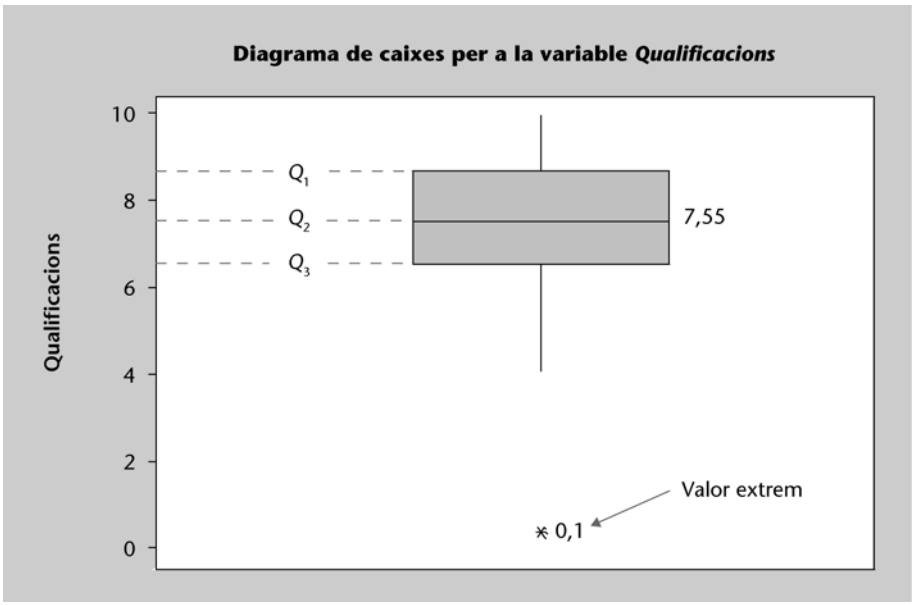
Figura 15. Estadístics descriptius calculats amb Excel

A	B	C	D	E	F	G
Calificacione						
9,5	9,50951406			Calificaciones		
6,7	0,08051406					
8,2	3,18176406			Media	6,41625	
3,6	7,93126406			Varianza	5,925798438	
4	5,83826406			Cuasivarianza	6,077741987	
5,5	0,83951406			desviación típica	2,465307686	
5	2,00576406			Mediana	6,9	
8,9	6,16901406			Moda	4	
6,5	0,00701406			Suma	256,65	
5	2,00576406			Cuenta	40	
4	5,83826406			Maximo	10	
4,3	4,47851406			Minimo	0,5	
0,5	35,0020141			Rango	9,5	
1,2	27,2092641					
2,25	17,3576391			Cuartil primero	4,5	
9,8	11,4497641			Cuartil segundo	6,8	
10	12,8432641			Cuartil tercero	8,45	
5,4	1,03276406					

Diagrama de caixa i bigotis (*boxplot*)

Usant els quartils és possible construir un tipus de gràfic, el diagrama de caixa (*boxplot*), que resulta molt útil a l'hora de visualitzar la distribució de les dades. Aquest diagrama està compost per una caixa central, definida pels quartils primers i tercers, que conté el 50% "central" de les observacions, i dos segments situats en els respectius extrems de la caixa, representant cada un el 25% de les observacions extremes (figura 16).

Figura 16. Diagrama de caixa i bigotis (*boxplot*) i valors extrems (*outliers*)

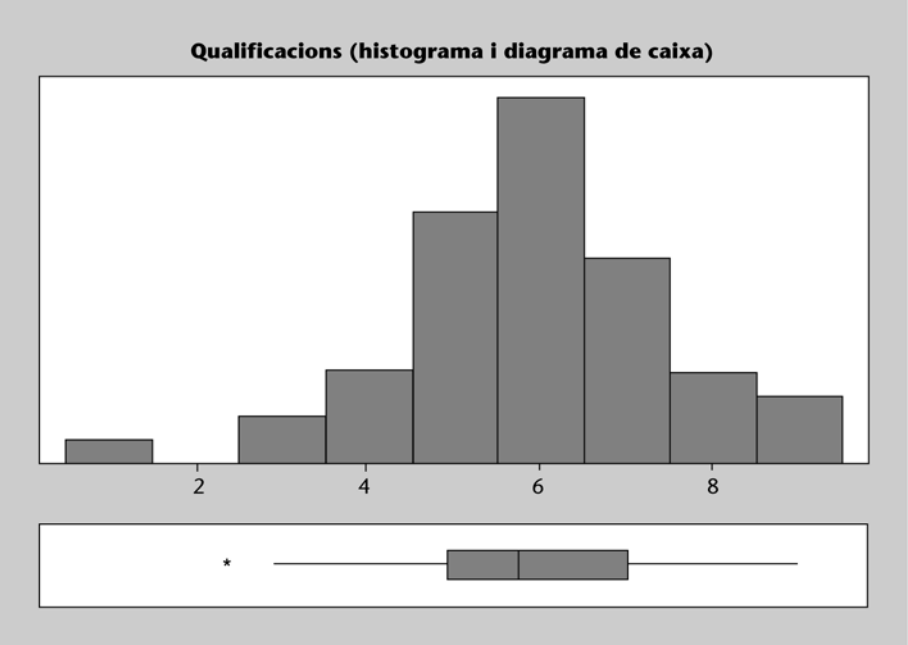


El diagrama de caixa i bigotis serveix també per identificar possibles valors anòmals (*outliers*), que estan excessivament allunyats de la resta de les dades, és a dir: o bé són extremadament grans o bé extremadament petits en comparació de la resta d'observacions. Aquests valors anòmals se solen re-

presentar mitjançant un asterisc, i poden ser deguts a un error en el registre de les dades o bé a valors que, en realitat, estan extremadament allunyats de la resta d'observacions (per exemple, el preu d'un Ferrari quan es compara amb preus de turismes de gamma mitjana). Identificar valors anòmals en un conjunt d'observacions és important, ja que l'anàlisi de les dades pot donar resultats molt diferents en funció que es considerin o no els esmentats valors en l'estudi (per exemple, la mitjana i la variància d'un conjunt de dades poden canviar de manera notable segons s'inclogui o no un d'aquests valors extrems).

L'estreta relació existent entre l'histograma i el diagrama de caixa es pot observar a la figura 17. En cert sentit, el diagrama de caixa es pot interpretar com un histograma vist des de dalt. En aquest cas, la zona del diagrama de caixa situada entre els quartils primer i tercer correspondria a la zona central de l'histograma. A més, en ambdós casos queda identificat el valor anòmal (*outlier*), així com la forma aproximadament simètrica de la resta de la distribució.

Figura 17. Relació entre histograma i diagrama de caixa



4. El concepte de probabilitat

Un **experiment aleatori** és aquell en el qual no és possible conèixer *a priori* l'esdeveniment resultant que s'esdevindrà però, tanmateix, sí que és possible observar un cert patró regular en els resultats que van succeint quan l'experiment es repeteix moltes vegades. Per exemple, quan es considera l'experiment aleatori consistent a llançar una moneda (o un dau) a l'aire, no és possible predir quin serà l'**esdeveniment resultant** de l'experiment, com ara, si sortirà cara o creu (o quin nombre sortirà en el cas del dau); tanmateix, sí que es pot afirmar que després de molts llançaments el percentatge o proporció d'esdeveniments *cara* obtinguts serà molt pròxim al 50% o 1/2 (en el cas del dau no trucat, el percentatge o proporció d'esdeveniments *3* obtinguts serà molt pròxim a 0,1667 o 1/6). Aquest percentatge o proporció d'aparició d'un esdeveniment després de moltes repeticions de l'experiment és el que dóna lloc a la idea de probabilitat:

Exemple

La **probabilitat** d'un succés és sempre un nombre entre 0 i 1. Així, per exemple, una probabilitat de 0,25 representa un percentatge d'aparició del 25% o, equivalentement, una proporció de 1/4.

> Es defineix la **probabilitat d'un esdeveniment** A, $P(A)$, com el percentatge o proporció d'aparició de l'esmentat esdeveniment en una sèrie extraordinàriament llarga de repeticions de l'experiment, totes independents entre si.

El requisit d'independència entre les diferents repeticions de l'experiment aleatori significa que el resultat de cada repetició de l'experiment no està condicionat pels resultats obtinguts en repeticions anteriors (per exemple, quan es llança diverses vegades una moneda a l'aire, l'esdeveniment resultant de cada nou llançament és independent dels resultats obtinguts en llançaments previs).

Exemple 1 de probabilitats

En l'experiment *llançament d'una moneda a l'aire*, és possible considerar els següents esdeveniments o potencials resultats: C = {cara}, X = {creu}, Ω = {cara o creu} i $\varnothing$ = {ni cara ni creu}. Els dos últims esdeveniments es coneixen, respectivament, com a esdeveniment segur Ω (que inclou tots els resultats possibles) o conjunt buit $\varnothing$ (que no inclou cap resultat derivat de l'execució de l'experiment). En aquest cas, sembla clar que $P(C) = 0,5$ (si es repetís l'experiment moltes vegades, aproximadament el 50% d'aquestes serien cares), $P(X) = 0,5$, $P(\Omega) = 1$ (en el 100% dels llançaments sortirà o bé cara o bé creu) i $P(\varnothing) = 0$ (en el 0% dels llançaments no s'obtindrà cap resultat).

Exemple 2 de probabilitats

En l'experiment aleatori *llançament d'un dau*, és possible considerar esdeveniments o potencials resultats com els següents: {1}, {2}, {3}, {4}, {5}, {6},

Ω = {un nombre entre 1 i 6}, $\varnothing$ = {cap nombre entre 1 i 6}. En aquest cas, $P(\{1\}) = 1/6$ (després de moltes repeticions, un de cada sis llançaments acabarà essent un 1), $P(\{2\}) = 1/6$, $P(\{3\}) = 1/6$, $P(\{4\}) = 1/6$, $P(\{5\}) = 1/6$, $P(\{6\}) = 1/6$, $P(\Omega) = 1$ i $P(\varnothing) = 0$.

Observeu, a més, que també és possible considerar esdeveniments compostos com, per exemple, parell = {2, 4, 6}, senar = {1, 3, 5}, major2 = {3, 4, 5, 6}, menor3 = {1, 2}, etc. En aquest cas, P(parell) = 3/6 = 1/2, P(imparell) = 1/2, P(major2) = 4/6 = 2/3, P(menor3) = 2/6 = 1/3.

Propietats bàsiques de les probabilitats

Hi ha una sèrie de propietats bàsiques que ha de satisfer qualsevol probabilitat. Aquestes propietats són molt útils a l'hora de calcular probabilitats d'esdeveniments complexos a partir de probabilitats ja conegudes o fàcils d'obtenir:

1) La probabilitat de qualsevol esdeveniment A sempre és un nombre situat entre 0 i 1 (ambdós inclusivament): $0 \leq P(A) \leq 1$.

Exemple: en els exemples anteriors, totes les probabilitats trobades eren valors entre 0 i 1.

2) La probabilitat de l'esdeveniment impossible o conjunt buit $\varnothing$ és sempre 0: $P(\varnothing) = 0$. En altres paraules, quan es realitza un experiment aleatori sempre s'obté algun resultat, per tant la proporció de *no resultats* és 0.

Exemple: en els exemples anteriors, $P(\varnothing) = 0$.

3) La suma de les probabilitats de tots els possibles resultats de l'experiment aleatori sempre val 1. En altres paraules, la probabilitat de l'esdeveniment segur és sempre 1.

Exemple: En l'exemple de la moneda, $P(\Omega) = 1 = P(C) + P(X)$; en l'exemple del dau, $P(\Omega) = 1 = P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\})$.

4) La probabilitat que un esdeveniment no ocorri és 1 menys la probabilitat que sí que ocorri: $P(no\ A) = 1 - P(A)$.

Exemple: en l'exemple de la moneda, $P(C) = 0{,}5 = 1 - P(no\ C) = 1 - P(X)$; en l'exemple del dau, P(parell) = 0,5 = 1 – P(no parell) = 1 – P(imparell); $P(\varnothing) = 1 - P(\Omega)$.

5) Si dos esdeveniments A i B no tenen resultats comuns (són disjunts), la probabilitat que ocorri $A \cup B$ és la suma de les probabilitats, si A i B són disjunts, $P(A \cup B) = P(A) + P(B)$.

Exemple: en l'exemple de la moneda, $P(C \cup X) = P(C) + P(X) = 1$; en l'exemple del dau, $P(\{1, 2\}) = P(\{1\}) + P(\{2\}) = 2/6 = 1/3$; $P(\Omega \cup \varnothing) = P(\Omega) + P(\varnothing) = 1 + 0 = 1$.

6) En general, per a qualsevol dos esdeveniments A i B, es complirà que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, on $A \cap B$ és el conjunt de possibles resultats que satisfan els esdeveniments A i B alhora. Notem que quan A i B són disjunts (no tenen resultats en comú), $A \cap B = \varnothing$ i, per tant, $P(A \cup B) = P(A) + P(B) - P(\varnothing) = P(A) + P(B) - 0 = P(A) + P(B)$, que és l'expressió vista en la propietat anterior.

Exemple: en l'exemple del dau, $P(\text{parell} \cup \text{major2}) = P(\text{parell}) + P(\text{major2}) - P(\text{parell} \cap \text{major2}) = 3/6 + 4/6 - 2/6 = 5/6$ (observarem que *parell* $\cap$ *major2* = {4, 6}).

5. Distribucions de probabilitat discretes

A l'inici d'aquest mòdul es va definir el concepte de variable quantitativa discreta com la variable quantitativa que podia prendre un nombre finit o comptable de valors diferents. Així, un exemple de variable discreta seria X = *resultat del llançament d'un dau*, ja que l'esmentada variable només pot prendre 6 possibles valors.

Cada un dels possibles valors d'una variable discreta tindrà associada una probabilitat d'ocurrència (per exemple, en el cas del dau, la probabilitat d'obtenir un 2 serà d'1/6), per la qual cosa sembla natural estudiar com es distribueixen o comporten les esmentades probabilitats. En concret, es pot definir una funció de probabilitat, f(x), que associï a cada valor x de la variable discreta X la probabilitat d'ocurrència, $P(x)$. Per exemple, en el cas de la variable anterior, associada a l'experiment aleatori *llançament d'un dau normal*, la corresponent funció de probabilitat seria: $f(1) = P(X = 1) = 1/6$, $f(2) = P(X = 2) = 1/6$, $f(3) = P(X = 3) = 1/6$, $f(4) = P(X = 4) = 1/6$, $f(5) = P(X = 5) = 1/6$, $f(6) = P(X = 6) = 1/6$.

Observem

Notem que si es fa servir un **dau trucat**, no totes les probabilitats d'ocurrència són iguals i, per tant, la funció de probabilitat torna valors diferents per a diferents valors per a diferents possibles valors de la variable.

Donada una variable aleatòria discreta, X, resulta útil conèixer la **distribució de probabilitat** de l'esmentada variable, com ara com es distribueixen o comporten les probabilitats d'ocurrència dels possibles valors. A tal efecte es defineixen les funcions següents:

La **funció de probabilitat** de X és la funció f(x) que assigna a cada possible valor x de X la seva probabilitat d'ocurrència, és a dir: $f(x) = P(X = x)$ per a tot valor possible x de X.

La **funció de distribució** de X és la funció $F(x)$ que assigna a cada possible valor x de X la probabilitat acumulada d'ocurrència, és a dir, $F(x) = P(X = x)$ per a tot valor possible x de X.

La taula 3 mostra la funció de probabilitat i la funció de distribució corresponents a la variable X anterior, però usant un dau trucat que té dos valors 6 i cap valor 2. Per la seva part, la figura 18 mostra ambdues funcions superposades al mateix gràfic. Observant detingudament la taula 3 i la figura 18 es poden deduir les característiques següents pròpies d'aquestes funcions:

- Com que representen probabilitats, ambdues funcions sempre prenen valors en l'interval [0, 1].

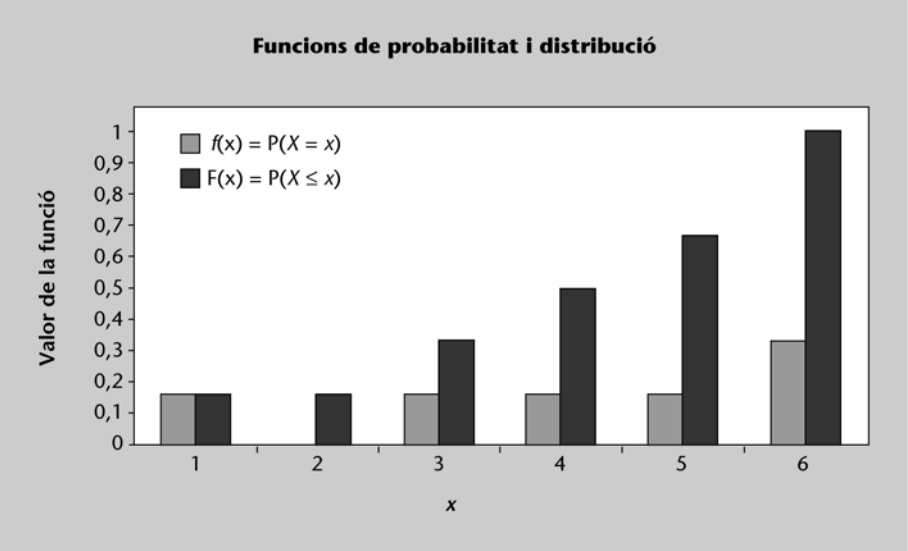
- La suma de tots els valors que pren la funció de probabilitat sempre ha de ser 1 (això es deu a les propietats de la probabilitat).

La funció de distribució sempre és una funció creixent que passa de valor 0 en el seu extrem esquerre ($F(0) = P(X \leq 0) = 0$) a valor 1 en el seu extrem dret ($F(6) = P(X = 6) = 1$).

Taula 3. Funcions de probabilitat i distribució per a una variable discreta

X	**Funció de probabilitat** $f(x) = P(X = x)$	**Funció de distribució** $F(x) = P(X \leq x)$
1	1/6	1/6
2	0	1/6
3	1/6	2/6
4	1/6	3/6
5	1/6	4/6
6	2/6	1
Total	**1**	

Figura 18. Funcions de probabilitat i distribució d'una variable discreta



Paràmetres descriptius d'una distribució discreta

Mentre que els estadístics descriptius i els gràfics/taules de freqüències s'utilitzen per analitzar el comportament (distribució) d'una mostra d'observacions empíriques, les distribucions de probabilitat són models estadístics que fa ús de paràmetres i de funcions de distribució per descriure el comportament teòric (distribució teòrica) de tota una població. De forma anàloga al que ocorria amb les mostres –que eren caracteritzades per estadístics descriptius com la mitjana o la variància mostral–, també les distribucions de probabilitat associades a poblacions se solen caracteritzar per paràmetres tals com la mitjana o la variància poblacional. Ara bé, ja que en general no es disposarà d'observacions sobre tota la població, sinó només d'una funció de distribució o de probabilitats, la forma de calcular els esmentats paràmetres és una mica diferent:

- **Mitjana o valor esperat d'una variable discreta:** La mitjana o valor esperat d'una variable discreta X que pot prendre els valors x_1, x_2 ..., es representa per μ o $E[X]$ i es calcula de la manera següent:

$$\mu = E[X] = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + ... = \sum_i x_i \cdot f(x_i)$$

on $f(x)$ denota la funció de probabilitat de X.

Exemple: notem que en el cas d'un dau equilibrat, el valor esperat o mitjana de X = *resultat del llançament* seria $\mu = 3$; tanmateix, en el cas del dau trucat que es mostra a la taula 3, la mitjana o valor esperat és:

$$\mu = 1 \cdot f(1) + 2 \cdot f(2) + 3 \cdot f(3) + 4 \cdot f(4) + 5 \cdot f(5) + 6 \cdot f(6) =$$
$$= 1 \cdot \frac{1}{6} + 2 \cdot 0 + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{2}{6} = 4{,}167$$

- **Variància i desviació estàndard d'una variable discreta:** La variància d'una variable discreta X que pot prendre els valors x_1, x_2 ..., es representa per σ^2 i es calcula de la manera següent:

$$\sigma^2 = (x_1 - \mu)^2 \cdot P(X = x_1) + (x_2 - \mu)^2 \cdot P(X = x_2) + ... = \sum_i (x_i - \mu)^2 \cdot f(x_i)$$

on $f(x)$ denota la funció de probabilitat de X. De manera anàloga a com ocorria amb els estadístics mostrals, la desviació estàndard d'una variable és l'arrel quadrada positiva de la seva variància, és a dir:

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

Exemple: en el cas del dau trucat que es mostra a la taula 3, la variància és:

$$\sigma^2 = (1 - 4{,}167)^2 \cdot \frac{1}{6} + (2 - 4{,}167)^2 \cdot 0 + (3 - 4{,}167)^2 \cdot \frac{1}{6} +$$
$$+ (4 - 4{,}167)^2 \cdot \frac{1}{6} + (5 - 4{,}167)^2 \cdot \frac{1}{6} + (6 - 4{,}167)^2 \cdot \frac{2}{6} = 3{,}139$$

I la corresponent desviació estàndard: $\sigma = \sqrt{3{,}139} = 1{,}772$

La distribució binomial

Una de les distribucions discretes més utilitzades a la pràctica és la distribució binomial. Aquesta distribució s'usa per a contestar a preguntes com les següents:

- Si cada vegada que un sistema informàtic és atacat per un virus la probabilitat que el sistema no falli és de 0,76, quina és la probabilitat que no s'hagi produït cap error en el sistema després de 5 atacs?
- Si cada vegada que es consulta una font d'informació la probabilitat que aquesta proporcioni una resposta satisfactòria és de 0,85, quina és la probabilitat que se n'obtingui alguna resposta satisfactòria després de 3 consultes?

Distribució de Poisson i la uniforme discreta

Altres distribucions discretes molt habituals són la distribució de Poisson i la uniforme discreta. És possible trobar a Internet força documentació sobre aquestes distribucions discretes i d'altres, així com els seus àmbits d'aplicació.

- Si després de l'administració d'un fàrmac a un pacient en estat crític la probabilitat de supervivència d'aquest és de 0,99, quina és la probabilitat que sobrevisquin els 14 pacients crítics que han rebut el tractament?
- Si la probabilitat d'obtenir una concessió per a un projecte d'investigació és de 0,20, quina és la probabilitat d'obtenir almenys una concessió després de tres intents?
- Si cada vegada que es tracta d'enquestar un transeünt elegit a l'atzar la probabilitat que aquest respongui és de 0,15, quina és la probabilitat que s'aconsegueixin obtenir 80 o més respostes a partir d'una mostra aleatòria de 150 transeünts?

La **distribució binomial** és un model estadístic que permet calcular probabilitats sobre la variable aleatòria X = nombre d'èxits aconseguits en n proves independents. Cada una d'aquestes n proves és una repetició d'un experiment aleatori el resultat del qual és binari (èxit/fracàs), on p és la probabilitat d'èxit en cada prova i $q = 1 - p$ la probabilitat de fracàs.

Resultat *èxit*

No s'ha de confondre el resultat ***èxit*** d'un experiment aleatori amb el fet que el resultat sigui desitjable des d'un punt de vista social o subjectiu. Així, per exemple, es podria considerar *èxit* de l'experiment aleatori la fallida del sistema informàtic que és atacat per un virus.

Observem que la variable X = *nombre d'èxits en n proves independents* pot prendre qualsevol valor k entre 0 i n (ambdós inclusivament). Se sol usar la notació $X \sim B(n, p)$ per a indicar que X es distribueix o es comporta segons una distribució binomial de paràmetres n (nombre de proves o repeticions) i p (probabilitat d'èxit en cada prova). En aquestes condicions, les probabilitats associades a l'esmentada variable són donades per l'expressió matemàtica següent:

Per a qualsevol k entre 0 i n, $P(X=k)=\binom{n}{k}p^k\cdot(1-p)^{n-k}$, on $\binom{n}{k}=\frac{n!}{k!(n-k)!}$,

on $0! = 1! = 1$ i $n! = n \cdot (n-1) \ldots 1$ per a tot $n > 1$.

Observem

L'expressió $n!$ es llegeix com a factorial de n o n factorial. Així, per exemple, $4! = 4 \cdot 3 \cdot 2 \cdot 1$ i $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$. No obstant això, $1! = 1$ i $0! = 1$.

Es compleix a més que la mitjana (valor esperat) i la variància d'una distribució binomial són, respectivament: $\mu = n \cdot p$ i $\sigma^2 = n \cdot p \cdot (1-p)$.

Exemple: la probabilitat que en introduir dades en un formulari web es cometi un error és de 0,1. Si deu persones emplenen l'esmentat formulari de manera independent, quina és la probabilitat que no hi hagi més d'un formulari erroni?, quin és el valor esperat i la desviació estàndard de la variable considerada?

Observem que, en aquest cas X = *nombre de formularis erronis en 10 proves* i $X \sim B(10, 0{,}1)$. A més, es demana $P(X \leq 1) = P(X = 0 \text{ o } X = 1) = P(X = 0) + P(X = 1)$ (ja que són esdeveniments disjunts). Ara bé:

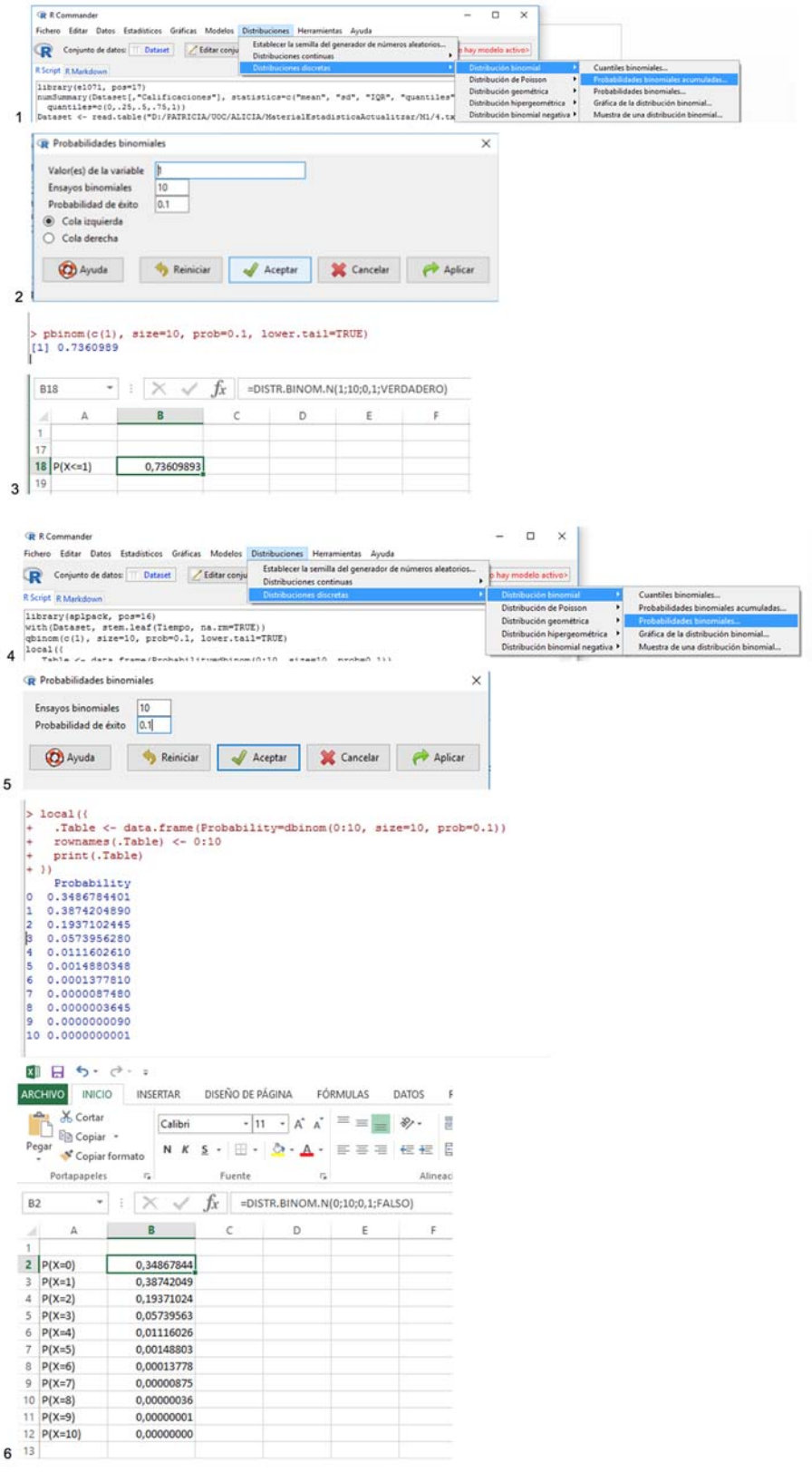
$$P(X=0)=\binom{10}{0}0{,}1^0\cdot(0{,}9)^{10}=\frac{10!}{0!10!}(1)(0{,}3487)=0{,}3487$$

$$P(X=1)=\binom{10}{1}0{,}1^1\cdot(0{,}9)^{9}=\frac{10!}{1!9!}(0{,}1)(0{,}3874)=0{,}3874$$

Per tant, $P(X \leq 1)$ =0,3874 + 0,3487 = 0,7361. Finalment, $\mu = 10 \cdot 0{,}1 = 1$ i $\sigma = \sqrt{10 \cdot 0{,}1 \cdot 0{,}9} = 0{,}9487$.

A la pràctica, els càlculs probabilístics anteriors se solen automatitzar amb l'ajuda d'algun programari estadístic o d'anàlisi de dades. La figura 19 mostra com es poden calcular probabilitats d'una distribució binomial amb ajuda de R Commander i Excel.

Figura 19. Càlcul de probabilitats en una distribució binomial amb R Commander i Excel

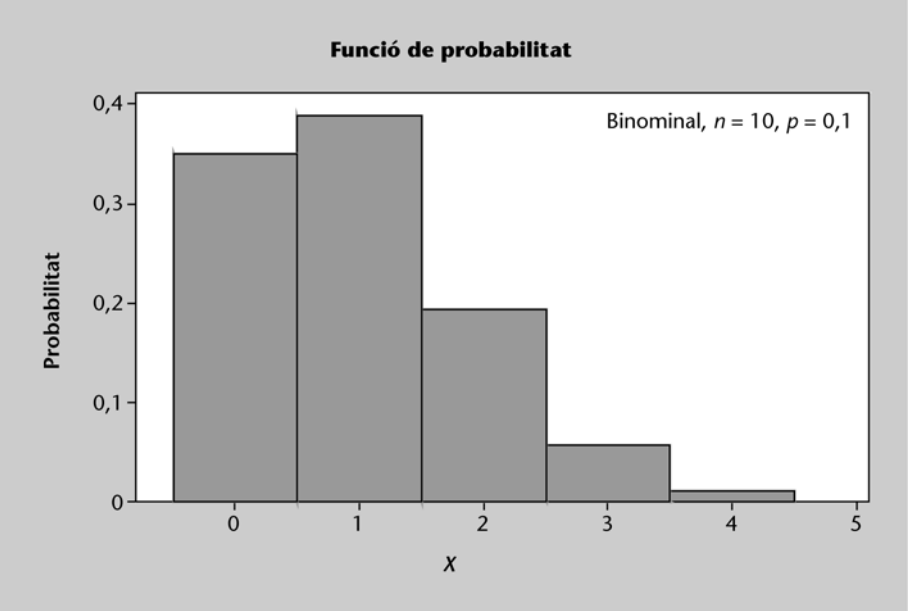


Passos a seguir

Se segueix la ruta *Distribuciones> Distribuciones discretas> Distribución binomial > Probabilidades binomiales acumuladas* **(1)** i es completen els paràmetres a la finestra corresponent **(2).** El resultat es mostra a **(3)**. Observem que, si en lloc d'escollir l'opció *Probabilidades binomiales acumuladas* a **(2)** s'hagués escollit l'opció *Probabilidades binomiales* (4) completant els paràmetres a la finestra corresponent (5), el programa hauria calculat $P(X = 1)$ en lloc de $P(X <= 1)$ **(6)**.

La figura 20 mostra la funció de probabilitat associada a la binomial de l'exemple anterior. S'observa que, encara que en teoria els possibles valors de la variable X anirien des de 0 fins a 10 (nombre de proves), a la pràctica els valors majors que 4 tenen probabilitat d'esdeveniment pràcticament nul·la (és molt poc freqüent que s'obtinguin valors superiors a 4). En efecte, $P(X > 4) = 1 - P(X <= 4) = \{\text{usant R Commander o Excel}\} = 1 - 0{,}9984 = 0{,}0016$.

Figura 20. Funció de probabilitat d'una $B(10, 0{,}1)$

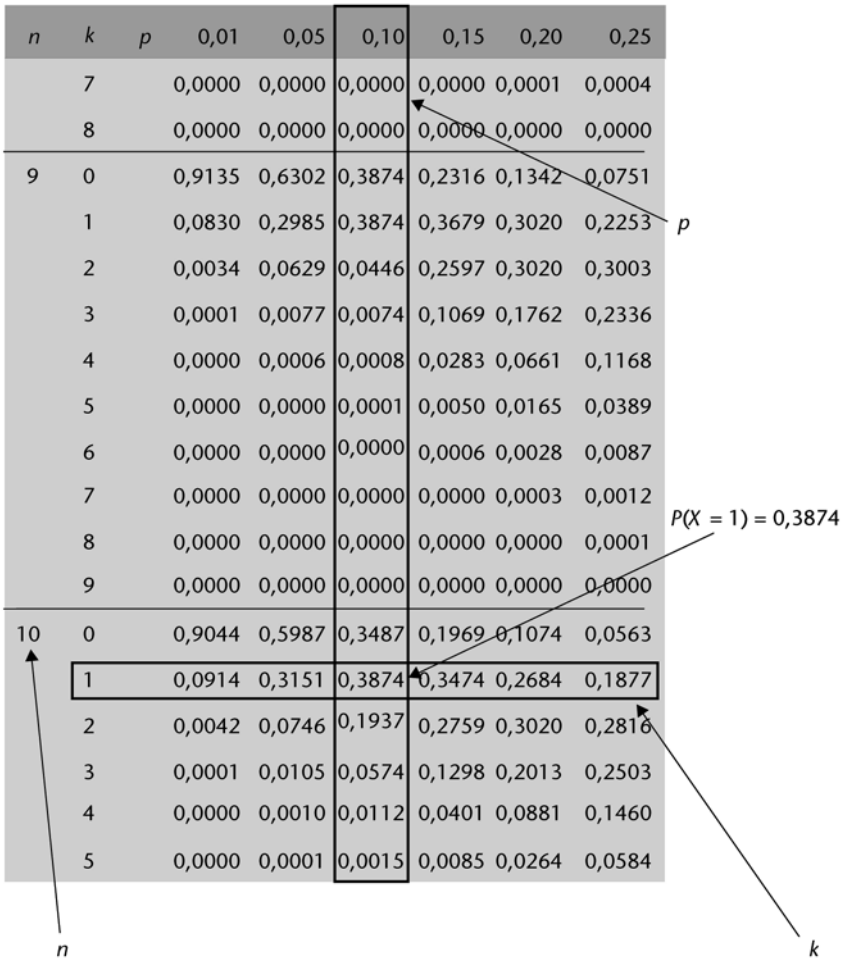


Les probabilitats anteriors es poden obtenir també amb taules estadístiques (sense necessitat d'utilitzar cap programari). Així, seguint l'exemple anterior, la figura 21 mostra com es calcula $P(X = 1)$ amb la taula binomial. En aquest cas, X és una $B(10, 0{,}1)$ i es vol trobar $P(X = k)$, essent $k = 1$. Per a això, es busca la secció de la taula corresponent a $n = 10$, i la intersecció entre la fila $k = 1$ i la columna $p = 0{,}1$.

Càlcul de probabilitats

Resulta fàcil trobar a Internet molts documents que expliquen molt detalladament l'ús de taules per a calcular probabilitats. Tanmateix, convé automatitzar els càlculs tant com sigui possible mitjançant l'ús de programari.

Figura 21. Càlcul de probabilitats binomials mitjançant taules



6. Distribucions de probabilitat contínues

A l'inici d'aquest mòdul es va definir el concepte de variable quantitativa contínua com la variable quantitativa que podia prendre un nombre infinit (no comptable) de valors diferents. Així, un exemple de variable contínua seria X = *temps que es triga a desenvolupar un portal web*, ja que l'esmentada variable pot prendre un valor real qualsevol entre 0 i infinit.

A diferència del que ocorria amb les variables discretes, quan es treballa amb variables contínues no és possible definir una funció de probabilitat que assigni probabilitats als diferents valors de la variable: si X és una variable contínua, X pot prendre un nombre infinit (no comptable) de valors, per la qual cosa la probabilitat teòrica que la variable X prengui un valor concret, x, són sempre 0, és a dir: $P(X = x) = 0$ per a qualsevol valor x de X. si és possible, tanmateix, assignar probabilitats a intervals de valors. Per exemple, si el 51% dels portals web triguen a desenvolupar-se entre 240 i 258 hores, aleshores $P(240 < X < 258) = 0{,}51$. Per a descriure la distribució de probabilitat d'una variable contínua es continua utilitzant la funció de distribució (encara que amb algun matís nou) i, a més, s'usa també l'anomenada *funció de densitat* en lloc de la funció de probabilitat típica de variables discretes:

Nota

En variables contínues, ja que $P(X = x) = 0$ per a qualsevol valor x de X, es complirà que:

a) $P(X \leq x) = P(X < x)$

b) $P(X \geq x) = P(X > x)$

La **funció de densitat** d'una variable contínua X és una funció $f(x)$ tal que la probabilitat que X prengui un valor en un interval (a, b) coincideix amb **l'àrea** compresa per l'esmentada funció entre els extrems de l'esmentat interval (figura 22), *i. e.*: $P(a < X < b)$ = àrea sota f(x) entre a i b.

La **funció de distribució** de X és aquella funció $F(x)$ que assigna a cada possible valor x de X la seva probabilitat acumulada d'ocurrència (figura 23), *i. e.* $F(x) = P(X \leq x)$ = àrea sota $f(x)$ des de $-\infty$ (menys infinit) fins a x.

Nota

La funció de densitat $f(x)$ sempre és postiva i comprèn una área total d'1.

Atenció

Observem l'equivalència entre els conceptes de "probabilitat" i "àrea".

La figura 22 mostra la funció de densitat d'una variable amb distribució simètrica i centrada en el valor 250 (atès que la funció és totalment simètrica, la mitjana i la mediana coincideixen en aquest punt). S'observa també l'àrea tancada sota funció de densitat entre els valors $a = 240$ i $b = 258$. Aquesta àrea correspon amb la probabilitat següent: $P(240 < X < 258)$. Per la seva part, la figura 23 mostra la funció de distribució associada a la mateixa variable. Novament s'aprecia la simetria respecte al valor central, així com el fet que la funció de distribució va creixent així que va acumulant probabilitats, i passa del valor 0 en l'extrem esquerre al valor 1 en l'extrem dret. A partir d'aquesta gràfica es poden estimar visualment probabilitats acumulades, com ara: $P(X <= 260)$ serà un valor molt proper a 0,8.

Figura 22. Funció de densitat d'una variable contínua i àrea tancada

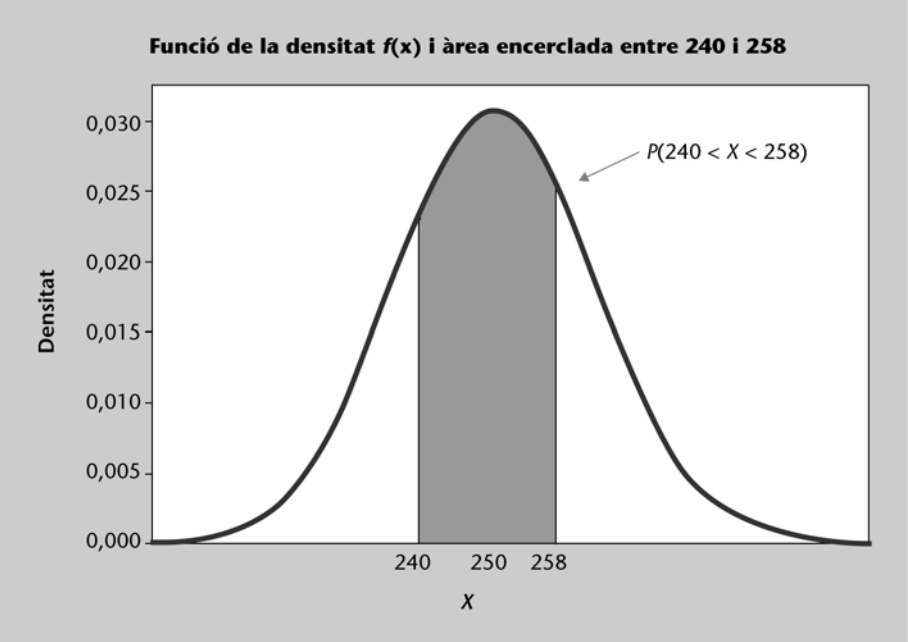
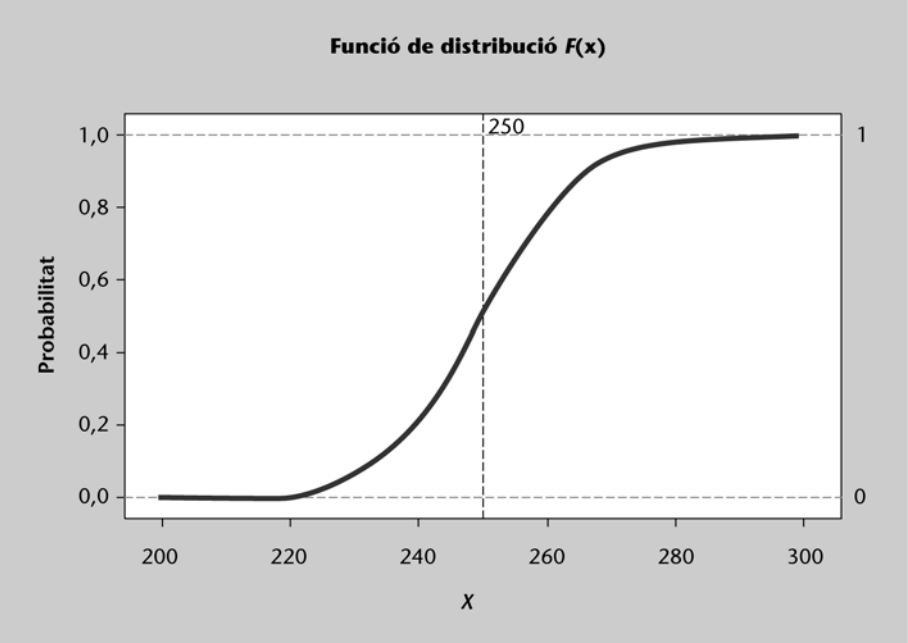


Figura 23. Funció de distribució d'una variable contínua



Funció de distribució

La funció de distribució és una funció acumulativa de probabilitats i, per tant, és sempre creixent, i passa de 0 (extrem esquerre) a 1 (extrem dret).

Paràmetres descriptius d'una distribució contínua

En el cas de distribucions contínues, la manera de calcular els paràmetres és similar a l'emprada per a distribucions discretes, si bé ara els sumatoris són substituïts per àrees (integrals definides en termes matemàtics) entre dos extrems:

Atenció

Tot i que a la pràctica es farà ús de programari estadístic per a fer els càlculs, és important conèixer quins conceptes es fan servir per a definir cada tipus de paràmetre.

- **Mitjana o valor esperat d'una variable contínua:** la mitjana o valor esperat d'una variable contínua X és representada per μ o $E[X]$ i es calcula de la manera següent:

$$\mu = E[X] = \text{àrea total sota } "x \cdot f(x)" = \int_{-\infty}^{+\infty} x \cdot f(x) dx$$

on $f(x)$ denota a la funció de densitat de X.

- **Variància i desviació estàndard d'una variable contínua:** la variància d'una variable contínua X és representada per σ^2 i es calcula de la manera següent:

$$\sigma^2 = \text{àrea total sota } "(x - \mu)^2 \cdot f(x)" = \int_{-\infty}^{+\infty} (x - \mu)^2 \cdot f(x) dx$$

on $f(x)$ denota la funció de densitat de X. Com sempre, la desviació estàndard d'una variable és l'arrel quadrada positiva de la seva variància, és a dir:

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

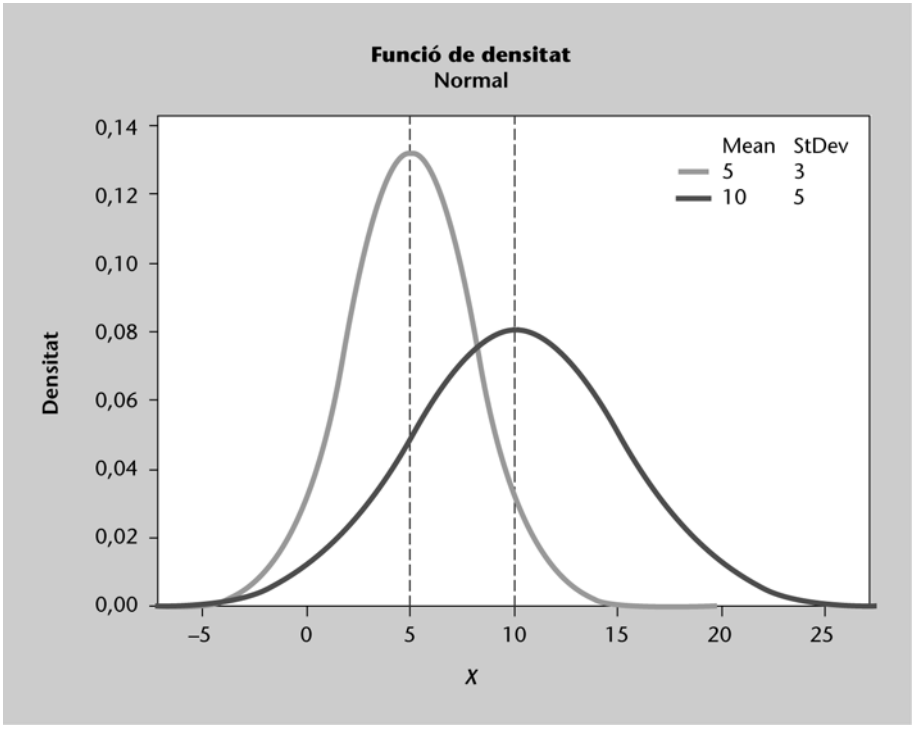
La distribució normal o gaussiana

La distribució normal o gaussiana és la distribució teòrica més important. Moltes variables contínues segueixen una distribució normal o aproximadament normal. Altres variables contínues i discretes també poden ser aproximades, en determinades circumstàncies, mitjançant una distribució normal. La normal, a més, és una distribució clau en l'estadística inferencial, ja que algunes de les seves propietats s'utilitzen per a obtenir informació sobre tota la població a partir d'informació sobre una mostra.

La forma concreta d'una distribució normal és caracteritzada per dos paràmetres: la mitjana, μ, que defineix on se situa el centre de la funció de densitat, i la desviació estàndard, σ, que defineix l'amplitud de la funció de densitat. Quan una variable contínua X segueix una distribució normal, se sol representar per $X \sim N(\mu, \sigma)$.

Les figures 22 i 23 mostren, respectivament, la funció de densitat i la funció de distribució d'una normal amb mitjana $\mu = 250$ i desviació estàndard $\sigma = 13$. La figura 24 mostra les funcions de densitat per a dues distribucions de tipus normal amb paràmetres $\{\mu = 5, \sigma = 3\}$ i $\{\mu = 10, \sigma = 5\}$ respectivament. S'observa que la funció de densitat de la normal té forma de campana de Gauss, elevada en el centre (el valor mitjà o esperat) i amb dues cues simètriques en els extrems. Notem, a més, com cada una de les corbes està centrada en la mitjana, així com el fet que la corba és més ampla com més gran és la desviació estàndard.

Figura 24. Funcions de densitat associades a dues normals



Com en qualsevol altra funció de densitat, l'àrea total compresa sota la corba és d'1. A la pràctica això significa que per a qualsevol valor x de X, $P(X > x) = 1 - P(X < x)$, és a dir, l'àrea a la dreta d'un valor és l'àrea total (que val 1) menys l'àrea a la seva esquerra i viceversa (figura 25). A més, ja que la normal és una distribució simètrica respecte a la seva mitjana, l'àrea compresa per una cua és igual a l'àrea compresa per la cua oposada (figura 26).

Figura 25. L'àrea total d'una funció de densitat és 1

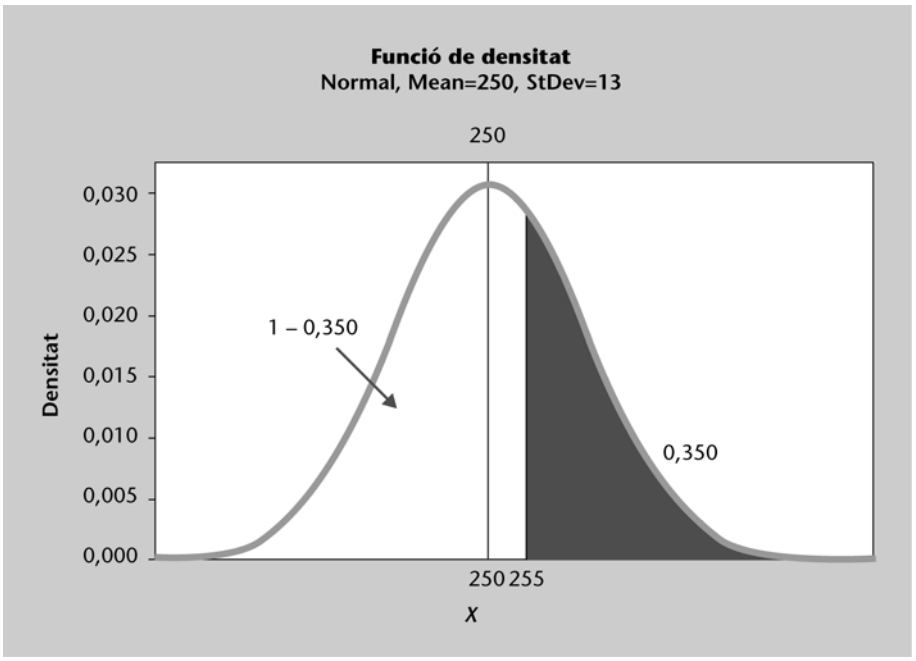
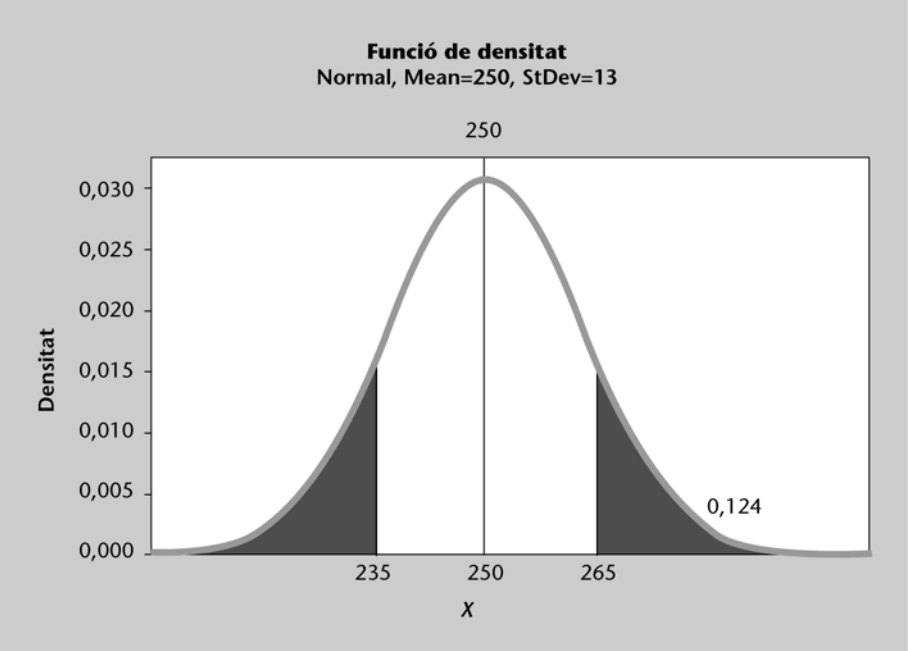


Figura 26. Dues cues simètriques comprenen la mateixa àrea



Qualsevol distribució normal compleix a més l'anomenada **regla 68-95-99,7,** segons la qual l'interval $(\mu - \sigma, \mu + \sigma)$ conté, aproximadament, el 68% de les observacions, l'interval $(\mu - 2, \mu + 2)$ conté, aproximadament, el 95% de les observacions, i l'interval $(\mu - 3, \mu + 3)$ conté, aproximadament, el 99,7% de les observacions. Així, per exemple, si el $X \sim N$ (250,13), es pot afirmar que un 68% de les observacions de X estaran en l'interval (237, 263), un 95% de les observacions estaran en l'interval (224, 276), i un 99,7% de les observacions estaran en l'interval (211, 289). Observem, per tant, que serà altament improbable trobar valors de X fora d'aquest últim interval.

D'entre les infinites distribucions normals que es poden considerar variant els paràmetres μ i σ convé esmentar l'anomenada **normal estàndard**, que té per paràmetres $\mu = 0$ i $\sigma = 1$. En altres paraules, una variable contínua Z es distribuirà segons una normal estàndard, $Z \sim N(0,1)$, si la seva funció de densitat és la d'una normal centrada en l'origen i amb desviació estàndard unitària. Aquesta distribució normal estàndard se sol utilitzar sovint en estadística inferencial, i també quan es volen calcular probabilitats d'una normal qualsevol mitjançant l'ús de taules de probabilitats ja calculades.

En efecte, atesa una variable normal qualsevol, $X \sim N(\mu, \sigma)$, és possible aplicar-li un **procés d'estandardització** per a obtenir una normal estàndard Z. Això s'aconsegueix restant a la variable X la mitjana μ (amb la qual cosa la funció de densitat és desplaçada al llarg de l'eix x fins que queda centrada en l'origen) i dividint el resultat per la desviació estàndard σ (amb la qual cosa la nova variable tindrà una desviació estàndard unitària), és a dir: $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$. Aquest procés d'estandardització permet, entre altres coses, calcular probabilitats per a una normal qualsevol a partir de les taules de probabilitats precalculades que existeixen per a la distribució normal estàn-

dard (figura 27), la qual cosa evita haver de resoldre integrals cada vegada que es vol obtenir una nova probabilitat. Suposem, per exemple, que X segueix una $N(1.500, 100)$ i es vol obtenir $P(X < 1.400)$ mitjançant l'ús de taules. El primer pas consisteix a estandarditzar els valors:

$$P(X < 1.400) = P\left(\frac{X - \overline{x}}{\sigma} < \frac{1.400 - \overline{x}}{\sigma}\right) = P\left(Z < \frac{1.400 - 1.500}{100}\right) = P(Z < -1)$$

En altres paraules, es vol calcular l'àrea a l'esquerra del valor –1 en una normal tipificada o estàndard. Normalment, la taula de la normal estàndard, Z, ofereix àrees (probabilitats) a l'esquerra de valors positius, per la qual cosa resultarà necessari fer una petita transformació tenint en compte que: (a) per simetria de la normal estàndard, l'àrea (probabilitat) a l'esquerra d'un valor negatiu k és igual a l'àrea (probabilitat) a la dreta del valor positiu corresponent, $|k|$ (*i. e.*, $P(Z < -1) = P(Z > 1)$), i (b) l'àrea (probabilitat) total continguda sota la corba és 1 (és a dir, l'àrea a l'esquerra d'un valor més l'àrea a la seva dreta suma 1, per exemple: $P(Z < 1) + P(Z > 1) = 1$). Tenint en compte l'anterior, es dedueix que $P(Z < -1) = P(Z > 1) = 1 - P(Z < 1) =$ {vegeu la taula de la figura 27)} $= 1 - 0,8413 = 0,1587$.

Figura 27. Càlcul de probabilitat en una normal mitjançant taules

	,00	,01	,02	,03	,04	,05
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265

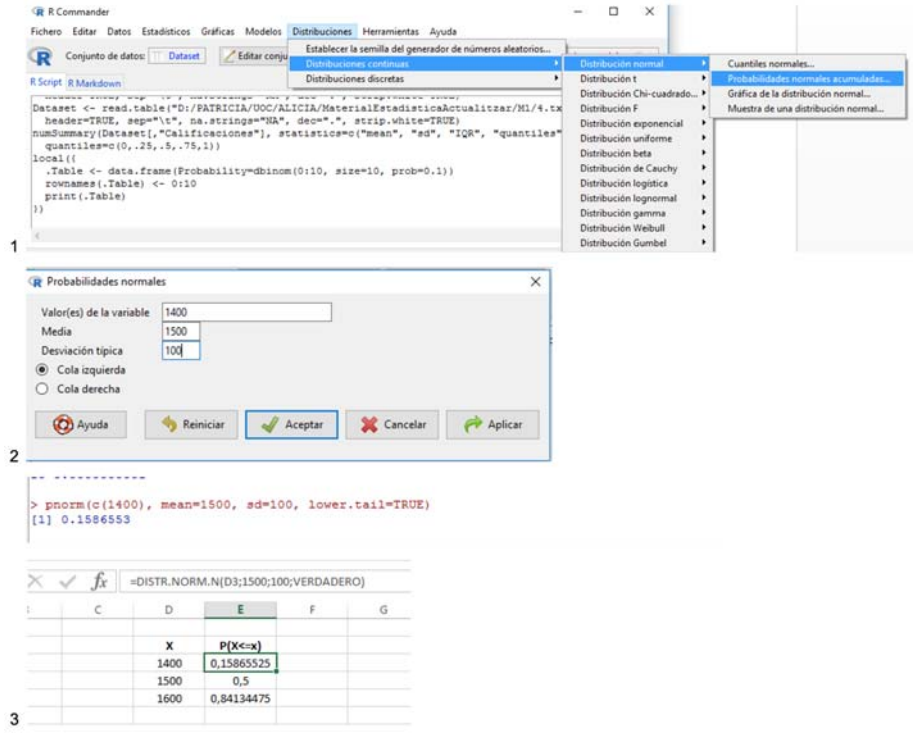
Nota

Noteu que per a trobar $P(Z < 1,00)$ usant la taula, s'ha de buscar el valor d'intersecció entre la fila 1,0 i la columna 0,00 (atès que 1,00 = 1,0 + 0,00). Si es demanés $P(Z < 1,24)$, llavors s'hauria de buscar la intersecció entre la fila 1,2 i la columna 0,04 (atès que 1,24 = 1,2 + 0,04), i s'obtindria el valor 0,8925.

D'altra banda, també és possible automatitzar el càlcul de probabilitats d'una normal qualsevol mitjançant l'ús de programari estadístic, amb la qual cosa s'elimina així la necessitat de resoldre manualment les integrals indefinides o d'haver d'usar taules de probabilitats precalculades. La figura 28 mostra com obtenir probabilitats d'una normal amb R Commander. En concret, per a una normal amb mitjana μ =1.500 i desviació estàndard σ = 100, s'obté que $P(X < 1.400) = 0,158655$. Alhora, la figura 28 mostra com s'han obtingut amb R Commander i Excel algunes probabilitats per a la ma-

teixa variable. Observem que $P(X < 1.500) = 0,5$, el qual és lògic, ja que 1.500 és la mitjana i, alhora, la mitjana de la distribució normal.

Figura 28. Càlcul de probabilitats en una normal amb R Commander i Excel



Passos a seguir

Se segueix la ruta *Distribuciones continuas > Distribución normal > Probabilidades normales acumuladas* **(1)** i es completen els paràmetres a la finestra corresponent **(2).** El resultat es mostra a **(3)**. Si s'hagués escrit el següent codi: *dnorm (1400, mean = 1500, sd = 100)* el programa hauria calculat el valor de la funció de densitat a $x = 1.400$ en lloc de $P(X < 1.400)$. Finalment, per a una probabilitat p donada, s'ha d'escriure el següent codi indicant els següents arguments: *qnorm(probability, mean, standard deviation)*, retorna aquell valor *c* de la variable *X* tal que $P(X < c) = p$.

Exemples d'aplicació d'una normal

- Segons un estudi realitzat pel Ministeri d'Educació, el nombre d'hores anuals que dediquen els nens espanyols a veure la televisió és una variable aleatòria que segueix una distribució normal de mitjana 1.500 hores i desviació estàndard de 100 hores. Quin percentatge de nens dediquen entre 1.400 i 1.600 hores anuals?

  En aquest cas, $X \sim N$ (1.500,100) i es demana $P(1.400 < X < 1.600)$. Per la regla 68-95-99,7, tenim que la probabilitat anterior serà, aproximadament, del 68% (ja que $\mu - \sigma = 1.400$ i $\mu + \sigma = 1.600$). Per a calcular de manera més exacta l'esmentada probabilitat, convé notar que $P(1.400 < X < 1.600) = P(X < 1.600) - P(X < 1.400)$, és a dir, l'àrea entre 1.400 i 1.600 coincideix amb l'àrea a l'esquerra de 1.600 menys l'àrea a l'esquerra de 1.400. Les probabilitats anteriors es poden calcular usant qualsevol programari estadístic (per exemple, R Commander o Excel), i donen com a resultat: $P(X < 1.600) = 0,8413$ i $P(X < 1.400) = 0,1587$, per la qual cosa la probabilitat buscada és de 0,6827, és a dir, un 68,27% dels nens dediquen entre 1.400 i 1.600 hores anuals a veure la televisió.

- Partint de les dades de l'Institut Nacional d'Estadística (INE), el sou mitjà anual d'un treballador és de 26.362 euros. Suposant que els esmentats sous segueixin una distribució normal amb una desviació estàndard de

6.500 euros, quin serà el percentatge de treballadors que superin els 40.000 euros?

En aquest cas, $X \sim N$ (26.362,6.500) i es demana $P(X > 40.000)$. Observem que, com que l'àrea total sota la corba normal és 1, $P(X > 40.000) = 1 - P(X < 40.000)$ = {R Commander o Excel} =1 – 0,9821 = 0,0179, amb la qual cosa només un 1,8% dels treballadors superarien la xifra dels 40.000 euros anuals.

- El temps que s'empra a emplenar un qüestionari en línia segueix una distribució aproximadament normal amb una mitjana de 3,7 minuts i una desviació estàndard d'1,4 minuts. Quin és la probabilitat que es tardi menys de 2 minuts a respondre l'esmentat qüestionari? I que es tardi més de 6 minuts? Trobeu el valor c tal que $P(X < c) = 0,75$ (percentil 75 de la variable).

  En aquest cas, $X \sim N$ (3,7, 1,4). En primer lloc, $P(X < 2)$ = {R Commander o Excel} = 0,1131, és a dir, un 11,31% dels individus que responguin el qüestionari empraran menys de 2 minuts a fer-lo. D'altra banda, $P(X > 6) = 1 - P(X < 6)$ = {R Commander o Excel} = 0,0505, i així, un 5% dels individus trigaran més de 6 minuts a respondre el qüestionari. Finalment, per a el valor c tal que $P(X < c) = 0,75$ s'ha d'escriure el següent codi: *qnorm(1-0.25,mean=3.7, sd=1.4*), amb la qual cosa s'obté un valor aproximat de 4,64 minuts, de tal manera que el 75% dels individus triguen menys de 4,64 minuts a completar el qüestionari (o, dit d'una altra manera, el 25% triguen més de 4,64 minuts a fer-lo).

Les distribucions *t* de Student i *F* de Snedecor

A més de la normal, hi ha moltes altres distribucions de probabilitat contínues que se solen utilitzar en estadística inferencial. N'hi ha una que és l'anomenada distribució t de Student, i una altra és l'anomenada F de Snedecor. Ambdues s'introdueixen a continuació:

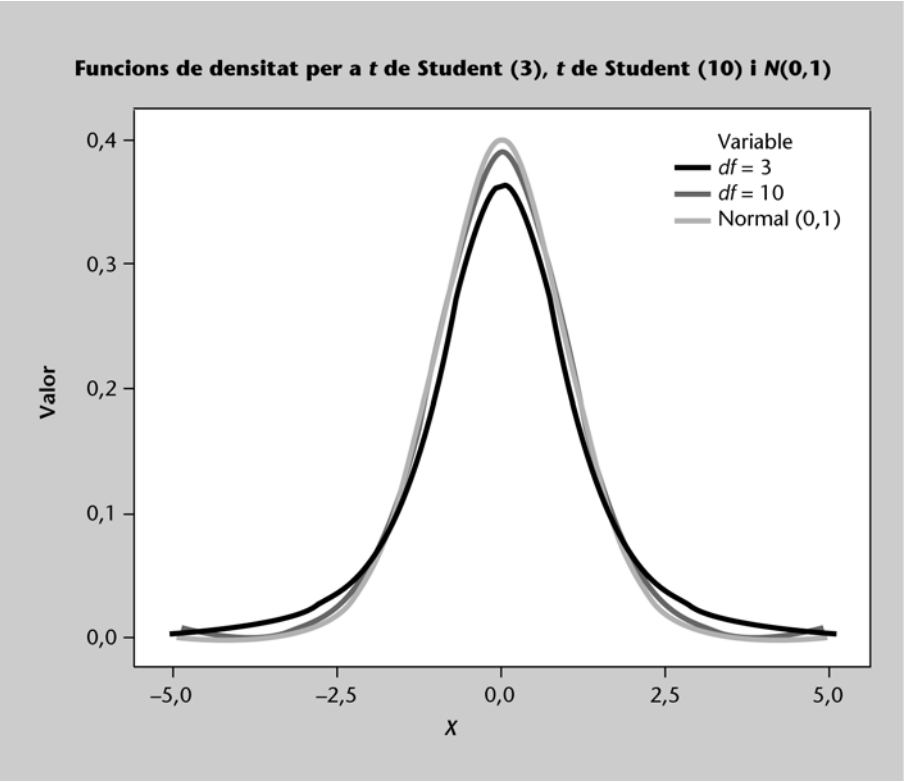
La distribució ***t* de Student** és una distribució simètrica i centrada en l'origen (la seva mitjana i la seva mediana són 0). Aquesta distribució és caracteritzada per un paràmetre anomenat **graus de llibertat** o ***df*** (*degrees of freedom*), on $df > 2$. A la pràctica, $df = n - 1$, on n és la grandària de la mostra que s'estigui analitzant. La figura 29 mostra diverses funcions de densitat de les t de Student, cada una de les quals estan associades a un valor concret del paràmetre df. S'observa com la t de Student s'assembla cada vegada més a una normal estàndard, així que es va incrementant el paràmetre graus de llibertat.

Graus de llibertat

En estadística, el concepte de **graus de llibertat** associats a un conjunt de dades es pot interpretar com el nombre mínim de valors que caldria conèixer per a determinar aques-

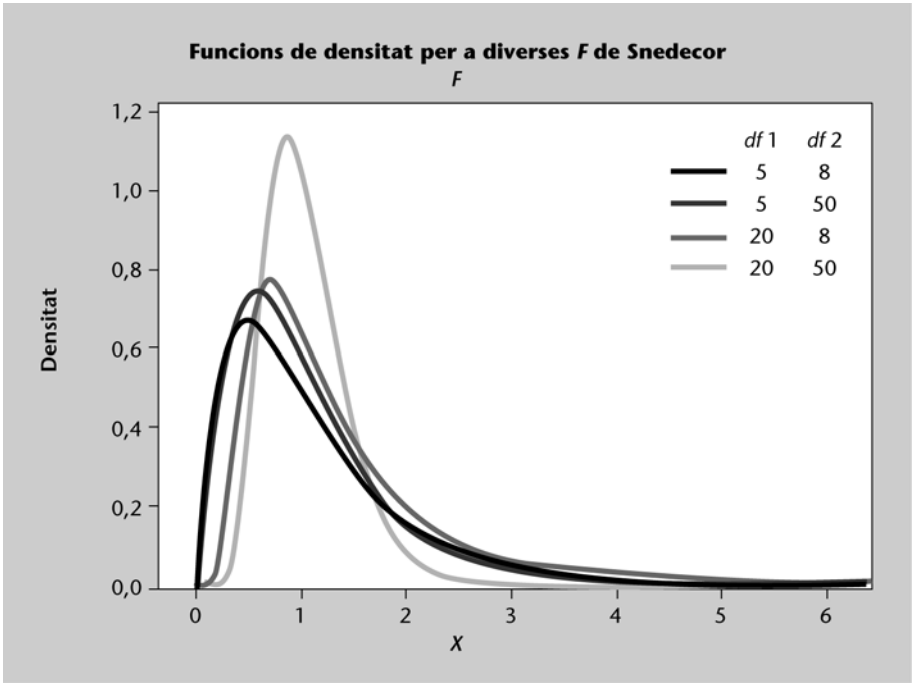
tes dades. Així, per exemple, en el cas d'una mostra aleatòria de grandària N hi hauria N graus de llibertat (no es pot determinar el valor de cap de les dades, fins i tot encara que es conegués el valor dels $N - 1$ restants). Tanmateix, per a un conjunt de N dades de les quals es coneguin $N - 1$, la mitjana mostral tindria $N - 1$ graus de llibertat (fixats els valors de les $N - 1$ dades i de la mitjana, ja quedaria fixat el valor desconegut restant). Així, si tenim un conjunt de 3 observacions de la variable X ($x_1 = 2$, $x_2 = -2$ i $x_3 = a$) (desconegut) i sabem que la mitjana dels tres valors és 0, necessàriament $a = 0$.

Figura 29. Funcions de densitat de *t* de Student segons *df*



Per la seva part, la distribució ***F* de Snedecor** és una altra distribució contínua. La *F* de Snedecor sempre pren valors no negatius (és a dir, una variable que sigui aquesta distribució només pot prendre valors iguals o més grans que 0, mai valors negatius). A més, aquesta distribució no és simètrica, sinó que està esbiaixada a la dreta (figura 30). Així com la normal era caracteritzada per dos paràmetres, μ (mitjana) i σ (desviació estàndard), la *F* de Snedecor també és caracteritzada per dos paràmetres: els **graus de llibertat del numerador**, *df* 1 i els **graus de llibertat del denominador**, *df* 2. Igual com ocorria amb la *t* de Student, per a cada valor d'aquests paràmetres s'obté una funció de densitat diferent i, per tant, una distribució *F* de Snedecor diferent.

Figura 30. Funcions de densitat de *t* de Student segons *df* 1 i *df* 2



Per a calcular probabilitats associades a una *t* de Student o a una *F* de Snedecor, es pot usar programari estadístic o d'anàlisi de dades (R Commander, Excel, etc.) de manera anàloga a com es feia en el cas de la normal. Així, per exemple, si *X* és una variable aleatòria que segueix una distribució *t* de Student amb 10 graus de llibertat, $P(-1{,}74 < X < 1{,}74) = P(X < 1{,}74) - P(X < -1{,}74) =$ {R Commander o Excel} = 0,9438 – 0,0562 = 0,8876 (figura 31).

Figura 31. Probabilitats en una *t* de Student

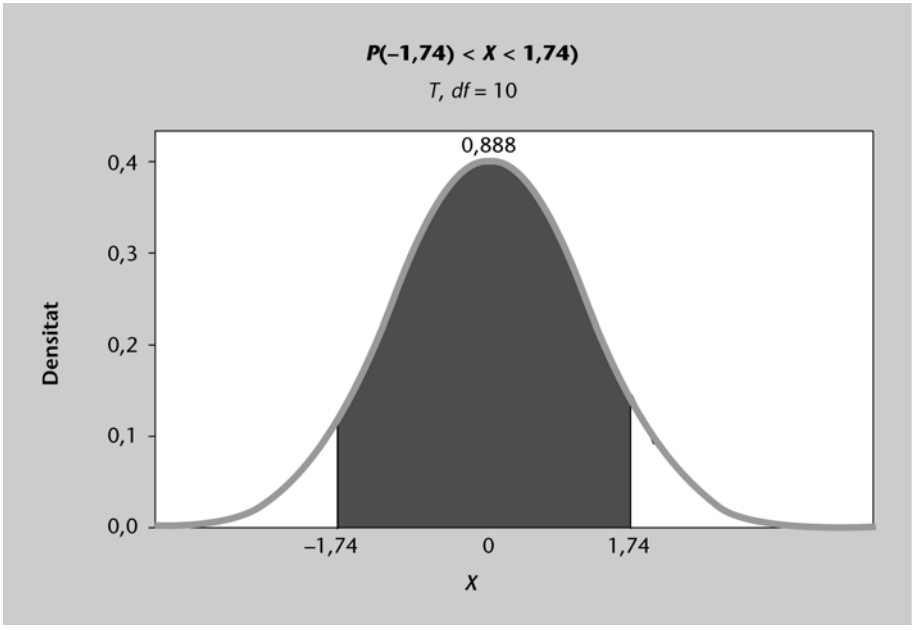
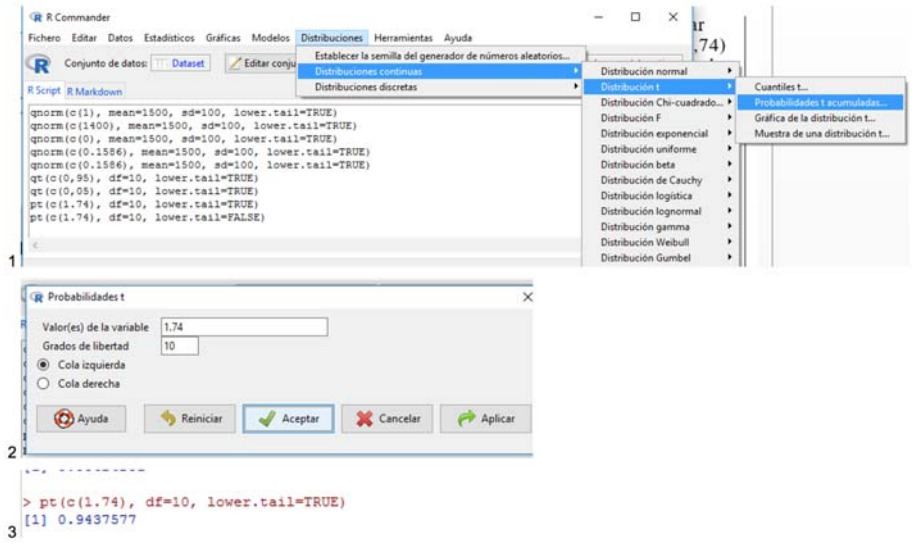


Figura 32. Càlcul de probabilitats en una distribució *t* de Student amb R Commander

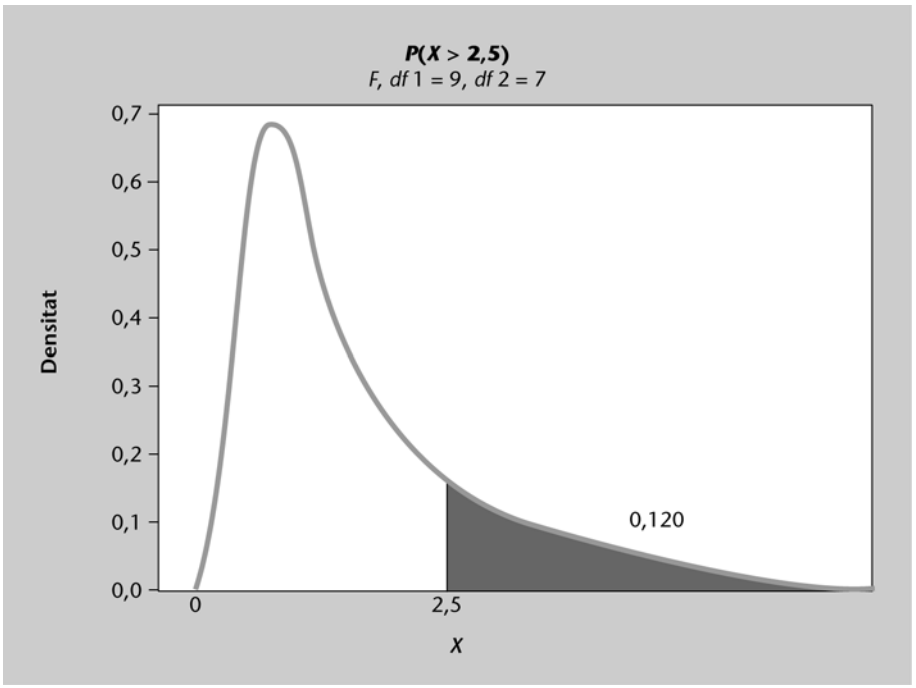


Nota

Noteu que $P(-1{,}74 < X < 1{,}74)$ és representada per l'àrea marcada en la figura 31 (això és, l'àrea compresa entre els valors –1,74 i 1,74). Per a calcular aquesta àrea, es fa $P(X < 1{,}74)$ (és a dir, l'àrea a l'esquerra de l'1,74) i al valor obtingut se li resta $P(X < -1{,}74)$ (és a dir, l'àrea a l'esquerra del –1,74). Per a calcular $P(X < 1{,}74)$ amb R Commander, es fa servir el menú *Distribuciones continuas*> *Distribucion t* > *Probabilidades t acumulades* **(1)**, i s'especifiquen els graus de llibertat (10 en aquest exemple) i el valor de la constant (1,74 en aquest cas) **(2)**. El valor de $P(X < -1{,}74)$ s'obtindria anàlogament.

Finalment, si X és una variable aleatòria que segueix una distribució F de Snedecor amb 9 graus de llibertat en el numerador i 7 graus de llibertat en el denominador, llavors $P(X > 2{,}5) = 1 - P(X < 2{,}5) =$ {R Commander o Excel} $= 1 - 0{,}8797 = 0{,}1203$ (figura 32).

Figura 33. Probabilitats en una *F* de Snedecor



Nota

De manera anàloga al cas de les distribucions binomial i normal, també hi ha taules que permeten calcular, sense necessitat d'utilitzar programari com R Commander o Excel, les probabilitats associades a una distribució *t* de Student o *F* de Snedecor (vegeu, per exemple, http://www.software.dell.com/textbook/distribution-tables).

Resum

En aquest mòdul hem introduït les tècniques bàsiques de l'estadística descriptiva univariant: representació gràfica de dades discretes i contínues, organització de les dades mitjançant taules de freqüències i ús d'estadístics descriptius per a resumir dades. Convé recordar que el tipus de gràfic, taula o estadístic a usar dependrà sempre del tipus de variable considerada (categòrica, quantitativa discreta o quantitativa contínua), així com del tipus d'informació que es vulgui obtenir.

A més, hem explicat també el concepte de probabilitat d'un esdeveniment, que té un paper rellevant en l'anàlisi i predicció del comportament de les variables aleatòries associades a fenòmens quotidians.

Finalment, hem presentat alguns dels principals models matemàtics que es fan servir per a descriure, de manera teòrica, el comportament de variables aleatòries. La distribució binomial, la normal, la t de Student i la F de Snedecor són alguns exemples dels esmentats models. El càlcul de probabilitats associades a variables que es comporten segons algun d'aquests models permet entendre millor el seu comportament i fer estimacions sobre la població d'individus de la qual provenen les dades.

Exercicis d'autoavaluació

1) La taula següent resumeix les respostes ofertes per 200 usuaris d'un portal web a la pregunta "el nivell d'usabilitat del portal és adequat":

Resposta	Freqüència
Totalment d'acord	50
D'acord	75
Lleugerament d'acord	25
Lleugerament en desacord	15
En desacord	15
Totalment en desacord	20

Es demana:
a) Construir un diagrama de barres que permeti visualitzar les respostes obtingudes.
b) Calcular la freqüència relativa d'aparició de cada resposta i construir un diagrama circular per a il·lustrar els esmentats valors.

2) La taula següent conté 40 observacions per al temps transcorregut (en hores) entre la tramesa d'un missatge a un fòrum en línia i la resposta corresponent.

4,0	3,5	3,1	6,0	5,6	3,1	2,9	3,8
4,3	3,8	4,5	3,5	4,5	6,1	2,8	5,0
5,4	3,8	6,8	4,9	3,6	3,6	3,8	3,7
4,1	2,0	3,7	5,7	7,8	4,6	4,8	2,8
5,0	5,2	4,0	5,4	4,6	3,8	4,0	2,9

A partir d'aquestes dades, es demana:
a) Construir un diagrama de tiges i fulles. Useu 1,0 com a unitat d'increment.
b) Construir un histograma.
c) Observeu en les dades algun patró clar? Quina és la moda de la distribució de les dades?

3) La taula següent mostra 20 observacions de la variable aleatòria *nombre de correus electrònics rebuts en un dia*.

3,9	3,4	5,1	2,7	4,4
7,0	5,6	2,6	4,8	5,6
7,0	4,8	5,0	6,8	4,8
3,7	5,8	3,6	4,0	5,6

Es demana:
a) Trobar els estadístics descriptius d'aquesta mostra. Quant val el rang interquartílic? Entre quins dos valors estan compreses el 50% de les dades centrals de la mostra?
b) Construir un diagrama de caixa i bigotis (*boxplot*). Hi ha algun valor anòmal (*outlier*) entre les observacions?

4) Quan s'efectua un control antidopatge a un atleta que no ha pres cap substància, la probabilitat que el test d'un fals positiu és de 0,006. Si durant una competició s'efectua el test a un total de 1.000 atletes que estan lliures de substàncies, quin serà el nombre esperat mitjà de falsos positius?, quina és la probabilitat que el nombre de falsos positius sigui superior a 15?, què es podria pensar si apareixen més de 15 positius?

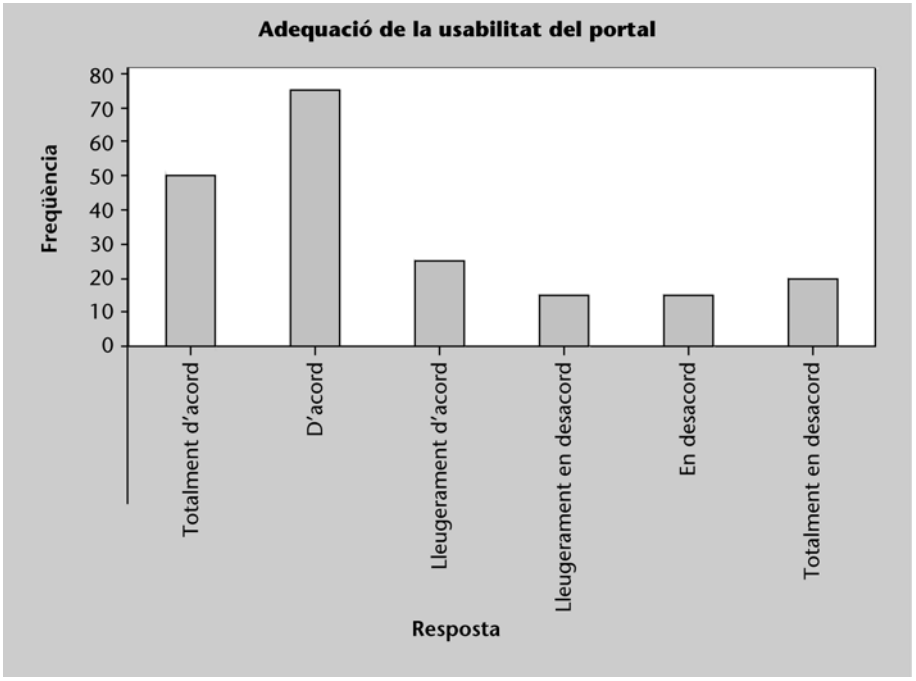
5) D'acord amb l'Institut Nacional d'Estadística, el 9,96% dels adults residents a Espanya són estrangers. A fi de realitzar una enquesta, es pretén contactar amb una mostra aleatòria de 1.200 adults residents a Espanya. Quin serà el nombre esperat (mitjà) d'estrangers que contindrà l'esmentada mostra?, quina és la probabilitat que la mostra contingui menys de 100 estrangers?

6) El temps de durada d'un embaràs és una variable aleatòria que es distribueix de manera aproximadament normal amb una mitjana de 266 dies i una desviació estàndard de 16 dies. Quin percentatge d'embarassos duren menys de 240 dies (uns 8 mesos)?, quin percentatge d'embarassos duren entre 240 i 270 dies (entre uns 8 i 9 mesos)?, a partir de quants dies se situen el 20% dels embarassos més llargs?

Solucionari

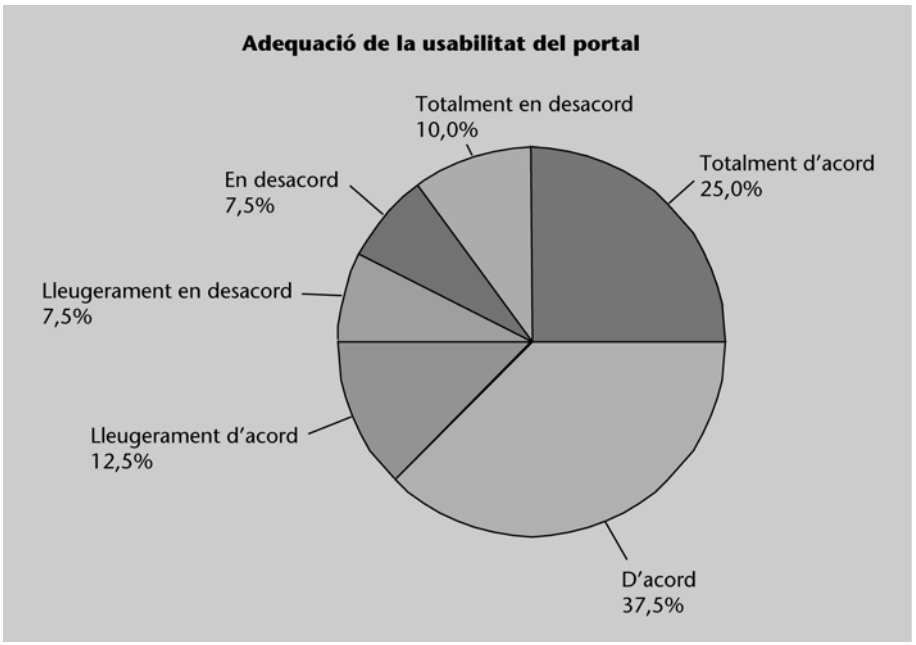
1)

a)



b)

Resposta	Freqüència	Freq. relativa
Totalment d'acord	50	25,0%
D'acord	75	37,5%
Lleugerament d'acord	25	12,5%
Lleugerament en desacord	15	7,5%
En desacord	15	7,5%
Totalment en desacord	20	10,0%
Totals	**200**	**100%**

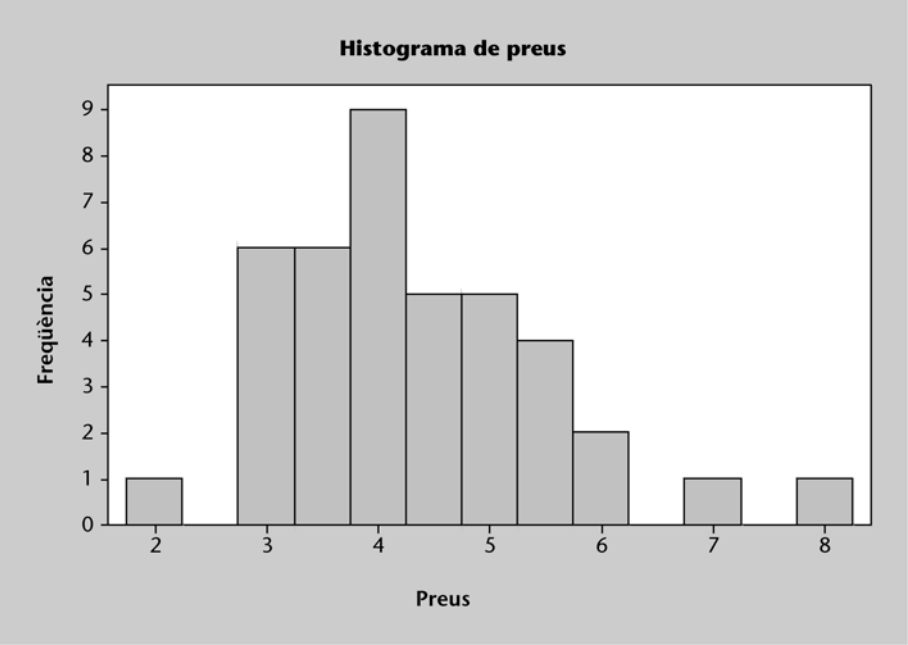


2)

a)

```
> with(Dataset, stem.leaf(Tiempo, na.rm=TRUE))
1 | 2: represents 1.2
 leaf unit: 0.1
            n: 40
   1     2* | 0
   5     2. | 8899
   7     3* | 11
  18     3. | 55667788888
  (5)    4* | 00013
  17     4. | 556689
  11     5* | 00244
   6     5. | 67
   4     6* | 01
   2     6. | 8
HI: 7.8
```

b)



c) Encara que no sembla haver-hi cap patró clar en les dades, sí que s'aprecia –tant a l'histograma com en el gràfic de tiges i fulles una certa forma de campana, amb la part central més elevada i uns extrems o cues més baixes. La moda d'aquest conjunt de dades és 3,8 ja que, com s'aprecia en el diagrama de tiges i fulles, és el valor que més apareix.

3)

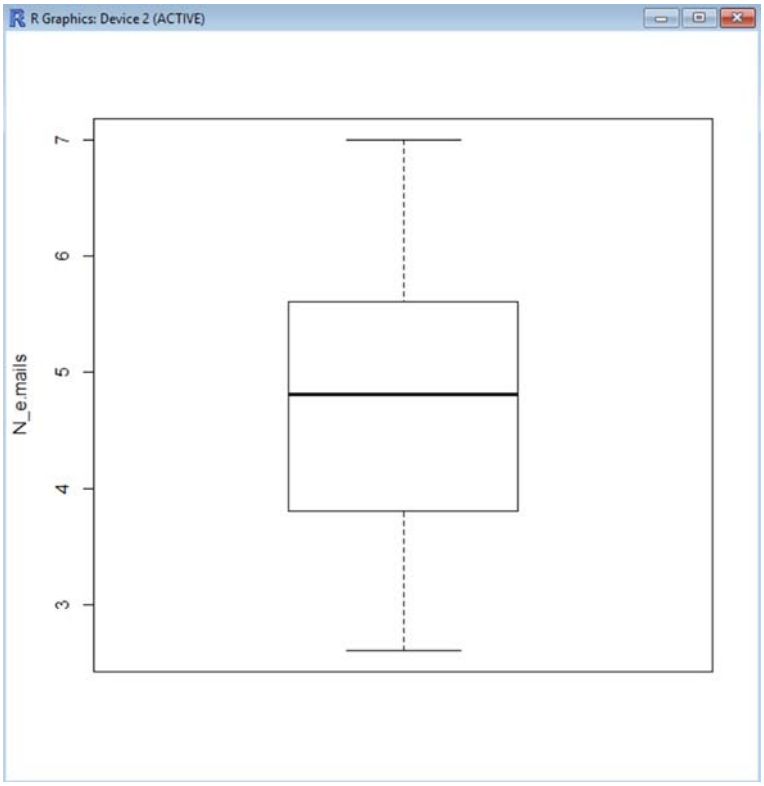
a)

```
> summary(Dataset)
   N_e.mails
 Min.   :2.60
 1st Qu.:3.85
 Median :4.80
 Mean   :4.81
 3rd Qu.:5.60
 Max.   :7.00

> numSummary(Dataset[,"N_e.mails"], statistics=c("mean", "sd", "IQR", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
 mean      sd  IQR  0%  25% 50% 75% 100%  n
 4.81 1.30178 1.75 2.6 3.85 4.8 5.6    7 20
```

El rang interquartílic és $Q3 - Q1 = 5{,}60 - 3{,}85 = 1{,}75$. Entre $Q1 = 3{,}85$ i $Q3 = 5{,}60$ hi ha compreses el 50% de les dades centrals.

b)



No s'observa, en aquest cas, cap valor anòmal (*outlier*), ja que el gràfic no mostra cap símbol *****.

4) En aquest cas, ja que el resultat de cada test pot ser positiviu (amb probabilitat 0,006) o "no positiu" (amb probabilitat 1 – 0,006 = 0,994), la variable aleatòria X = *nombre de falsos positius en 1.000 proves a atletes nets* segueix una distribució binomial de paràmetres n = 1.000 i p = 0,006. En el cas de la binomial, la mitjana o valor esperat és $\mu = n \cdot p = 6$, així, es pot esperar que en aplicar el test a 1.000 atletes *nets* hi hagi 6 falsos positius.

D'altra banda, $P(X > 15) = 1 - P(X <= 15)$ = {R Commander o Excel} =1 – 0,9995 = 0,0005. Per tant, si apareixen més de 15 positius es podria pensar que molt probablement no tots ells siguin falsos.

5) En aquest cas, la variable aleatòria X = *nombre d'estrangers a la mostra* segueix una distribució binomial de paràmetres n = 1.200 i p = 0,0996. Per tant, el valor esperat d'estrangers a la mostra és $\mu = n \cdot p = 119{,}52$, amb la qual cosa la mitjana d'estrangers per a les mostres d'aquestes característiques és d'aproximadament 120.

D'altra banda, $P(X < 100) = P(X <= 99)$ = {R Commander o Excel} = 0,0245. És molt poc probable que una mostra contingui menys de 100 estrangers si aquesta és realment aleatòria.

6) Es considera la variable aleatòria X = *dies que dura un embaràs*. Observem que $X \sim N(266,16)$.

$P(X < 240)$ = {R Commander o Excel} = 0,0521, és a dir, el 5,2% dels embarassos duren menys de 8 mesos.

$P(240 < X < 270) = P(X < 270) - P(X < 240)$ = {R Commander o Excel} = 0,5987 – 0,0521 = 0,5466, de tal manera que el 55% dels embarassos duren entre 8 i 9 mesos.

Finalment, es demana el valor c tal que $P(X > c) = 0{,}20$, és a dir: $P(X < c) = 1 - P(X > c) = 0{,}80 \rightarrow c$ = {R Commander o Excel} = 279,47, amb la qual cosa, el 20% dels embarassos supera els 279 dies.