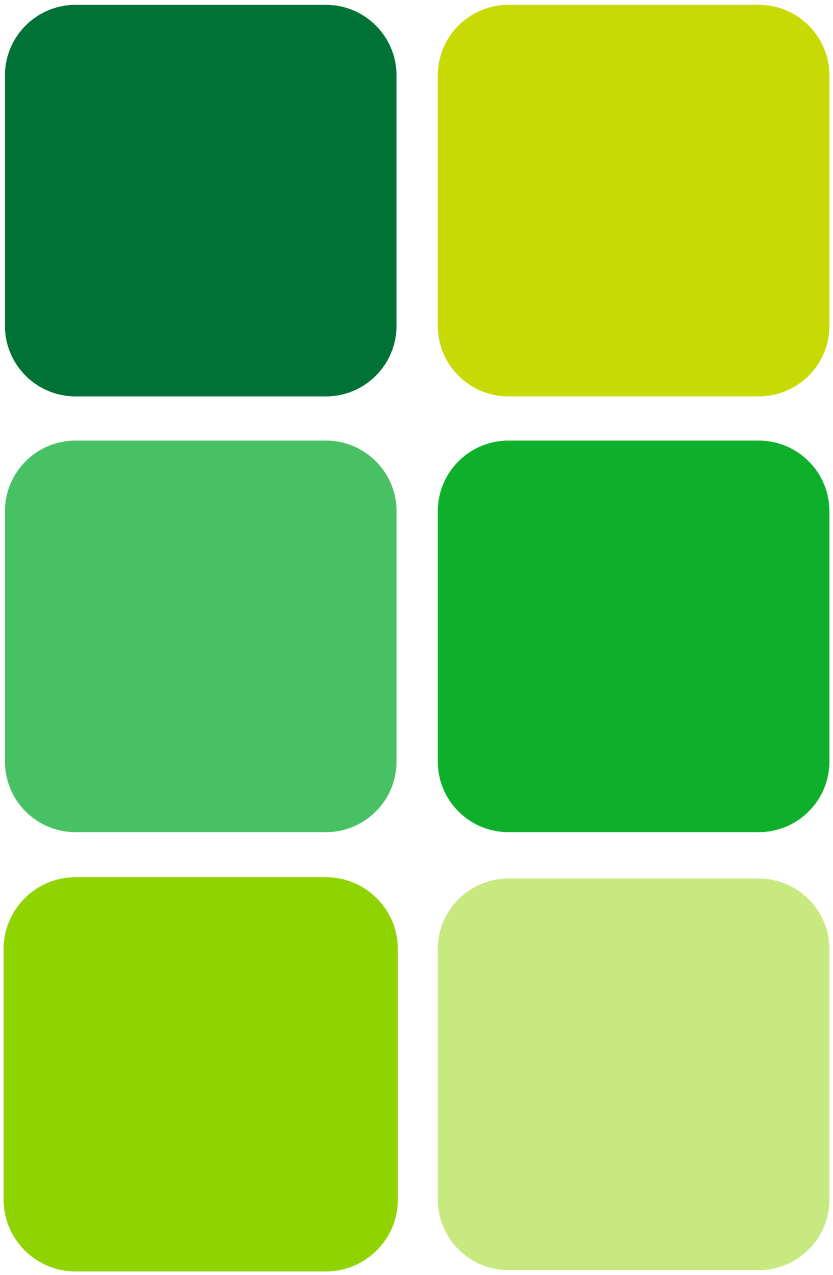


www.sapientia.uji.es | 34

UNIVERSITAT JAUME·I

Estadística bàsica per a l'Enginyeria Tècnica en Informàtica de Gestió

Pablo Gregori Huerta

Irene Epifanio López

Estadística bàsica per a l'Enginyeria Tècnica en Informàtica de Gestió

Pablo Gregori Huerta
Irene Epifanio López

UNIVERSITAT JAUME·I

DEPARTAMENT DE MATEMÀTIQUES

■ Codi d'assignatura IG12

Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
http://www.tenda.uji.es e-mail: publicacions@uji.es

Col·lecció Sapientia, 34
Primera edició, 2010
www.sapientia.uji.es

ISBN: 978-84-693-0998-8

Índex

PART I
INTRODUCCIÓ

Capítol 1

Introducció a l'Estadística

1.1 Breu història de l'Estadística

- L'Estadística va nàixer com la "ciència de l'Estat":
  - Censos de població per poder a formar els exèrcits.
  - Censos de béns (collita, ramaderia, etc.) per a una adequada recaptació d'impostos...

  Només descrivia la realitat.
- D'altra banda, excavacions arqueològiques apunten que els jocs d'atzar tenen més de 40000 anys (vegeu la Figura 1.1)

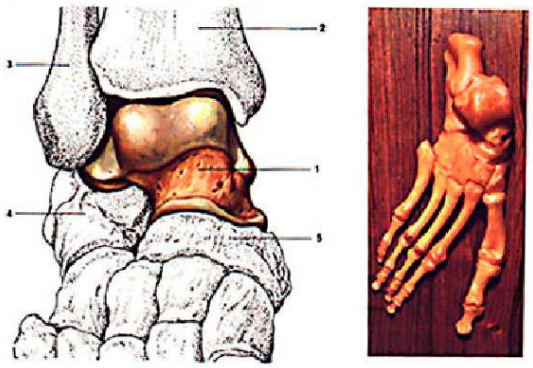


Figura 1.1: Os astràgal, i la seua posició al peu. Precursor del dau actual a la prehistòria

- En les antigues Grècia i Roma s'associava l'atzar a la voluntat divina.
- En el Renaixement es qüestionen les explicacions teològiques. L'atzar s'estudia des d'un nou punt de vista. Apareix el concepte d'**equiprobabilitat**, en els resultats de llaçar un dau ben construït.
- Es pensa que l'origen del **càlcul de probabilitats** és la resolució del problema (1654) de repartir els diners de les apostes si els jugadors es

veuen forçats a finalitzar la partida sense haver-hi un guanyador (batuda policial, les apostes estaven prohibides).

- La Física i l'Astronomia van impulsar el càlcul de probabilitats: com es podien combinar diferents mesuraments d'una magnitud per a obtenir-ne un de més precís? (D. Bernoulli)

- P. Laplace va ser el primer a definir el concepte de **Probabilitat**, va desenvolupar la llei **Normal** i es va plantejar el problema de pronosticar el valor d'una variable a partir de valors coneguts d'altres variables relacionades.

- La teoria de la **selecció natural** de C. Darwin està molt lligada a l'Estadística (variabilitat en l'espècie —atzar—, i supervivència —correlació— entre variables).

- F. Galton, cosí de Darwin, va encunyar el terme de **regressió**, estudiant la relació entre les alçades de pares i fills.

- Fisher introdueix la metodologia actual:

  1. L'elecció d'un model a partir de les dades empíriques.
  2. La deducció de les propietats matemàtiques del model.
  3. L'estimació dels paràmetres del model si es considera com convenient.
  4. La validació final del model mitjançant un contrast d'hipòtesi.

- Al segle XX, els mètodes estadístics s'estenen a àrees molt diverses:

  – Enginyeria (control de qualitat, predicció i control de processos, codificació de senyals).
  – Física (teoria cinètica dels gasos).
  – Antropologia, psicologia, medicina, economia, etc.

- La investigació en problemes militars durant la Segona Guerra Mundial va derivar en un nou camp conegut com **Investigació Operativa**.

- El desenvolupament de les computadores ha fet que molts mètodes només teòrics en el passat es puguen aplicar amb bons resultats.

1.2 Objectius de l'Estadística

- En una situació concreta, on cal prendre una decisió que pot comportar conseqüències positives (guanys) o negatives (pèrdues), tenir informació és millor que no tenir-la (és a dir, amb la informació és més fàcil encertar la decisió).

- En moltes situacions, la informació està feta de milers o milions de peces d'informació (dades), que és intractable de manera directa, sense processar.

- L'objetiu fonamental de l'aplicació de l'Estadística és que el seu usuari (siga un investigador, una empresa o un jugador) tinga els elements necessaris per prendre una decisió el més encertada possible. Per açò, l'Estadística li pot aportar:

    1. Una forma adequada d'obtenir les dades del problema que es planteja (**mètodes de mostratge**).
    2. Si es disposa d'una gran quantitat de dades, una forma d'assimilarles i extraure'n la informació més rellevant (**estadística descriptiva**).
    3. Una forma d'"intuir" totes les dades que no haja sigut capaç de recollir (**models de probabilitat, estadística inferencial**).
    4. Una forma de calcular els riscos que impliquen les distintes decisions que puga prendre a partir de les dades obtingudes (**càlcul de probabilitats**).

1.3 Exemples de problemes que involucren grans poblacions

L'Estadística serveix per a tractar problemes que afecten grans poblacions.

- Exemple 1

    - **Responsable:** El de govern d'un país, regió, ciutat...
    - **Problema:** administració d'un país (millora econòmica, social, salut, etc.)
    - **Poblacions d'interès:** població total, població activa, població de persones majors, població infantil, població de recursos naturals, població d'industries, població de preus, etc.

- Exemple 2

    - **Responsable:** una empresa de producció, de transformació, distribuïdora, de serveis...
    - **Problema:** obtenir beneficis (comprar més barat, vendre més car, guanyar clients, retallar despeses...).
    - **Poblacions d'interès:** població de proveïdors, població de clients potencials, població de clients reals, població d'unitats produïdes o venudes, població de despeses per manteniment, població d'ingressos per vendes, població d'indústries competidores...

- Exemple 3

  - **Responsable:** investigadors de ciències de la salut (farmàcia, medicina...)
  - **Problema:** trobar mitjans per a millorar la salut.
  - **Poblacions d'interès:** població d'éssers humans (com a beneficiaris), població de malalts d'una malaltia concreta, població d'òrgans humans concrets, població de virus, població de bacteris, població d'elements orgànics, població de substàncies químiques, població de malalties humanes, població de variants d'una malaltia concreta...

- Exemple 4

  - **Responsable:** investigadors de ciència de materials.
  - **Problema:** trobar els millors materials per a tasques concretes.
  - **Poblacions d'interès:** població de diversos tipus de materials, població de característiques (resistència, torsió, fusió, etc.)...

- Exemple 5: un exemple més proper.

  - **Responsable:** professor.
  - **Problema:** transferir a l'alumnat la sèrie de competències de la seua assignatura.
  - **Poblacions d'interès:** població d'alumnes matriculats (els seus coneixements, actituds, recursos d'aprenentatge, etc.).

  Quins factors dels alumnes (i que presenten variabilitat) influencien el correcte desenvolupament de la situació?

  - Coneixements previs.
  - Actitud vers les Matemàtiques.
  - Hàbits d'estudi i de treball.
  - Ús de la calculadora.
  - Ús de l'ordinador.
  - Temes d'interès personals.
  - Sexe, edat, idees polítiques, equip de futbol... (?)

  Aleshores el professor demana aquests aspectes en una enquesta que pot tenir com a resultat la Taula 1.1, on les columnes indiquen la informació dels alumnes segons:

  - **nivell:** resultat d'una xicoteta prova relacionada amb els coneixements que el professor suposa que té el seu alumnat (0 = mín, 10 = màx).

– **actitud:** actitud vers les assignatures de caire matemàtic (0 = mín, 4 = màx).
– **estudi:** nombre d'hores per setmana que es té previst dedicar a l'assignatura.
– **calc:** autoavaluació respecte al domini de la calculadora (0 = mín, 2 = màx).
– **ordin:** autoavaluació respecte al domini de l'ordinador (0 = mín, 2 = màx).
– **oci:** afició preferida, només una (lliure).

Taula 1.1: Taula de dades arreplegades mitjançant una enquesta.

id	nivell	actitud	estudi	calc	ordin	oci
01	7	3	3	2	2	esports
02	1	0	1	1	2	internet
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮

1.4 Vocabulari bàsic

Les següents paraules formen el primer glossari de paraules especialitzades de l'Estadística:

- **Individu:** Entitat mínima que conforma, juntament amb moltes altres, l'objecte de l'estudi.
- **Població:** Conjunt d'individus que forma l'objecte de l'estudi.
- **Variable:** Factor concret que es desitja analitzar, i que presenta variabilitat en els distints individus de la població.
- **Dada:** Valor concret de la variable prés per a un individu concret. També es diu "observació" o "medició".
- **Mostra:** Conjunt d'individus dels quals es coneix la dada, o també "conjunt de les dades arreplegades" (la mostra és un subconjunt de la població).
- **Variable qualitativa:** Variable les dades de la qual corresponen a una llista de possibles estats (o atributs), anomenats **categories**.
- **Variable quantitativa:** Variable les dades de la qual expressen quantitats de comptar o dins d'una escala de mesurament. Es poden comparar (per exemple, $3 \leq 4$) i apreciar distàncies (per exemple, 3 està més prop de 4 que de 5), i fins i tot operar.

- **Mostra univariant:** Mostra amb dades corresponents a una sola variable.

- **Mostra multivariant:** Mostra amb dades corresponents a més d'un factor (cas particular, les **bivariants**).

Amb aquestes comença la primera part, que es dedica a la descripció sistemàtica de les mostres de dades.

1.5 Pràctica R: 1. Introducció a R amb un exemple il·lustratiu

Objectius

El programa R està avalat per la comunitat científica. Els objectius d'aquest capítol són:

- Conéixer la web de desenvolupament del programa (instal·lació, documentació, etc.).

- Iniciar una primera sessió de contacte (obrir, ajudes, gràfics, directori de treball, assignació de variables, tancament de sessió, arxius de treball) per a familiaritzar-se amb l'entorn de treball.

- Observar un exemple molt complet de la potencialitat del programa (loteries de New Jersey).

El projecte R

A l'adreça `http://www.r-project.org/` es troba tot el material relatiu al desenvolupament del programa. Des d'aquesta pàgina es pot accedir a la rèplica (*mirror*) a Espanya (`http://cran.es.r-project.org/`) i descarregar l'instal·lador de la versió més recent per a Windows o Linux, i a `Documentation > Manuals`, on es recomana *"An Introduction to R"*. Una traducció al castellà d'eixe llibre es troba seguint l'enllaç `contributed documentation` (concretament a `http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf`), a més del document *"R para principiantes"* (`http://cran.r-project.org/doc/contrib/rdebuts_es.pdf`).

Per altra banda, existeix un paquet particular, disponible a la web, anomenat `Rcmdr` (*R commander*) que crea una interfície gràfica d'usuari en forma de menús que facilita la utilització a aquells que es troben insegurs amb la programació.

Funcionament general: exemple de sessió

Algunes característiques generals de R són:

- R és un programa de llicència lliure desenvolupat per la comunitat científica i que gaudeix d'un gran prestigi.
- R és un programa amb llenguatge de programació propi. Té implementats molts tipus d'objectes i funcions convenients per a les labors estadístiques.
- En una sessió de R, l'usuari ha d'invocar les funcions que realitzen les tasques desitjades, o programar-les prèviament si no formen part encara de la distribució.
- Quan una tasca implica l'execució de diverses línies de codi, és molt aconsellable escriure en un editor de textos (notepad, emacs, vi...), per anar corregint els possibles errors.
- El programa té un directori de treball (path) per defecte on busca o guarda arxius requerits. És important tenir aquest directori controlat, per a recollir els productes obtinguts. Es pot canviar el directori de treball amb les opcions del menú `File > Change dir...`. Si algun fitxer està en un altre lloc, cal indicar tot el `path` per què el trobe.
- L'operador d'assignació (per a atribuir valors constants a variables) és `<-`, encara que també es pot fer servir el `=`.

La millor manera de comprendre la idiosincràsia del programa és encetar una sessió:

- **Obrir sessió:** premeu sobre la icona del programa. Una interfície gràfica d'usuari (GUI) s'obri, amb una finestra de comandamanents.
- **Alguns comandaments:** copia i desa al prompt de R.

```
# les línies que comencen amb #
# són considerades comentaris
dni <- 12345678    # escriu el teu dni sense lletra
nom <- 'Pere'      # escriu el teu nom
cognom1 <- "Giner" # escriu el teu cognom
edat <- 21         # escriu la teua edat
dni
nom
cognom1
edat
# el nom de les variables fa que es mostre
# el seu contingut a la pantalla
ls()               # mostra variables definides
help(plot)         # nova finestra amb ajuda
dades <- scan('dades-s1.txt') # trobes l'arxiu?...
dades
sum(dades)         # suma els valors
dades < 3          # compara els valors amb 3
```

```
plot(dades)          # nova finestra gràfica
q()                  # tancament de sessió
```

- **Tancar la sessió:** la funció `q()` tanca la sessió i pregunta a l'usuari si vol desar l'espai de treball. L'espai de treball (`Workspace`) consisteix en un parell d'arxius:
  - `.RData`: arxiu que serveix per a encetar una nova sessió de R (fent doble clic) amb totes les variables definides a la memòria. Serveix per a continuar la sessió anterior com si no s'haguera tancat mai.
  - `.RHistory`: arxiu de text que conté totes les línies introduïdes al prompt de R.

En principi, si es treballa amb un editor de text i es guarda el codi, no és necessari guardar l'espai de treball (només cal copiar i desar tot el codi d'una vegada i es processarà). Només és útil per a reprendre grans projectes que empren molt de temps de càlcul.

Exemple il·lustratiu

Descarreguem l'arxiu `lottery.zip` del web [13] i extraiem el seu contingut en el directori de treball de la sessió de R actual. El text següent es basa en un exemple bastant interessant d'una anàlisi de dades i la forma en què es treballa amb R (veure [5]). No és important comprendre del tot els comandaments que apareixeran, encara que és recomanable parar atenció com a primer contacte. Copieu i deseu a l'editor de comandaments de R cada bloc de codi que va apareixent i observeu els resultats.

Una loteria de New Jersey consisteix en un sorteig diari, on el jugador aposta 50 cèntims per un número qualsevol entre el 000 i el 999 (pot fer tantes apostes com vulga, a 50 cèntims cadascuna, repetint número o no).

Una volta fet el sorteig, la meitat de la recaptació va a les arques públiques i l'altra meitat és per a pagar els guanyadors. El premi es reparteix a parts iguals.

Llegim les dades dels números guanyadors i els premis per aposta del període maig 1975 - març 1976 (254 dies)

```
lottery.number <- scan("lotterynumber.txt")
lottery.payoff <- scan("lotterypayoff.txt")
```

Inspeccionem els 254 números premiats

```
lottery.number
```

i mirem quant s'ha guanyat en cada sorteig

```
lottery.payoff
```

(per exemple, en el primer sorteig va eixir premiat el número 810 i va donar un premi de \$190).

Ara podem veure com estan distribuïts els números premiats (si el joc és legal, tots els números són igualment probables, i no hauria d'haver-hi desequilibris), en un gràfic molt útil

```
hist(lottery.number)
```

Si mirem la distribució dels premis per aposta (això ja no depén de la legalitat, sinó de les eleccions dels apostants) en un gràfic, podem observar que en poques ocasions el premi ha estat major de \$600 o menor de \$100

```
hist(lottery.payoff)
```

El premi més alt concedit en aquell període va ser

```
max(lottery.payoff)
```

i el número guanyador va ser

```
lottery.number[ lottery.payoff == max(lottery.payoff) ]
```

Per contra, el menor premi concedit va ser

```
min(lottery.payoff)
```

corresponent al número

```
lottery.number[ lottery.payoff == min(lottery.payoff) ]
```

Alguna intuïció? Anem a investigar si hi ha relació entre el valor del número premiat i el valor del premi corresponent (en principi, números "populars" premiats resultaran en premis menuts, mentre que números "impopulars" premiats donaran lloc a premis majors

```
plot(lottery.number, lottery.payoff)
```

Com que no s'observa bé si hi ha una pauta, passem una línia "mitjana" que mostra la tendència

```
lines( lowess(lottery.number, lottery.payoff, f=.2) )
```

Sembla que els números de 000 a 099 tenen premis molt alts, per què? I els números de 100 a 300 tenen els menors premis. Per què?

Anem a identificar els números amb majors i menors premis. Passeu el ratolí sobre els punts del gràfic

```
identify( lottery.number, lottery.payoff, lottery.number )
```

499, 767, 020, 077, 919... què tenen en comú? Passem als sorteigs d'altres períodes

```
lottery2.number <- scan("lottery2number.txt")
lottery2.payoff <- scan("lottery2payoff.txt")
lottery3.number <- scan("lottery3number.txt")
lottery3.payoff <- scan("lottery3payoff.txt")
```

Fem un gràfic comparatiu dels premis concedits en els 3 períodes

```
boxplot(lottery.payoff, lottery2.payoff, lottery3.payoff,
        ylab="Premios", xlab="Año")
```

Marquem una línia que indica $500

```
abline(h=500)
```

La caixa indica la meitat central dels premis (perc 25 a perc 75), la ratlla dins la caixa indica un valor central representatiu dels premis (mediana), mentre que els bigots arriben fins on els valors de premis es consideren "normals". Els punts aïllats són valors considerats estranyament "anormals" (atípics).

Conclusió: Els premis mitjans es mantenen, però la gent (com nosaltres) s'adona de les pautes (números impopulars) i es diversifiquen les apostes, raó per la qual cada vegada hi ha menys números "impopulars".

1.6 Pràctica R: 2. Mostres de dades univariants (la classe `vector`)

Objectius

L'Estadística tracta l'anàlisi de dades. Les dades més senzilles són aquelles que es poden representar com una etiqueta o valor numèric. Una llista ordenada de dades d'aquest tipus pot formar un vector. Per això, R té implementat la classe `vector` com la més bàsica. Per tant, l'objectiu d'aquest capítol és el domini en la manipulació i exploració de les dades contingudes en un vector.

Declaració de vectors

Un vector és una llista ordenada, d'objectes del mateix tipus. En R estan implementats tres tipus de vectors: numèrics (`numeric`), de cadenes de caràcters (`character`) i de valor lògics `TRUE-FALSE` (`logical`).

Observeu el següent codi per a comprendre com es declaren i funcionen els vectors:

```
v1 <- 3             # el vector més curt té 1 component només
v1
v2 <- c(v1, 0, 7, 9)     # la funció c(...) concatena vectors
v2
length(v1)           # llargària de v1 = nombre de components
length(v2)
v3 <- c('XXX', 'A', 'B')      # vector de cadenes de caràcters
v3
v4 <- c(TRUE, TRUE, FALSE)           # vector de valors lògics
v4
v5 <- runif(n=100)                      # 100 valors aleatoris
v5
```

Escriure el nom d'una variable fa que es mostre el seu contingut a la pantalla, i és molt recomanable per a comprovar que els comandaments funcionen com s'espera. Quan es mostra un vector a la pantalla, el número que ix en cada línia, al costat esquerre i entre claudàtors (`[ ]`), indica la component (posició) del primer valor d'aqueixa línia.

Principals funcions que actuen sobre vectors

- `c()`: funció que admet com a arguments qualsevol nombre de vectors.
  Torna un únic vector, resultant de concatenar els vectors arguments de la funció.
  Exemple: `v2 <- c(v1, 0, 7, 9)`

- `length()`: funció que admet com a argument un únic vector.
  Torna la llargària del vector argument (és a dir, el seu nombre de components).
  Exemple: `length(v2)`

- `[ ]`: funció que admet com a argument dos vectors: (1) un vector, del qual es va a seleccionar unes components i (2) un altre vector que indicarà quines components es seleccionen
  Torna un vector que és el "subvector" del primer argument indicat per les components del segon argument.
  Exemples: `v5 <- v2[3] # v5 te 1 comp: la 3a comp de v2`
  `v6 <- v2[c(1,4)] # v6 te 2 comp's: la 1a i 4a de v2`
  També es pot fer servir per modificar components concretes d'un vector.
  Exemple: `v5[1] <- 9 # la 1a comp de v5 ara val 9`

- `rev()`: funció que admet un únic vector com a argument.
  Torna un vector, l'argument del qual amb les components capgirades (la que era la primera de l'argument és l'última del resultat, i la que era última de l'argument és la primera del resultat).
  Exemple: `v7 <- rev(v2)`

- `sort()`: funció que admet un únic vector com a argument.
  Torna un vector, l'argument del qual amb les components ordenades de menor a major.
  Exemple: `v7 <- sort(v2)`

- `unique()`: funció que admet un únic vector com a argument.
  Torna un vector format només per les distintes components del vector argument, sense repetir.
  Serveix per a examinar els valors diferents d'un vector llarg.
  Exemple: `v7 <- unique(v4)`

- `!, &, |, xor`: operadors lògics (negació, i, o, o exclusiu). **S'apliquen sobre vectors lògics**

- **A partir d'ací les funcions s'apliquen sobre vectors numèrics**

- `+, -, *, /, ^...`: operadors aritmètics (suma, resta, producte, divisió, potència,... i altres).

- `==, !=, <, <=, >, >=`: operadors de comparació (igual, distint, menor, menor o igual, major, major o igual).

Els operadors admeten com a arguments dos vectors, i actuen component a component i torna, com resultat, el vector dels resultats. Si els vectors operats no són de la mateixa llargària, el més curt s'autoreplica per a poder fer l'operació. Dóna un missatge d'avís si les llargàries no coincideixen.

- `log()`, `exp()`, `log10()`, `log2()`, `sin()`, `cos()`, `tan()`, `asin()`, `acos()`, `atan()`, `abs()`, `sqrt()...`: funcions matemàtiques usuals. Admeten com a argument un vector i actuen sobre cada component separadament, i torna un vector de la mateixa llargària que l'original.
- `sum()` suma les components del vector argument.
- `prod()` multiplica les components del vector argument.
- `max()` torna el valor de la major component del vector argument.
- `min()` torna el valor de la menor component del vector argument.
- `which.max()` torna la posició de la major component del vector argument.
- `which.min()` torna la posició de la menor component del vector argument.

Creació de vectors: replicació i progressió aritmètica

Les següents funcions ajuden a crear dos vectors particulars de manera automàtica:

- `rep()`: replicar valors constants. Agafa com primer argument un vector, i el concatena amb si mateix el nombre de voltes especificat pel segon argument `times`.

  Exemple: `rep(x=c(1,2,3), times=20)`

- `seq()`: seqüència. Fa una progressió aritmètica segons els valors dels arguments:

  - `from` = des de,
  - `to` = fins a,
  - `by` = salt o diferència entre valors successius,
  - `length` = llargària total del vector.

  Per a definir una successió aritmètica **n'hi ha prou a fixar tres dels quatre arguments** anteriors.

  Exemples: `seq(from=0, to=10, by=1)`,
  `seq(from=0, to=10, length=5)`.

Exploració de les dades contingudes en un vector

L'operador `[ ]` aplicat a un vector forma un subvector amb les components indicades dins dels claudàtors, i es pot fer de tres formes. Per exemple, si partim del vector $v30 = (91, 1, 19, 59, 40, 96, 79, 16, 17, 25)$, podem seleccionar un subvector de $v30$...

- ...indicant dins de `[ ]` un vector de components seleccionades.

  ```
  v30[c(1,5,6)]
  ```

- ...indicant dins de `[ ]` un vector de components que es volen excloure, i escrivint un signe negatiu al davant.

  ```
  v30[-c(1,5,6)]
  ```

- ...indicant dins de `[ ]` un vector lògic que tinga el valor `TRUE` a les components seleccionades i `FALSE` a les excloses.

  ```
  v30[c(FALSE, FALSE, TRUE, TRUE, FALSE, TRUE, TRUE, FALSE,
  TRUE, FALSE)]
  ```

Aquesta darrera forma d'obtenir subvectors és molt convenient per a l'exploració de vectors. Observeu el següent exemple:

```
set.seed(123456789)
v31 <- runif(n=100)           # generem el vector de partida
v31           # aquest és el vector amb les dades a explorar
#
# quantes dades del vector són inferiors o iguals a 0.5?
v31 <= 0.5                    # escrivim la condició buscada
v31[v31 <= 0.5]           # aquest es el subvector d'interès
length(v31[v31<=0.5])             # i aquesta és la resposta
#
# quantes dades del vector són superiors a 0.33?
length(v31[v31 > 0.33])                          # directament
```

Lectura-escriptura de dades en arxius

La funció `write()` escriu un arxiu amb les dades d'un vector. Els arguments principals són `x` (el vector les dades del qual volen escriure en arxiu) i `file` (l'arxiu que serà creat)

```
v1 <- runif(n=100)                       # un vector numeric
write(x=v1, file='dades-1-s2.txt')
v2 <- rep(x=c('SI', 'NO'), times=100)  # un vector character
write(x=v2, file='dades-2-s2.txt')
v2
```

La funció `scan()` llegeix un arxiu i carrega les dades a un vector. Els arguments principals són `file` (el nom de l'arxiu) i `what` (quin tipus de dades va a llegir, numèriques si no se n'especifica un altre).

```
v3 <- scan(file='dades-1-s2.txt')
v3
v4 <- scan(file='dades-2-s2.txt', what='character')
v4
```

Representació gràfica de dades numèriques

La funció `plot()` s'encarrega de dibuixar gràfics en dues dimensions. Encara que és una funció molt completa, per a representar les dades (numèriques) d'un vector, ja que només hi ha una dimensió (la de les dades), el gràfic situa:

- A l'horitzontal, la posició (component)
- A la vertical, el valor numèric de la component

D'aquesta manera, les dades es mostren ordenades per component. Si es vol accentuar l'evolució dels valors quan avança la component, es pot utilitzar l'argument `type='l'`. Observeu l'exemple:

```
v <- runif(n=10)   # el vector numèric que s'ha de dibuixar
v
plot(v)            # el gràfic normal
plot(v, type='l')  # el gràfic amb els valors units per línies
```

Si les dades s'han recopilat en un ordre temporal, que queda traslladat a l'ordre de les components, aquest gràfic pot usar-se per interpretar la influència de l'ordre temporal sobre el valor de les dades (molt important en l'anàlisi de dades).

Exercicis d'ensinistrament

1. Escriviu el codi que declara en R el vector `w1=(1, 3, 5, ..., 67)`
2. Escriviu el codi que declara en R el vector `w1=(1, 1/2, 1/3, ..., 1/128)`
3. Escriviu el codi que declara en R el vector `w1=(sen(`$\frac{2\pi}{100}$`), sen(`$\frac{4\pi}{100}$`), sen(`$\frac{6\pi}{100}$`), ..., sen(`2π`))`
4. Escriviu el codi que declara en R el vector
   `w2 = (`$\underbrace{\texttt{'A', ..., 'A'}}_{50}$`,`$\underbrace{\texttt{'B', ..., 'B'}}_{23}$`,`$\underbrace{\texttt{'C', ..., 'C'}}_{15}$`)`.
5. Escriviu el codi que declara en R el vector
   `w3 = (`$\underbrace{\texttt{TRUE, ..., TRUE}}_{200}$`,`$\underbrace{\texttt{FALSE, ..., FALSE}}_{100}$`)`.
6. Calculeu amb R la suma $1 + 2 + \cdots + 10000$

   Sol.: 50005000

7. Calculeu amb R una aproximació de la sèrie $\sum_{n=1}^{\infty} \frac{1}{n^2}$, per exemple amb 1000 termes (és a dir, $1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{1000^2}$).

   Sol.: 1.643935

8. Escriviu el codi

```
set.seed(123)
w4 <- runif(n=1000)
```

   i contesteu les següents qüestions:

   (a) Quin és el valor de la 237-èsima component de `w4`?
   Sol.: 0.1977447

   (b) Quant sumen totes les components de `w4`?
   Sol.: 497.2778

   (c) Quantes components de `w4` són inferiors o iguals a 0.3333?
   Sol.: 334

   (d) Quantes components de `w4` estan entre 0.5 i 0.75? (ambdos inclosos)
   Sol.: 245

   (e) Quant sumen les components de `w4` que són inferiors o iguals a 0.3333?
   Sol.: 55.84107

9. Emmagatzemeu les dades contingudes a l'arxiu `s2-concentracio.txt` a una variable. Si aquestes representen els valors de concentració de plom a l'aire durant les observacions fetes un dia, cada 5 minuts:

   (a) Quants mesuraments s'han fet al llarg del dia? Sol.: 288

   (b) Quin ha sigut el mesurament màxim registrat? Sol.: 47.34

   (c) En quants mesuraments s'ha superat la concentració de 40.0? Sol.: 61

   (d) Si el primer mesurament va ser a les 00:00 i s'obté un nou cada 5 minuts, a quina hora es va arribar a la concentració màxima? Sol.: A les 11:50

   (e) Quant val la concentració mitjana de tot el dia? (Ajuda: sumeu totes les dades i dividiu-les entre el nombre d'observacions) Sol.: 24.07229

   (f) Visualitzeu les dades i interpreteu si el factor temps afecta el valor d'aquestes? Sol.: Sí.

   (g) Quines van ser les 10 medicions més baixes del dia? Sol.: 0.93, 1.07, 1.77, 2.03, 2.58, 2.73, 2.75, 2.88, 2.88 i 2.91

10. Emmagatzemeu les dades contingudes a l'arxiu `s2-defecte.txt` a una variable. Si aquestes representen el tipus de defecte observat en les peces defectuoses de la producció d'un dia:

(a) Quantes peces defectuoses s'han analitzat? Sol.: 87

(b) Quins són el diferents tipus de defecte trobats? Sol.: Fractura, porus i rebava.

(c) Quantes peces estan afectades de cada tipus de defecte? Sol.: 4 de fractura, 50 de porus i 33 de rebava.

(d) Construïu un vector numèric que codifique cada 'fractura' com '0', cada 'porus' com '1' i cada 'rebava' com '2'.

(e) Fes un gràfic del tipus de defecte usant el vector codificat (l'original, de tipus `character` no funciona). S'observa al gràfic alguna tendència del tipus de defecte? Sol.: No

1.7 Pràctica R: 3. Mostres de dades multivariants (la classe `data.frame`)

Objectius

Les dades que són analitzades en les situacions reals solen ser més complexes que un únic valor numèric, o una única etiqueta. Hi ha dades de tipus "les coordenades geogràfiques d'esdeveniments", o "la composició centesimal de minerals en roques", o "el tipus d'objecte i les seues dimensions en l'anàlisi d'imatges", etc.

La primera complexitat en les dades consisteix en agrupar dades més senzilles: per exemple, si en un experiment s'analitza la llargària d'una peça i, al mateix temps, la temperatura ambient existent en el moment de mesurar-la, aleshores una dada podria ser (100.02 mm, 25.7°C), mentre que altra seria (99.98 mm, 21.9°C). Així doncs, les dades són, de fet, parelles de dades.

Les dades multivariants es poden arranjar escrivint en cada fila, una dada multivariant (és a dir, un vector de dades mixtes). A continuació, les següents dades s'escriurien en les línies successives, formant una matriu, a la qual podem afegir capçaleres per a aclarir la natura de les variables, com es mostra a la Taula 1.2.

Taula 1.2: Taula de dades multivariants. Cada columna té la puntuació de tota la mostra respecte d'una variable senzilla, i cada línia té la puntuació d'un individu de la mostra en les distintes variables.

id	variable.1	variable.2	...	variable.k
1	valor.1.variable.1	valor.1.variable.2	...	valor.1.variable.k
2	valor.2.variable.1	valor.2.variable.2	...	valor.2.variable.k
⋮	⋮	⋮	⋱	⋮
n	valor.n.variable.1	valor.n.variable.2	...	valor.n.variable.k

La classe de variable implementada a R per a gestionar aquest tipus de dades és el full de dades (`data.frame`). Per tant, l'objectiu d'aquest capítol

és el domini en la manipulació i en l'exploració de les dades contingudes en un full de dades.

Declaració de `data.frame`

Un full de dades (*data frame* en anglès) és una classe especialitzada dins la classe `list` (de la qual no donarem detalls), una llista amb l'especialitat de tenir vectors de la mateixa llargària en totes les components.

La funció `data.frame()` serveix per a declarar un objecte de la classe full de dades, i té com a arguments una sèrie de vectors, tots de la mateixa dimensió. Per exemple:

```
f1 <- data.frame( nom=c("Pere", "Joan"), edat=c(33, 22, 10),
                  professor=c(T, T, F) )
# dóna error perquè no són vectors igual de llargs !
f1 <- data.frame( nom=c("Pere", "Joan", "Eva"),
                  edat=c(33, 22, 10), professor=c(T, T, F) )
f1
f2 <- data.frame( nom=c("Pere", "Joan", "Eva", "Maria"),
                  edat=c(33, 22, 10, 6),
                  professor=c(T, T, F, F) )
f2
```

Les etiquetes que es donen a cada vector seran les capçaleres (*header*) informatives del significat de les dades de cada columna.

R té implementades algunes bases de dades (que podeu indagar amb la funció `data()`, sense arguments). La funció `data()` amb el nom d'una de les bases de dades com a argument, fa que aqueix nom siga una variable que conté el *data.frame* d'aqueixes dades. Provem amb:

```
data()      # obri una finestra amb els noms de les dades a R
data(iris)  # ara 'iris' es una variable amb dades de flors
iris        # mostra les dades per pantalla
```

Principals funcions que actuen sobre `data.frame`

- `dim()`: funció que admet com a argument un data.frame.
  Torna un vector de dues components: el nombre de files i el de columnes de l'argument.
  Exemple: `dim(f2)`, `dim(iris)`

- `length()`: funció que admet com a argument un data.frame.
  Torna només el nombre de columnes de l'argument.
  Exemple: `length(f2)`, `length(iris)`

- `[ ]`: funció que admet com a argument un data.frame i un o dos vectors: (1) el data.frame és la variable de la qual es va a seleccionar una part i (2) si es passen dos vectors, només es seleccionen els valors que ocupen les files i columnes indicades pels vectors respectius. Si es passa només

un vector, es seleccionen les columnes completes que indiquen els valors del vector passat.
Torna un data.frame que és el "subdataframe" del primer argument indicat per les components del segon argument (encara que siga una columna simple).
Exemples:
```
f3 <- iris[1]   # f3 te 1 col: la 1a col de 'iris'
f3 <- iris['Sepal.Length']              # lo mateix
f4 <- f2[c(3,1)]         # f4 te 2 col's: la 3a i 1a de f2
f5 <- iris[c(3,1), c(1,2)] # f5 te 2 fil's i 2 col's:
                         # la 1a i 2a fila de 'iris'
                         # i la 3a i 1a fila
```
També es pot fer servir per a modificar valors del data.frame.
Exemple: `f5[1, 1] <- 5.0 #`

- `[[ ]]`: funció que admet com a argument un data.frame i un valor o un parell de valors: (1) el data.frame és la variable de la qual es va a seleccionar una part i (2) si es passa un valor, aquest indicarà la columna que es selecciona, si es passa un parell de valors, aquests indicaran l'element (fila, columna) que es selecciona.
  Torna un vector (columna) o un únic valor del data.frame argument. Ja no és un data.frame, sinó un vector o un únic valor.
  Exemples:
  ```
  v6 <- iris[[1]]  # v6 es un vector: la 1a col de 'iris'
  v7 <- iris[['Sepal.Length']]               # lo mateix
  v8 <- iris[[5,'Sepal.Length']]  # la 5a dada de la 1a col
  ```
  També es pot fer servir per a modificar valors del data.frame.
  Exemple: `f5[[1]][1] <- 7.0`

- `subset()`: funció que admet, com a arguments, un data.frame, una condició i un vector de columnes.
  Torna un data.frame, la part del data.frame argument corresponent a les línies de dades que verifiquen la condició, i només en les columnes marcades en l'últim argument (si s'especifica).
  Exemple:
  ```
  f8 <- subset(x=iris, subset= (Species=='versicolor'),
               select=c('Sepal.Length', 'Species'))
  ```

- $: és un àlies de la funció `[[ ]]`, quan només es selecciona una columna del data.frame. L'equivalència és:
  `dataframe[['capçalera']] == dataframe$capçalera`
  Exemple:
  ```
  iris[[1]] # una forma de referir el vector 1a col
  iris[['Sepal.Length']]              # el mateix vector
  iris$Sepal.Length                    # el mateix vector
  ```

Exploració de les dades d'un `data.frame`

Explorar les dades d'un full de dades és tant senzill com explorar les dades de vectors, ja que cada columna del full de dades és un vector.

Per exemple, usant les dades de la variable `iris` podem esbrinar:

- Quantes flors s'han analitzat?

```
dim(iris)[1]     # o length(iris[[1]])
```

- Quantes variables s'estudien sobre aquestes flors?

```
dim(iris)[2]     # o length(iris)
```

- Quines espècies de flors s'han analitzat?

```
unique(iris$Species)  #  o unique(iris[[5]])
```

- Quantes flors de l'estudi són 'setosa'?

```
length( iris$Species[ iris$Species=='setosa' ] )
```

- Quantes flors tenen una llargària de sèpal inferior a 6.0 i amplària de sèpal superior a 2.5?

```
length( iris$Species[ (iris$Sepal.Length < 6.0) &
                              iris$Sepal.Width > 2.5) ] )
```

- Quina és la mitjana de les llargàries de pètals per a les flors de l'espècie 'versicolor'? (ajuda: sumar i dividir)

```
suma <- sum( iris$Petal.Length[ iris$Species==
               'versicolor' ] )
n <- length( iris$Petal.Length[ iris$Species==
               'versicolor' ] )
mitjana <- suma/n
mitjana
```

- Com fariem un full de dades amb les dades de les flors 'setosa' respecte a les dimensions de pètals només?

```
submostra <- subset( x=iris, subset=(Species=='setosa'),
                select=c('Petal.Length', 'Petal.Width') )
```

Lectura-escriptura de dades en arxius

La forma més usual per a obtenir un full de dades no és teclejar totes les dades, sinó llegir-les d'un arxiu. La funció més habitual (amb els arguments més usats) per a fer-ho és:

`read.table(file, header, dec,...)`

on els arguments més usats tenen la interpretació:

- `file`: nom de l'arxiu, entre cometes i amb extensió.

- `header`: té capçaleres? Si la primera fila de l'arxiu té els noms de les variables (i a partir de la segona línia les dades), aleshores té capçaleres (i cal advertir-ho amb `header=TRUE`). R ho sabrà i les usarà com etiquetes. Si l'arxiu de dades no té capçaleres, R començarà a llegir les dades des de la primera línia, i establirà com a capçaleres els noms `V1`, `V2`, etc. El valor per defecte és `FALSE`.

- `dec`: signe que indica el nombre decimal. Per defecte és el punt (`.`). Si en l'arxiu hi ha comes decimals (`,`), cal indicar-ho, o R pensarà que les dades són de tipus `"character"`.

Practiqueu amb els arxius disponibles a la web [13] `s3-dades2v-dataframe-dades-1.txt` i `s3-dades2v-dataframe-dades-2.txt`. Editeu-los primer per comprovar el tema de les capçaleres (si cal imposar `header=TRUE` o no) i els decimals (si cal imposar `dec=","` o no).

```
f6 <- read.table(file="s3-dades2v-dataframe-dades-1.txt", ...
f6
f7 <- read.table(file="s3-dades2v-dataframe-dades-2.txt", ...
f7
```

Per a comprovar que les variables `f6` i `f7` contenen les dades de manera correcta, podeu accedir a cada vector columna usant l'operador `$` seguit del nom de la columna.

L'operació contrària a llegir un full de dades d'un arxiu és precisament crear un arxiu de text amb el contingut d'una variable de tipus full de dades. La funció:

```
write.table(x, file = "")
```

on els arguments més usats tenen la interpretació:

- `x`: variable que té l'objecte `data.frame` que voleu guardar a l'arxiu.

- `file`: nom de l'arxiu, entre cometes i amb extensió.

Escriviu el contingut de la variable `f3` en un arxiu anomenat `flors.txt`.

La funció `write.table()` és la funció recíproca de `read.table()`. Recordeu que si no s'indica un `path` diferent, l'arxiu s'escriurà al directori de treball.

Representació gràfica de dades d'un data.frame

La representació gràfica més conveninent de les dades d'un data.frame depèn del tipus de variable que formen les seues columnes.

La funció `plot()`

La funció `plot()` s'encarrega de dibuixar gràfics en dues dimensions. La sintaxi usual és:

```
plot(x, y, type, pch, col, main, sub, xlab, ylab, ...)
```

on els arguments més usats tenen la interpretació següent:

- `x`: vector amb les abscisses dels punts que s'han de dibuixar.
- `y`: vector amb les ordenades dels punts que s'han de dibuixar.
- `type`: forma d'unir els punts. Per defecte els dibuixa aïllats, però si fem `type="l"` una línia anirà unint els punts successius.
- `pch`: aparença del punt. Per defecte és un punt, però podeu triar un número del 0 al 26 o qualsevol caràcter entre cometes.
- `col`: color del punt.
- `main`: títol per al gràfic.
- `sub`: subtítol per al gràfic.
- `xlab`: etiqueta de l'eix d'abscisses.
- `ylab`: etiqueta de l'eix d'ordenades.

La funció `plot()` està implementada de manera especial als data.frame. Quan s'aplica la funció `plot()` a un data.frame amb dues columnes, la primera fa de `x`, la segona de `y` i les capçaleres de les variables s'usen com etiquetes dels eixos. Quan s'aplica a un data.frame amb més de dues columnes, el gràfic que es genera conté els gràfics bidimensionals de totes les combinacions de parelles de columnes.

```
plot(iris[c(1,2)])
plot(iris)
```

Quan les dades d'una columna no són numèriques, R les codifica a numèriques (des de 0 endavant) seguint l'ordre alfabètic, la qual cosa dificulta una miqueta la lectura del gràfic.

Aplicació a la representació gràfica de funcions d'una variable

La representació gràfica d'una funció $y = f(x)$ en un rang de valors de $x \in [a, b]$ es pot fer mitjançant la funció `plot()` de R.

Per a aconseguir-ho, primerament s'ha de definir el vector de valors de x, amb la major densitat possible. Després es calcula el valor de la funció a tots els valors de l'interval i, finalment, es dibuixa el gràfic bidimensional. Per exemple, per a dibuixar la paràbola $y = f(x) = x^2$ a l'interval $x \in [-1, 1]$ es podria fer:

```
x <- seq(from=-1, to=1, length=10)
fx <- x^2
plot(x, fx, main="Paràbola", sub="La típica paràbola",
     xlab="Ací van les abscisses", ylab="Ací les ordenades")
plot(x, fx, type="l", main="Paràbola",
     sub="La típica paràbola", xlab="Ací van les abscisses",
     ylab="Ací les ordenades")
```

Per a guardar el gràfic a arxiu, es pot fer polsant el botó secundari del ratolí, i triant el tipus d'arxiu gràfic preferit.

Afegir més gràfics a un gràfic existent

La funció `plot()` tanca la finestra gràfica actual, si és oberta, i n'inicia una de nova. Si es vol afegir més informació gràfica a una finestra oberta, s'ha d'usar la funció `points()`, la qual afegeix punts:

```
points(x, y, type , pch,  col, ...)
```

Els arguments són els mateixos que per a `plot()`.

Exercicis d'ensinistrament

Els següents exercicis són d'autoaprenentatge, per a ser realitzats individualment i amb l'assistència del professor si ho considereu necessari. No s'avaluen, però sí que podeu autoavaluar-vos, és a dir, podeu comprovar la correcció de l'exercici al mateix R: (1) escrivint el nom de la variable resposta si es demana la definició d'un objecte), (2) mirant la solució indicada si es demana el resultat d'una operació, o (3) testejant una funció si l'exercici demana la definició d'una d'aquestes.

1. Emmagatzemeu el full de dades contingut a l'arxiu `s3-dades2v-dataframe-dades-1.txt` en la variable `f6` (recordeu p. 26). Si la columna "hores" indica el nombre d'hores setmanals dedicades a l'assignatura, i la columna "notes" indica la nota final:

   (a) De quants alumnes s'ha arreplegat la informació sobre hores d'estudi i nota? Sol.: 133

   (b) Quina és la nota màxima obtinguda pels alumnes de la mostra? Sol.: 10

   (c) Quants alumnes han estudiat 3 hores o més a la setmana? Sol.: 62

   (d) Quant val la nota mitjana dels alumnes que han estudiat més de 3 hores a la setmana? I la dels que han estudiat 3 o menys hores per setmana? Sol.: 7.648387 i 4.7765

   (e) Quina és la mitjana d'hores d'estudi dels alumnes que han aprovat l'examen? I la dels que han suspés? Sol.: 3.152174 i 0.7804878

2. Emmagatzemeu en la variable `f7` el full de dades contingut a l'arxiu `s3-dades2v-dataframe-dades-2.txt` (recordeu p. 26). Si les dades corresponen a un estudi sobre la velocitat d'Internet, on s'han pres dades sobre descàrregues d'arxius anotant la seua mida (en MB), el "temps" de descàrrega (en segons) i el "proveïdor" d'internet

   (a) Quantes descàrregues s'han analitzat en l'estudi? Sol.: 145

   (b) De quina mida és l'arxiu més menut? Sol.: 0.9183045 MB

   (c) Quin ha sigut el temps de descàrrega màxim? Sol.: 9.57556 segons

   (d) Quants arxius s'han analitzat de cada proveïdor? Sol.: 59 de l'ISP1, 49 de l'ISP2 i 37 de l'ISP3

   (e) Si definim la velocitat de descàrrega com la divisió entre mida i temps, calculeu la velocitat màxima que ha oferit cada proveïdor. Sol.: 1.990711 per l'ISP1, 1.886546 per l'ISP2 i 2.124815 per l'ISP3 (tot mesurat en MB/seg)

3. Dibuixeu el gràfic on apareguen les dades de l'exercici 1: que en les abscisses aparega el temps d'estudi i en les ordenades la nota final. Què ens ensenya el gràfic?

4. Dibuixeu el gràfic on apareguen les dades de l'exercici 2: que en les abscisses aparega la mida de cada arxiu i en les ordenades el temps de descàrrega. S'intueix alguna relació entre les variables?

5. Dibuixeu el gràfic de la funció $y = \mathrm{e}^{-|x|}\cos(10x)$ a l'interval $x \in [-2, 2]$. Fica el títol "Esmorteïment" al gràfic.

6. Afegiu al gràfic anterior el gràfic de la funció $y = \mathrm{e}^{-|x|}$ en el mateix interval, però en color diferent, i amb tret discontinu.

PART II
MOSTRES DE DADES (ESTADÍSTICA DESCRIPTIVA)

Capítol 2

Descripció de mostres de dades qualitatives

2.1 Què són i com es representen

Una **mostra univariant qualitativa** és una llista de dades tipus "paraula" que pertanyen a una sèrie de categories. Són dades corresponents a un concepte de tipus qualitatiu, per contraposició als conceptes que s'expressen amb quantitats o valors d'escales de mesura numerals. En aquest cas, dues dades de la mostra són iguals o diferents, no hi ha més (a excepció de casos com el de l'Exercici 2.2.1, on hi ha un ordre natural entre les dades, que pot expressar-se amb l'adjectiu d'**ordinal**).

És molt fàcil tractar-les: només cal comptar quantes dades hi ha de cada categoria. El resultat es pot maquetar en forma de:

Definició 2.1.1 *La presentació eficient de mostres de dades qualitatives es pot fer en forma numèrica o gràfica:*

- ***Taula de freqüències**: taula on figuren les quantitats i percentatges de cada categoria apareguda en la mostra.*
- ***Diagrama de barres i sectors**: representació on l'abundància de dades de cada categoria es visualitza en la mida d'una barra o d'un sector de cercle que la representa.*

Exemple 2.1.1 *Una enquesta de "Llenguatges de programació favorits" es porta a terme sobre un grup d'alumnes. Els resultats s'anoten en una llista que després se simplifica en forma de taula tal com apareix a la Taula 2.1.*

Taula 2.1: Taula de freqüències de la mostra de dades recollida a l'Exemple 2.1.1

Llenguatge	C++	Python	Delphi	Java	C	Pascal	TOTAL
Freqüència	23	21	17	11	9	6	87
%	26.44	24.14	19.54	12.64	10.34	6.89	100.00

La forma en què es confecciona la taula de freqüències és prou intuïtiva, i no necessita explicació. De totes formes, fixem notació per poder treballar de manera més abstracta amb mostres de dades qualitatives:

- x_i: cada **categoria** (o dada diferent) de la variable (suposem que hi ha k categories per la notació de les fórmules). Aleshores tenim les categories $x_1, x_2, \ldots, x_k$.
- n_i: **freqüència absoluta** de la categoria x_i. És el nombre de dades de la mostra que coincideixen amb x_i (o nombre de repeticions de la dada x_i).
- n: **mida de la mostra**. Per tant, $n = \sum_{i=1}^{k} n_i$.
- f_i: **freqüència relativa** de la dada o categoria x_i. És la porció de mostra que representa la dada x_i (similar al percentatge, però sobre un màxim d'1 en lloc de 100). Per tant,

$$f_i = \frac{n_i}{n}, \qquad \% = f_i \times 100, \qquad \sum_{i=1}^{k} f_i = 1.$$

  Encara que la freqüència relativa és més "universal", les taules de freqüències solen mostrar els percentatges, per ser aquestos més populars.

Sobre les taules de freqüències només cal destacar un altre detall, al següent exercici.

Exercici 2.1.1 *Observeu que els llenguatges mostrats en la Taula 2.1 es podien haver ordenat d'altres formes. Per què és convenient aquesta forma?*

Les taules de freqüències són molt precises perquè indiquen quantitats exactes. La representació gràfica de la taula de freqüències és més imprecisa, però més poderosa, perquè transmet la informació més immediatament (vegeu la Figura 2.1).

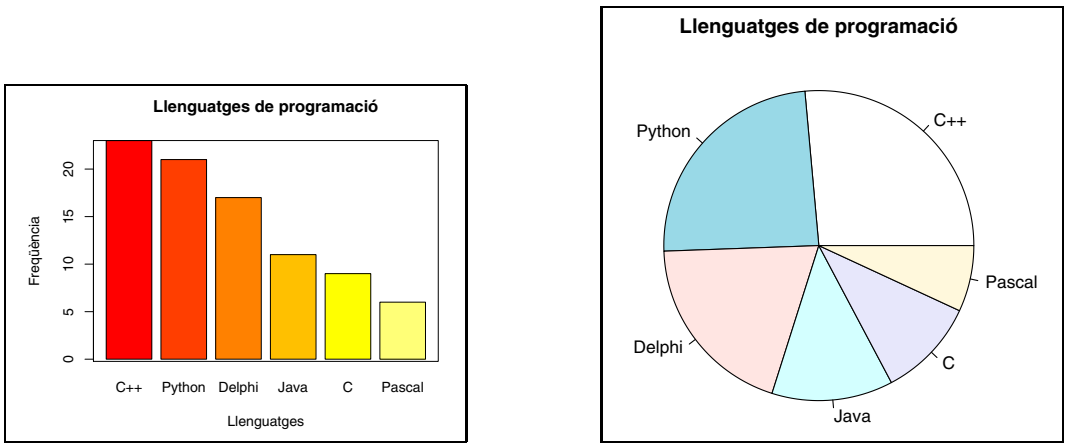


Figura 2.1: Diagrama de barres i de sectors per a la mostra recollida a l'Exemple 2.1.1 i recopilada a la Taula 2.1

La realització dels gràfics també és molt intuïtiva, i no necessita explicació. Actualment els ordinadors realitzen gràfics de molta qualitat.

La manera en què les dades de la mostra estan repartides s'anomena **distribució** (es pot dir distribució de freqüències).

2.2 Més exercicis

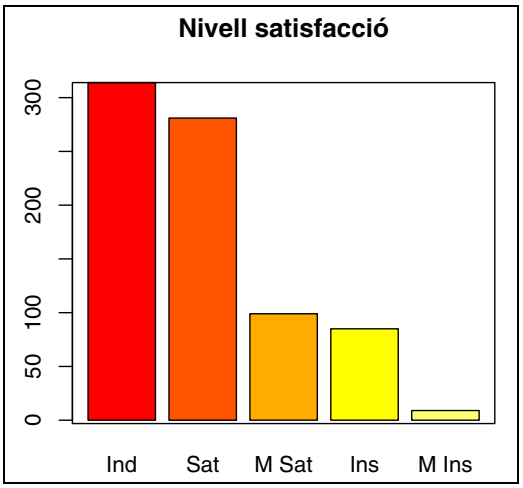
Exercici 2.2.1 *Una empresa de serveis demana opinió als usuaris sobre* ***nivell de satisfació****. La resposta s'ha de triar de les opcions:*

☐ *Molt insatif.* ☐ *Insatisfet* ☐ *Indiferent* ☐ *Satisfet* ☐ *Molt satisf.*

La mostra es resumeix de dues formes diferents, però similars, a les Taules 2.2 i 2.3. Quina de les dues representacions és més informativa o útil? (compara amb l'Exemple 2.1.1).

Taula 2.2: Taula de freqüències i gràfic associat a l'Exercici 2.2.1

Nivell	Freq.	%
Indif.	314	39.85
Sastisf.	281	35.66
Molt sat.	99	12.56
Insat.	85	10.79
Molt insat.	9	1.14
Total	788	100.00



Exercici 2.2.2 *Observeu els següents taula i gràfic (Figura 2.2) corresponents a la mateixa mostra sobre tipus de defecte en la fabricació de peces de plàstic. Són compatibles o hi ha alguna discordància?*

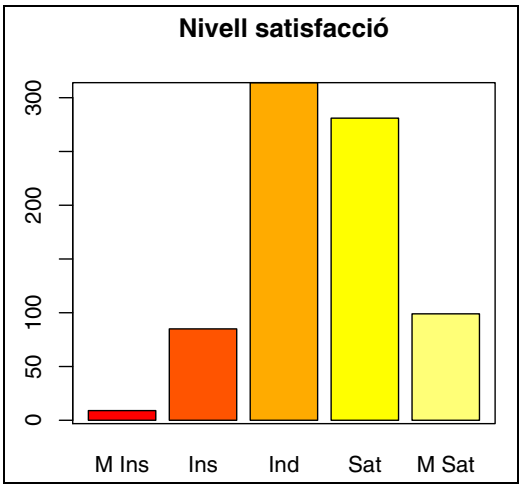
Exercici 2.2.3 *Un professor posa un examen tipus test de* 150 *preguntes amb* 5 *opcions per pregunta. Analitzem les solucions finals obtenint el diagrama de sectors de la Figura 2.3. Escriu una taula de freqüències compatible amb el gràfic. Si fas un altre examen amb aquest professor i trobes una qüestió que no saps, quina estratègia de resposta agafaries?*

2.3 La moda

Definició 2.3.1 (Moda) *Categoria (de la variable) més abundant en la mostra. És, per tant, el valor més representatiu d'una mostra qualitativa.*

Taula 2.3: Una altra taula de freqüències i gràfic associat a l'Exercici 2.2.1

Nivell	Freq.	%
Molt insat.	9	1.14
Insat.	85	10.79
Indif.	314	39.85
Sastisf.	281	35.66
Molt sat.	99	12.56
Total	788	100.00



Tipus defecte	Freq.	%
Porus	99	66.89
Rebava	32	21.62
Fractura	17	11.49
Total	148	100.00

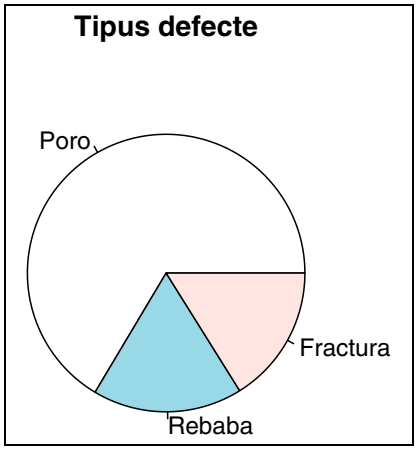
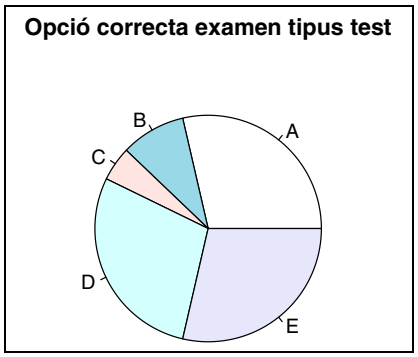


Figura 2.2: Figura de l'Exercici 2.2.2



Total	150	

Figura 2.3: Figura de l'Exercici 2.2.3

Exercici 2.3.1 *Dues (sub)mostres qualitatives X i Y es junten per a formar una sola mostra $Z = X \cup Y$. Explica quina relació hi ha (o pot haver-hi) entre les modes de les submostres i la moda de la mostra total.*

Per exemple: "la moda de la mostra és sempre una de les modes de les submostres?"

(Raoneu si creieu que sempre és així, o useu contraexemples si creieu que de vegades no és veritat).

Capítol 3

Descripció de mostres de dades quantitatives

3.1 Introducció

Una mostra de dades univariants quantitatives és una llista de dades de tipus "nombre" (que expressa una quantitat de comptar o d'una escala de mesures). En aquests casos, dues dades de la mostra es poden identificar com iguals, diferents, major o menor, més allunyades que d'altres, doble o triple que l'altra... Hi ha més joc.

Exemple 3.1.1 *Un usuari d'internet programa el seu ordinador per comprovar la velocitat de baixada del seu proveïdor cada minut. Els resultats (en Mbps) els últims 20 minuts han sigut els de la Taula 3.1.*

Taula 3.1: Dades arreplegades de l'Exemple 3.1.1

1.72	1.77	2.03	1.81	1.82	2.06	1.87	1.61	1.70	1.73
1.98	1.85	1.86	1.82	1.72	2.07	1.87	1.51	1.91	1.73

En els exemples reals la llista pot ser de milers de dades.

3.2 La mostra a primera vista

3.2.1 Interpretació geomètrica de les dades quantitatives

Per una banda, es pot fer la cadena d'identificacions següent:

$$\textbf{dada} = \textbf{nombre} = \textbf{posició} = \textbf{punt} \text{ (recta real)}$$

Per exemple, la dada $x_1 = 1.72$ queda situada a la recta real com es veu a la Figura 3.1.

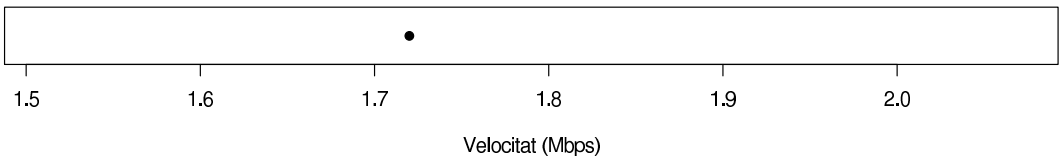


Figura 3.1: Una dada numèrica vista com a posició a la recta real

Les identificacions anteriors ens porten a les següents: **mostra = conjunt de nombres = núvol de punts** (en la recta real).

Per exemple, la mostra sencera de l'Exemple 3.1.1 es veuria com es mostra a la Figura 3.2.

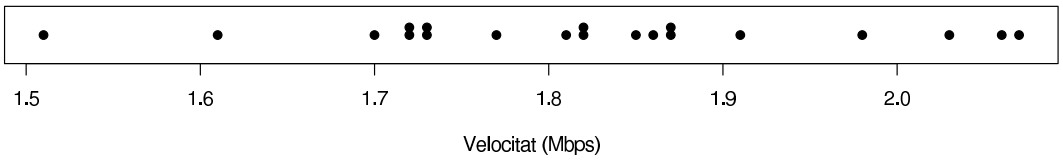


Figura 3.2: La mostra completa de les dades de l'Exemple 3.1.1

La informació d'una gran mostra és difícil d'assimilar observant les dades una per una. Per això és important organitzar-les:

- Numèricament, amb les dades com nombres que són **(Taula de freqüències)**.
- Gràficament, amb les dades com posicions dins la recta real **(Diagrama de punts o Histograma)**.

3.2.2 Taula de freqüències

És un concepte molt intuïtiu. Quan es tenen moltes dades i no hi ha prou repeticions entre elles, és necessari tallar la recta dels nombres reals en intervals, i que els intervals facen de categories. Comprovem-ho amb un exemple.

Exemple 3.2.1 *Per fer un estudi més complet, l'usuari recull* 500 *comprovacions de la velocitat de baixada de l'ADSL. En observar-les, comprova que hi ha massa valors diferents per tractar de comptar les repeticions de cadascuna. Mirant els valors extrems, veu que totes les dades estan entre* 1.4 *i* 2.3 *Mbps. Per tant decideix tallar l'interval* $[1.4, 2.3]$ *en trossos iguals i comptar quantes dades hi ha a cada interval. Amb això obté les dues primeres columnes de la Taula 3.2*

Per crear una eina més informativa, a la taula de freqüències s'afegeixen més columnes:

- x_i: cada **categoria** (interval) de la variable (suposem que hi ha k categories per la notació de les fórmules). Aleshores tenim les categories $x_1, x_2, \ldots, x_k$.

Taula 3.2: Taula de freqüències de la mostra de 500 dades de velocitats de descàrrega de l'Exemple 3.2.1

Veloc. (x_i)	Fr.abs. (n_i)	Fr.rel. (f_i)	Fr.abs.ac. (N_i)	Fr.rel.ac. (F_i)
[1.4–1.5]	11	0.022	11	0.022
(1.5–1.6]	30	0.060	41	0.082
(1.6–1.7]	75	0.150	116	0.232
(1.7–1.8]	136	0.272	252	0.504
(1.8–1.9]	125	0.250	377	0.754
(1.9–2.0]	78	0.156	455	0.910
(2.0–2.1]	33	0.066	488	0.976
(2.1–2.2]	11	0.022	499	0.998
(2.2–2.3]	1	0.002	500	1.000

- n_i: **freqüència absoluta** de la categoria x_i.

  És el nombre de dades de la mostra que cauen dins la categoria x_i.

- n: **mida de la mostra**. Per tant: $\sum_{i=1}^{k} n_i = n$.

- f_i: freqüència relativa de l'interval x_i. És la proporció de dades de la mostra dins la categoria x_i (similar al percentatge, però sobre un màxim d'1 en lloc de 100). Per tant:

$$f_i = \frac{n_i}{n}, \qquad \% = f_i \times 100, \qquad \sum_{i=1}^{k} f_i = 1.$$

- N_i (i F_i): **freqüències absolutes (i relatives) acumulades** fins la categoria x_i.

  Compten el nombre (i proporció) de dades de la mostra dins la categoria x_i o qualsevol anterior. Per tant:

$$N_i = \sum_{j=1}^{i} n_j, \qquad F_i = \sum_{j=1}^{i} f_j,$$

Les preguntes bàsiques que ajuden a contestar les taules de freqüències són les del tipus: "Quantes dades de la mostra...

- ... estan dins de l'interval de tolerància...?"

- ... són superiors a...?"

- ... són inferiors o iguals a...?"

Si els intervals no coinciden amb els valors de les preguntes, només es poden donar respostes aproximades. Amb la gestió de dades mitjançant programes

informàtics, cadascuna de les preguntes anteriors té una resposta immediata (amb un clic o darrere d'una línia de codi de programació).

Hi ha prou criteris sobre quants intervals han de formar la taula de freqüències (normalment, en funció de la mida de la mostra), però no els presentem ací.

La manera en què les dades de la mostra estan repartides s'anomena **distribució** (es pot dir distribució de freqüències), i es pot visualitzar bé a la taula de freqüències, bé als gràfics corresponents.

3.2.3 Gràfics

1. Diagrama de punts: s'utilitza quan la mostra té poques dades i es pot visualitzar la distribució del núvol de punts (vegeu la Figura 3.2).

2. Histograma: quan la mostra té una quantitat de dades que fa que el diagrama de punts siga una línia quasi contínua, és necessari elevar columnes que indiquen quantes dades hi ha a cada interval. És un gràfic a l'estil del diagrama de barres format a partir de la taula de freqüències (vegeu la Figura 3.3).

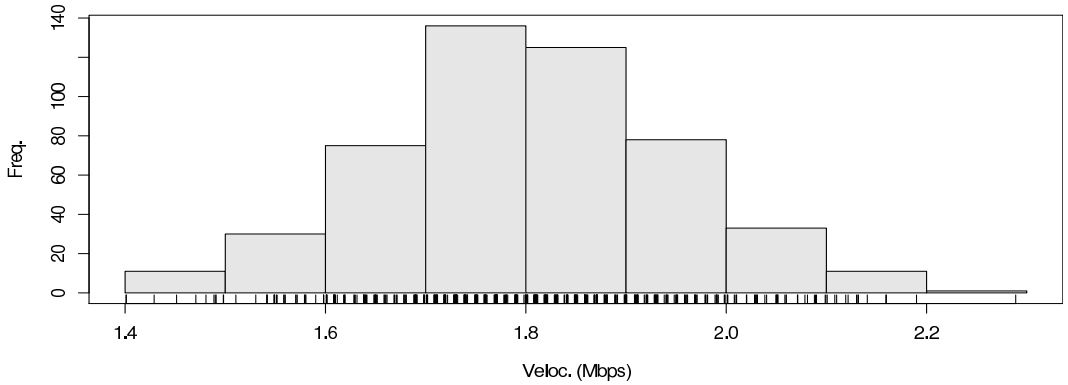


Figura 3.3: Histograma corresponent a la mostra de l'Exemple 3.2.1

3.3 Resum d'una mostra usant estadístics

3.3.1 Estadístics

Definició 3.3.1 (Estadístic) *Funció que calcula un valor a partir de les dades de la mostra.*

Usant la identificació següent,

$$\textbf{mostra} = \textbf{núvol de punts},$$

descriure la mostra significa descriure el núvol de punts. Aleshores, es pot descriure, *grosso modo*, donant:

- Una posició central representativa del núvol de punts: posició al voltant de la qual s'apinyen les dades (**estadístics de posició central**).

- Un nivell de dispersió de les dades: indicador de com d'allunyades entre si estan les dades (**estadístics de dispersió**).

Hi ha d'altres estadístics més elaborats, com els de forma, de concentració (de riquesa, quan les dades són economiques), etc., en els quals no entrem.

3.3.2 Estadístics de posició

Exemple 3.3.1 *Tres proveïdors d'ADSL (*1*,* 2 *i* 3*) que anuncien la mateixa velocitat són comprovats amb 20 proves (3 mostres amb 20 dades per mostra). El resultat gràfic de la comprovació es mostra a la Figura 3.4.*

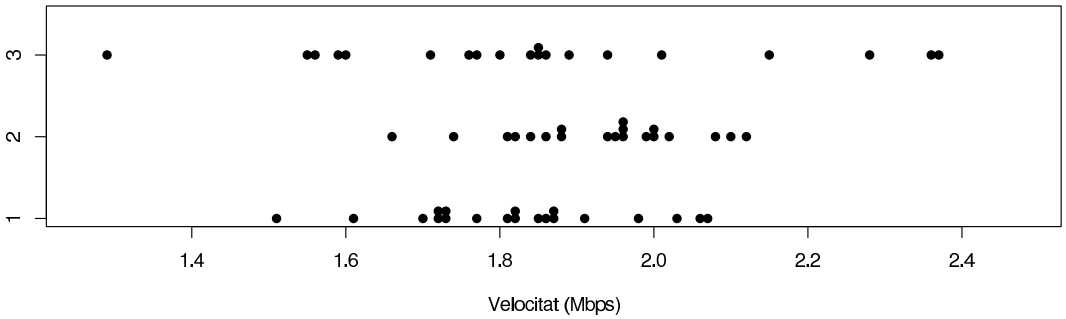


Figura 3.4: Diagrames de punts comparatius entre 3 mostres de dades sobre velocitats d'ADSL de l'Exemple 3.3.1

A igualtat de preus, quin proveïdor et sembla millor? Per què?

Definició 3.3.2 (Estadístics de posició central i exemples) *Valors que indiquen una posició* ***representativa*** *de la mostra completa de dades. Hi ha moltes filosofies per definir aquests valors. Dues de les més populars són:*

- *Les dades* extremes *són les* menys representatives *de la mostra, alesho-res les anem retirant progressivament fins que ens quedem amb un valor central, que serà, doncs, el més representatiu de tots (es denota per* $\widetilde{x}$ *i s'anomena* ***mediana****).*

- *Les dades de la mostra donen lloc a un* total *(vendes, notes...). Aquell valor que compensa totes les dades de la mostra, donant* el mateix total*, serà el valor que millor represente les diverses dades de la mostra. S'a-nomena* ***mitjana****. N'hi ha de diversos tipus segons es calcule el* total*. La més usada és l'****aritmètica*** *que es denota per* $\overline{x}$*, perquè és molt fre-qüent que el total es calcule* sumant*, però n'hi ha d'altres (la* geomètrica*, l'*harmònica*...).*

Si la mostra de dades és $x_1, x_2, \ldots, x_n$, i la seua ordenació (creixent) es denota per $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, aleshores, les definicions rigoroses són:

$$\widetilde{x} = \begin{cases} x_{(\frac{n+1}{2})}, & \text{si } n \text{ senar} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right), & \text{si } n \text{ parell} \end{cases} \qquad \overline{x} = \frac{\sum_i x_i}{n}$$

Exemple 3.3.2 *Un alumne trau, en 10 proves, les puntuacions de la Taula 3.3. Té un nivell global suficient per passar? Segons quin criteri?*

- *La mitjana és $\overline{x} =$ i la mediana $\widetilde{x} =$*
- *Escriu el cas de quatre alumnes, un que aprove amb els dos criteris, un altre que suspenga, i els altres que depenguen del criteri per aprovar... Quin seria el criteri general més just segons la teua opinió?*

Taula 3.3: Dades (notes) de l'alumne de l'Exemple 3.3.2

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
7.6	7.9	3.4	5.8	0.8	8.9	7.3	8.9	5.0	1.9

Definició 3.3.3 (Estadístics de posició relativa i exemples) *Valors referits a la posició que ocupen dins de la mostra completa de dades.*

- ***Mínim**: és el menor valor de la mostra i es denota per $x_{mín}$.*
- ***Màxim**: és el major valor de la mostra i es denota per $x_{màx}$.*
- ***Quantil d'ordre** p (on $p \in [0,1]$): és el valor que deixa per davall, almenys una porció de la mostra igual a p, i per damunt, almenys, una porció de la mostra igual a $1-p$. Es denota per x_p.*

Com exemple, el valor $x_{0.40}$ (quantil d'ordre 0.40) és aquell valor tal que, aproximadament, el 40% de les dades de la mostra són valors inferiors o iguals a ell mentre que la resta (el 60% de la mostra) són valors superiors o iguals a ell.

S'anomenen també percentils, decils o quartils (si divideixen la mostra en 100, 10 o 4 parts). Per exemple, percentil 75 ($P_{75} = x_{0.75}$), tercer decil ($D_3 = x_{0.30}$), quartil inferior ($Q_1 = x_{0.25}$), quartil superior ($Q_3 = x_{0.75}$), etc.

Per exemple, si has obtingut un 8.7 en un examen estaràs molt content, però si et diuen que el teu 8.7 només és el percentil 5, això significa que el 95% dels examinats ha tingut la teua nota o més. Aleshores ja no estaràs tant content perquè el teu nivell no és alt respecte al grup. Per exemple, a la guia d'usuari del sistema de crèdits ECTS de 2005 (veure `http://ec.europa.eu/education/programmes/socrates/ects/doc/guide_en.pdf`), es proposava la següent forma de qualificar als alumnes (*ECTS grading scale*): entre la mostra d'alumnes no suspensos, les notes del professor s'assignen com: A (10% superior), B (25% següent), C (30% següent), D (25% següent), E (10% següent). Així, la nota 8.7 podria ser qualsevol qualificació, depenent de les notes dels altres.

Exercici 3.3.1 *Es passa un test a dos grups d'alumnes (variable NOTA) i ens faciliten la Taula 3.4.*

- *On estan les dades de les dues mostres?*
- *Què indica cadascun dels nombres que apareixen?*
- *Quin grup és més fort en global? Per què?*
- *Completeu amb imaginació uns estadístics per a un grup virtual C que siga molt bo en general, però amb la quarta part de la classe molt roïna (vegeu la Taula 3.5).*

Taula 3.4: Taula facilitada per l'Exercici 3.3.1

Grup	n	$x_{\text{mín}}$	$x_{0.25}$	$\widetilde{x}$	$x_{0.75}$	$x_{\text{màx}}$	$\overline{x}$
A	121	0.0	1.7	5.3	6.1	8.3	5.7
B	89	0.3	0.9	5.3	7.2	8.2	5.7

Taula 3.5: Taula per completar de l'Exercici 3.3.1

Grup	n	$x_{\text{mín}}$	$x_{0.25}$	$\widetilde{x}$	$x_{0.75}$	$x_{\text{màx}}$	$\overline{x}$
C	57						

3.3.3 Estadístics de dispersió

Exemple 3.3.3 *Es continuen analitzant els tres proveïdors d'ADSL de l'Exemple 3.3.1 (vegeu la Figura 3.4).*

Quin proveïdor et sembla més regular? Per què?

Si hagueres de pronosticar una nova dada per cada proveïdor, amb quin tindries més confiança? I amb quin menys?

Definició 3.3.4 (Estadístics de dispersió i exemples) *Valors que indiquen el nivell de variabilitat de les dades d'una mostra. A major valor, major disparitat de dades en la mostra. Amb aquesta filosofia es defineixen:*

- ***Recorregut**: amplària abastada per les dades. Es denota i calcula com $Re = x_{màx} - x_{mín}$.*
- ***Recorregut interquartíl·lic**: amplària abastada per les dades des del quartil inferior al superior. No considera les dades més extremes perquè poden ser degudes a errors en la presa de dades (per exemple). Es denota i calcula com $RQ = x_{0.75} - x_{0.25}$.*

- ***Variància (i desviació típica) mostral**: interpretant la mitjana com a valor correcte i les dades com a intents d'encertar la mitjana, la variància s'interpreta com una espècie d'error quadràtic mitjà, i la desviació típica és la seua arrel quadrada, per tornar a les unitats originals de les dades. Es denoten i calculen com*

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2}{n-1} \qquad\qquad s = +\sqrt{s^2}$$

*El més utilitzat és la **desviació típica**, encara que tots aporten informació.*

La raó de trobar el denominador $n-1$, en lloc de n, a la definició de variància mostral és una mica tècnica, i es comprendrà enterament al capítol d'inferència, on les dades són una mostra d'una població amb mitjana i variància desconegudes, que s'estimaran amb la mitjana i variància mostrals. Intuïtivament, dividir per n la suma dels errors quadràtics infravaloraria la variància poblacional, ja que la mitjana mostral està "forçosament més prop de les dades de la mostra" que la mitjana real de la població sencera.

A l'hora de comparar els graus de dispersió de mostres molt diferents, els estadístics anteriors són injustos, ja que mostres de dades molt altes resultaran en dispersions altes, i mostres de dades molt baixes (en valor absolut) donaran dispersions baixes, sense entrar en si les mostres són més o menys homogènies. Un exemple ho aclarirà.

Exemple 3.3.4 *S'agafen dues mostres:*

- *Els temps (X) dels 8 corredors d'una carrera de 100 metre llisos.*
- *Els temps (Y) dels 8 corredors d'una carrera de marató.*

Suposem que $Re_X = 2$ s i $Re_Y = 5$ s. Quina carrera et sembla que ha tingut un final més "ajustat" (emocionant)?

Definició 3.3.5 (Coeficient de variació de Pearson) *És un estadístic de dispersió relativa, per a poder comparar dispersions de mostres, sense veure's afectades per les mides dels valors de les dades. Es denota i calcula com*

$$CV = \frac{s}{|\overline{x}|}$$

En dividir per $\overline{x}$, s'elimina el factor "mida de les dades", i es fa possible la comparació de mostres de dades encara que siguen molt diferents.

Exercici 3.3.2 *Les dades sobre el temps d'accès de lectura al disc (en milisegons) de dos discos durs A i B es mostren a la Taula 3.6. Compareu les mostres "a ull" en un primer intent, i després amb els estadístics que figuren (completant els que falten) a la Taula 3.7.*

Quin disc és més ràpid? I quin més fiable? Per què?

Taula 3.6: Dades de l'Exercici 3.3.2

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
A	7.85	8.01	7.82	7.63	8.14	8.08	8.82	8.17	7.05	8.64
B	7.86	8.17	7.45	8.51	7.91	7.67	9.83	7.32	8.86	8.92

Taula 3.7: Estadístics de les mostres de l'Exercici 3.3.2

	$x_{\text{mín}}$	$x_{0.25}$	$\widetilde{x}$	$x_{0.75}$	$x_{\text{màx}}$	$\overline{x}$	Re	RQ	s^2	s	CV
A		7.828		8.163							
B		7.718		8.772							

3.3.4 Propietats dels estadístics

Presentem sense demostració algunes propietats matemàtiques dels estadístics que poden ser útils per a aplicacions pràctiques.

Definició 3.3.6 (Transformació lineal) *Si tenim una mostra de dades x_1, x_2, $x_3, \ldots$, x_n, aleshores una nova mostra de dades y_1, y_2, $y_3, \ldots$, y_n és transformació lineal de l'anterior si hi ha dos nombres coneguts a i b tal que*

$$y_i = a + bx_i \qquad \text{per } i = 1, 2, \ldots, n.$$

Si X denota la variable de la primera mostra i Y la variable de la segona, es pot escriure $Y = a + bX$.

Exemple 3.3.5 *Variables que són transformació lineal d'altres:*

- *IVA = 0.16 * PREU*
- *COST.TELEFÒNIC. = 0.12 + 0.18 * TEMPS*
- *PREU.ACADÈMIA. = MATRÍCULA + PREU.MES * TEMPS*
- *MIDA.DESCÀRREGA. = VELOC.CONNEXIÓ. * TEMPS*

Propietat 3.3.1 *Si Y és transformació lineal de X, és a dir, $Y = a + bX$, aleshores:*

- *Mitjana:* $\overline{y} = a + b\overline{x}$
- *Quantils:* $y_p = \begin{cases} a + bx_p, & b > 0 \\ a + bx_{1-p}, & b < 0 \end{cases}$
- *Recorregut:* $Re_Y = |b| Re_X$
- *Recorregut interquartíl·ic:* $RQ_Y = |b| RQ_X$
- *Variància i desviació típica:* $s_Y^2 = b^2 s_X^2$ *i* $s_Y = |b| s_X$

Definició 3.3.7 (Suma de variables) *Si tenim dues mostres de la mateixa mida, x_1, x_2, $x_3, \ldots$, x_n i y_1, y_2, $y_3, \ldots$, y_n de variables X i Y, podem formar una nova mostra z_1, z_2, $z_3, \ldots$, z_n de variable Z com suma de X i Y si*

$$z_i = x_i + y_i \qquad \text{per } i = 1, 2, \ldots, n.$$

En aquest cas, només hi ha una relació segura: $\overline{z} = \overline{x} + \overline{y}$. Dels altres estadístics no es pot assegurar res.

Exemple 3.3.6 *Variables que són suma d'altres variables*

- *PES.TOTAL = PES.RECIPIENT + PES.CONTINGUT*
- *TEMPS.PROCÉS= TEMPS.SUBPROCÉS.1 + TEMPS.SUBPROCÉS.2*

Definició 3.3.8 (Unió de submostres) *Si tenim dues mostres x_1, $x_2 \ldots$, x_n i y_1, $y_2, \ldots$, y_m de variables X i Y, podem fer la unió Z les variables X i Y com*

$$\underbrace{z_1, \ldots, z_n}_{x_1,\ldots,x_n}, \underbrace{z_{n+1}, \ldots, z_{n+m}}_{y_1,\ldots,y_m}$$

Propietat 3.3.2 *Si la mostra Z és unió de X i Y, aleshores:*

- *Mitjana:* $\overline{z} = \dfrac{\sum_i z_i}{n+m} = \dfrac{n\overline{x} + m\overline{y}}{n+m}$
- *Mínim:* $z_{mín} = \min(x_{mín}, y_{mín})$
- *Màxim:* $z_{màx} = \max(x_{màx}, y_{màx})$

Exemple 3.3.7 *Variables que són unió d'altres variables: un estudi necessita la recopilació de dades, que s'encarreguen a diferents tècnics. Després de recopilar les dades, cada tècnic pot calcular els estadístics de la seua submostra, però al final interessa calcular el de la mostra sencera (que és la unió de les submostres).*

3.4 Avaluant mostres amb nous gràfics

3.4.1 Histograma

Per a un ull entrenat, l'histograma és una bona eina per captar i comparar les posicions centrals i dispersions d'una mostra. A la Figura 3.5 es mostra un exemple de comparació efectiva des dels gràfics, sense necessitat de calcular estadístics.

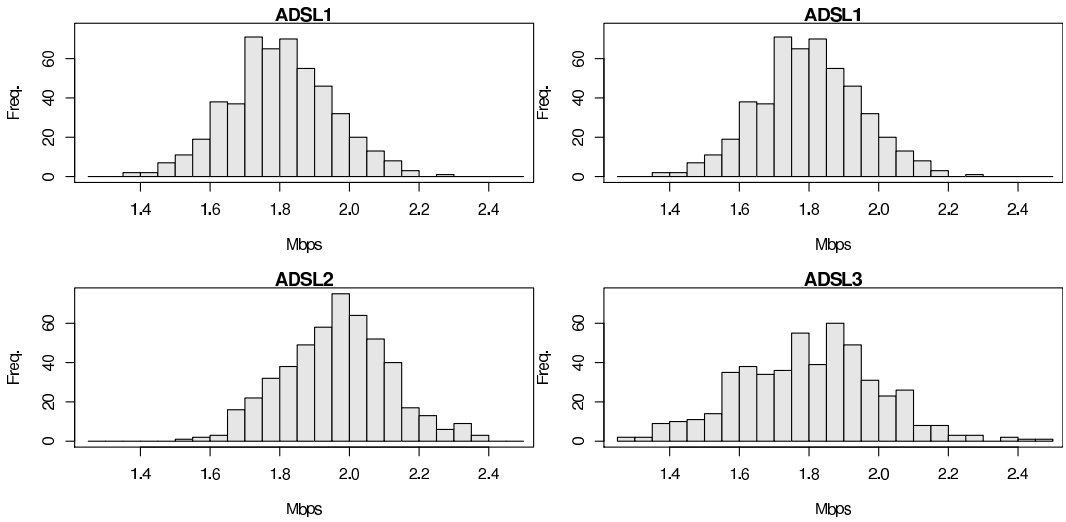


Figura 3.5: Comparació d'histogrames (en vertical només) per avaluar mostres de dades intuïtivament (però eficaçment al mateix temps). Columna esquerra: la posició central és diferent i la dispersió similar. Columna dreta: la posició central és similar i la dispersió diferent

3.4.2 Diagrama de caixa (boxplot)

És un gràfic que transmet d'un cop d'ull els quartils inferior i superior (costats de la caixa), la mediana (marcada a l'interior de la caixa), i un recorregut de valors considerats *normals* (bigots), així com un possible conjunt de valors considerats *anormals* (anomenats *valors atípics* o *outliers*).

De fet, el bigot inferior arriba fins a la menor de les dades que està a una distància del primer quartil inferior a 1.5 voltes el recorregut interquartíl·lic, mentre que el superior ho fa de igual manera respecte del tercer quartil. Les dades que queden fora d'aquests extrems dels bigots es consideren atípics i queden clarament detectats al gràfic.

Per tant, al diagrama de caixa, el grau de dispersió s'intueix a les amplàries de caixes i bigots, i és molt més utilitzat que l'histograma a l'hora de comparar mostres relacionades (com a tractaments, metodologies, condicions de fabricació, etc.).

A la Figura 3.6 es pot veure els diagrames de caixa de les mostres de l'Exemple 3.3.1.

3.4.3 Diagrama de quantils

Transmet d'un cop d'ull el valor de tots els quantils, encara que només de manera aproximada (vegeu la Figura 3.7).

3.5 Exercicis proposats

Exercici 3.5.1 *Una acadèmia que prepara per a fer oposicions fa simulacres d'examen al final del curs. També aconsegueix registrar la nota dels seus alum-*

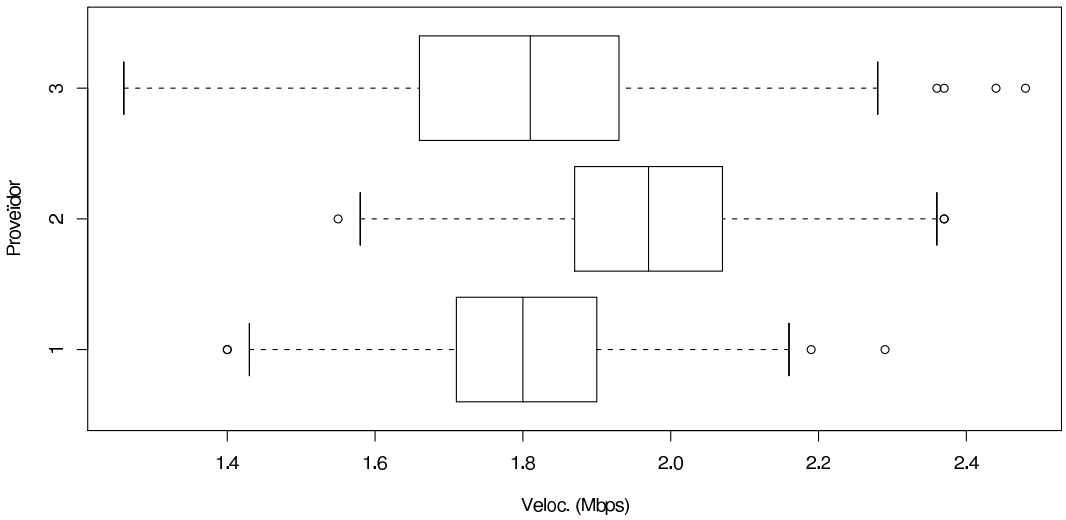


Figura 3.6: Diagrames de caixa de les mostres usades a l'Exemple 3.3.1

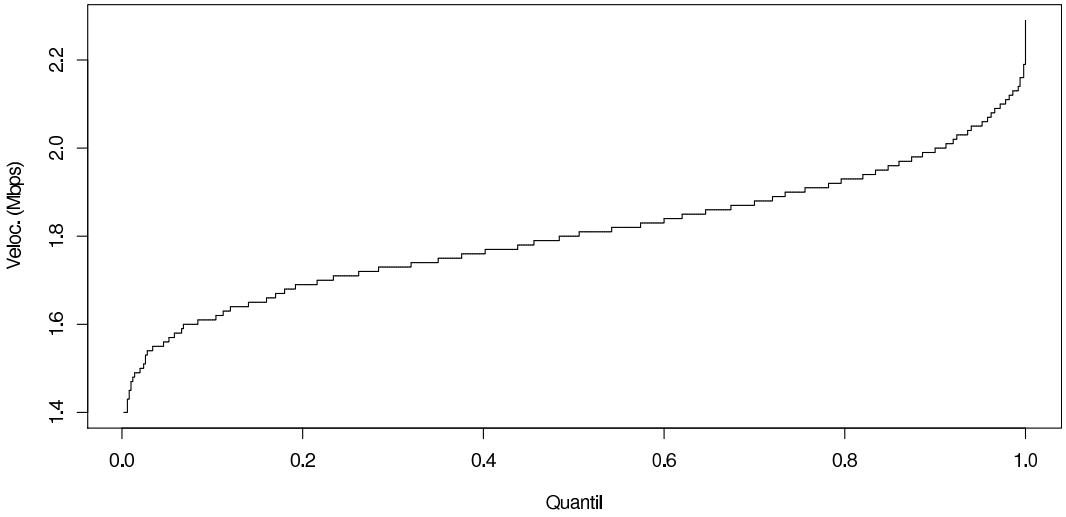


Figura 3.7: Diagrama de quantils de la mostra de dades de l'Exemple 3.2.1

nes quan passen per l'oposició. Així, tenen les dades del curs passat (que van seguir el curs 8 alumnes) a la Taula 3.8.

Taula 3.8: Taula de dades de l'Exercici 3.5.1

Alumne	1	2	3	4	5	6	7	8
Nota acadèmia	8.8	8.8	10.0	8.1	9.6	9.6	8.4	7.7
Nota oposició	9.0	8.5	9.6	8.2	9.6	9.7	8.7	7.9

1. *Quines notes van ser globalment més altes, les del simulacre o les de l'oposició? En què et bases?*

2. *Quines notes van ser globalment més homogènies, les del simulacre o les de l'oposició? En què es baseu?*

Exercici 3.5.2 *Es mesuren amb un aparell les resistències, en Nw, d'un adhesiu en dues condicions ambientals diferents (fred i calor), amb els resultats mostrats a la Taula 3.9.*

Taula 3.9: Taula de l'Exercici 3.5.2

Estad.	n	$\overline{x}$	s_X	$x_{\text{mín}}$	$x_{0.25}$	$\widetilde{x}$	$x_{0.75}$	$x_{\text{màx}}$
Fred	50	1802.14	12.23	1775.03	1792.02	1803.28	1810.72	1829.44
Calor	50	1790.71	11.75	1757.50	1782.14	1791.25	1800.85	1814.99

*Contesteu, **raonant** l'ús dels valors de la taula:*

1. *Generalment, en quines condicions funciona millor l'adhesiu? Fred o calor?*

2. *En quines condicions s'ha observat la millor adherència de la mostra?*

3. *El 75% de les ocasions sota condicions fredes, l'adhesiu té una resistència major als ... Nw?*

4. *La resistència de l'adhesiu és més predible en condicions fredes o calentes?*

Exercici 3.5.3 *La valoració que cada client té sobre un servei d'atenció consisteix en un valor enter de l'1 al 5. Durant els 6 primers mesos de l'any es van arreplegar 326 valoracions, que es van presentar com a positives, perquè es va aconseguir una valoració mitjana de 4.2, amb desviació típica de 0.41. En el segon semestre es van arreplegar 299 valoracions més, que es van presentar com a bones, perquè es va obtenir una valoració mitjana de 3.95, amb desviació típica de 0.98.*

Al final de l'any es vol presentar la valoració global del servei, però s'han perdut les dades originals. Contesteu de manera raonada, si és possible deduir-ho de les dades de què disposeu:

1. *Quina ha sigut la valoració mitjana de tot l'any?*

2. *Quina ha sigut la seua desviació típica?*

3. *Podeu deduir quina ha sigut la valoració màxima?*

4. *Podeu deduir quina ha sigut la valoració mínima?*

Exercici 3.5.4 *La producció diària d'una planta de fabricació (en milers d'unitats) s'arreplega cada dia a les 23:59. Es vol comparar la producció diària de dos anys consecutius, i per això s'han processat estadísticament les dades, donant lloc a la Taula 3.10. Responeu a cada pregunta raonant la resposta i*

Taula 3.10: Estadístics de l'Exercici 3.5.4

	n	$\overline{x}$	$\widetilde{x}$	$x_{\text{mín}}$	$x_{0.25}$	$x_{0.75}$	$x_{\text{màx}}$	s_X
Any 2004	366	181	186	135	151	204	245	32.4
Any 2005	365	188	184	101	176	194	252	23.5

justificant l'ús dels valors de la taula:

1. *La producció diària ha sigut major, en general, durant l'any 2005?*

2. *El pitjor dia en termes de producció, va ser en 2004?*

3. *La producció diària en aquests dos anys ha sigut sempre inferior a les 250000 unitats?*

4. *La producció diària ha sigut més homogènia en 2004?*

5. *Quantes unitats s'han produït en 2005?*

Exercici 3.5.5 *Per comparar dos models d'impressora de color, es registren els valors de* ***temps*** *(X, en segons) que tarda cadascuna en imprimir una sèrie de pàgines. Per raons lògiques s'estudia l'impressió de negre i la de color per separat. Els valors recollits es passen a un programa estadístic que calcula els valors de la Taula 3.11.*

Taula 3.11: Dades de l'Exercici 3.5.5

Marca	Tipus	$\overline{x}$	s	$x_{\text{mín}}$	$x_{0.25}$	$\widetilde{x}$	$x_{0.75}$	$x_{\text{màx}}$
A	B/N	1.296	0.113	1.064	1.225	1.274	1.355	1.514
A	Col	2.741	0.184	2.380	2.665	2.819	2.865	2.938
B	B/N	1.518	0.148	1.237	1.435	1.493	1.599	1.869
B	Col	2.420	0.149	2.229	2.326	2.429	2.469	2.728

Contesteu les preguntes amb justificació breu però suficient:

1. *Quina impressora és més ràpida en cada tipus d'impressió, en general, segons la mostra?*

2. *Quina impressora és més regular en cada tipus d'impressió, en general, segons la mostra?*

3. *La meitat de les pàgines impreses en color per la marca A prenen un temps inferior a* 2.741 *seg. Vertader o fals?*

4. *Quina impressora és més convenient (ràpida) per a una persona que sol imprimir un* 60% *en negre i el* 40% *restant en color?*

Exercici 3.5.6 *Per comparar dos algorismes d'ordenació de bases de dades, es registren els valors de* ***temps*** *(X, en segons) que tarda cadascú en ordenar les dades usant un banc de bases de dades. Per raons lògiques s'estudia el funcionament de l'algorisme per separat en bases de dades menudes i grans. Els valors recollits es passen a un programa estadístic que calcula els valors de la Taula 3.12.*

Taula 3.12: Dades de l'Exercici 3.5.6

Algorisme	Mida	$\overline{x}$	s	$x_{\text{mín}}$	$x_{0.25}$	$\widetilde{x}$	$x_{0.75}$	$x_{\text{màx}}$
A	Menuda	1.518	0.148	1.237	1.435	1.493	1.599	1.869
A	Gran	2.420	0.149	2.229	2.326	2.429	2.469	2.728
B	Menuda	1.296	0.113	1.064	1.225	1.274	1.355	1.514
B	Gran	2.741	0.184	2.380	2.665	2.819	2.865	2.938

Contesteu les preguntes amb justificació breu però suficient:

1. *Quin algorisme és més ràpid en cada tipus de base de dades, en general, segons la mostra?*

2. *Quin algorisme és més regular en cada tipus de base de dades, en general, segons la mostra?*

3. *La meitat de les grans bases de dades ordenades per l'algorisme B prenen un temps inferior a* 2.429 *segons. Vertader o fals?*

4. *Quin algorisme és més convenient (ràpid) per a una empresa que sol treballar amb un* 40% *de bases de dades grans i el* 60% *restant de menudes?*

Exercici 3.5.7 *Es comprova l'eficiència d'un algorisme de càlcul aplicant-lo sobre una bateria de* 583 *problemes. Es registra el temps X, en segons, que tarda l'algorisme a resoldre cada problema.*

Problema	1	2	3	4	$\cdots$
Temps	14.37	15.03	18.61	15.64	$\cdots$

Un ordinador processa les dades recollides donant lloc a la taula:

Estadístics	$\overline{x}$	s	$x_{mín}$	$x_{0.25}$	$\widetilde{x}$	$x_{0.75}$	$x_{màx}$
X	15.55	1.92	10.18	14.33	15.54	16.86	21.98

Contesteu les preguntes amb justificació breu però suficient:

1. *Més de la meitat dels problemes es van resoldre en menys de* 15.0 *segons. Vertader o fals?*

2. *Si l'experimentació va consistir a programar un bucle que prenia cada problema (del primer a l'últim) i l'aplicava l'algorisme, i el temps entre la solució d'un problema i l'aplicació del següent era de* 0.03 *segons, quin ha estat el temps total exacte que ha emprat l'experimentació (des de l'inici del primer problema fins a la solució del problema* 583*)?*

3. *Un altre algorisme, que realitza la mateixa tasca, i al qual se li ha aplicat la mateixa bateria de problemes, dóna un temps mitjà de* 17.3 *segons i una desviació típica de* 0.99 *segons. Si ambdós algorismes han d'aplicar-se a un nou problema, pronostiqueu el temps de resolució de cada algorisme i expliqueu quin dels pronòstics és més fiable.*

Exercici 3.5.8 *Compareu la dispersió de les dues mostres presentades a la Taula 3.13 (és a dir,* ***digues quina és més homogènia****) tenint en compte la diferència de magnitud de les dades de cadascuna.*

Taula 3.13: Dades de l'Exercici 3.5.8

A	5.46	7.75	4.89	6.24	7.90	5.99	5.77	4.79
B	241.03	264.66	221.67	255.64	230.36	233.25	265.83	297.17

Exercici 3.5.9 *Es registra* ***diàriament*** *el nombre total de taulells produïts (en milers d'unitats) per dues línies de producció (A i B) d'una fàbrica, durant l'any 2004. Una volta obtingudes totes les dades, un programa ofereix els estadístics per a cada mostra que es poden veure a la Taula 3.14*

Taula 3.14: Estadístics de l'Exercici 3.5.9

Línia	$\overline{x}$	$x_{\text{mín}}$	$x_{0.25}$	$\widetilde{x}$	$x_{0.75}$	$x_{\text{màx}}$	s
A	37.50	18.14	33.03	37.52	41.98	58.25	6.11
B	38.77	18.26	35.12	38.66	42.33	57.66	5.56

Responeu presentant xifres o raonaments suficients que donent suport a la vostra resposta:

1. *Quina línia ha funcionat 'millor' durant 2004 per la quantitat de taulells produïts diariament?*

2. *Quina línia ha funcionat 'millor' durant 2004 per l'estabilitat en el nombre de taulells produïts diariament?*

3. *Quants taulells es produïren a la línia A el dia de major producció?*

4. *Si es defineix com 'dolent' el dia en el qual la producció és inferior a les 35000 unitats, quina línia de producció va tenir més dies dolents en 2004?*

5. *Quants taulells s'han produït en cada línia durant 2004?*

6. *Si es defineix com 'bo' el dia en el qual la producció és superior a 50000 unitats, quina línia de producció va tenir més dies bons en 2004?*

Exercici 3.5.10 *Completeu la taula de freqüències mostrada a la Taula 3.15.*

Taula 3.15: Taula de freqüències de l'Exercici 3.5.10

x_i	n_i	f_i	N_i	F_i
0				
50	10		34	
100			56	
200		0.224	84	
500			105	
1000				

Exercici 3.5.11 *Els ingressos diaris a una botiga durant els 284 dies d'apertura en 2003 es recullen a la Taula 3.16, on INGRESSOS són els ingressos diaris en euros. Calculeu de manera exacta o aproximada:*

1. *L'ingrés diari mitjà.*

2. *La desviació típica.*

3. *Diuen que almenys la meïtat dels dies es va ingressar per davall dels 60 euros. És açò cert segons les dades?*

Taula 3.16: Dades de l'Exercici 3.5.11

INGRESSOS	0–50	50–100	100–200	200–500	500–1000	1000–5000
N. DIES	24	67	131	46	12	4

Exercici 3.5.12 *S'estudia la variable* ***nombre de membres que componen les vivendes particulars*** *en dos barris de la ciutat de Castelló. Per a tal fi es pren del cens la informació, que es processa estadísticament.*

Les variables BARRI_A i BARRI_B indiquen el nombre de membres per vivenda a cada barri, respectivament. Els resultats eixen resumits a la Taula 3.17.

*Respondre (**justificant detalladament el motiu de la resposta**) a cada pregunta:*

1. *Si les mostres recullen la totalitat de famílies d'ambdós barris, quin barri està més poblat? (és a dir, a quin barri hi ha més persones).*

2. *Quin barri té la vivenda més nombrosa i de quants membres consta?*

Taula 3.17: Dades de l'Exercici 3.5.12

	BARRI_A	BARRI_B
Mida mostra	255	227
Mitjana	2,29412	2,3304
Mediana	2,0	3,0
Moda	2,0	2,0
Desv. típica	1,1449	1,07304
Mínim	1,0	1,0
Màxim	7,0	7,0
Percentil 25	1,0	2,0
Percentil 75	3,0	3,0
Asimetria	1,00673	0,805318
Curtosi	1,54478	1,0633
C. de variació	0,4991	0,4605

3. *Quin dels dos barris és més homogeni en relació al nombre de membres de les seues vivendes?*

Exercici 3.5.13 *Responeu a les següents preguntes basant-te en la Taula (de freqüències) 3.18, on la variable és el* ***nombre de telefonades realitzades per cada abonat*** *de certa companyia de telefonia mòbil.*

Taula 3.18: Taula de freqüències de l'Exercici 3.5.13. Els punts suspensius indiquen una o més files que existeixen però que no podem consultar

x_i	n_i	f_i	N_i	F_i
0	58	0,006818716	58	0,006818716
1	284	0,033388197	342	0,040206913
2	643	0,075593699	985	0,115800612
3	1151	0,135316247	2136	0,251116859
⋮	⋮	⋮	⋮	⋮
10	135	0,01587115	8171	0,960616036
11	78	0,009169998	8249	0,969786033
12	49	0,00576064	8298	0,975546673
13	43	0,005055255	8341	0,980601928
14	27	0,00317423	8368	0,983776158
15	26	0,003056666	8394	0,986832824
⋮	⋮	⋮	⋮	⋮

1. *Si la mostra recull la totalitat d'abonats de la companyia, de quants abonats es tracta?*

2. *Quants abonats han realitzat 3 o menys telefonades?*

3. *Quin percentatge d'abonats va realitzar menys de 12 telefonades?*

4. *Quin percentatge d'abonats va realitzar més de 15 telefonades?*

Exercici 3.5.14 *Es realitza un estudi sobre els salaris mensuals de dos col·lectius de becaris: els becaris de Ministeri i els de Conselleria. Els resultats de l'enquesta es recullen en la Taula 3.19.*

Taula 3.19: Taules de freqüències de les dades arreplegades a l'Exercici 3.5.14

MINISTERI					
SOUS	450–500	500–550	550–600	600–650	650–700
N. BECARIS	8	10	23	5	2

CONSELLERIA				
SOUS	400–450	450–500	500–550	550–600
N. BECARIS	4	28	21	4

*Responeu, **raonant els càlculs que feu**, a les següents qüestions:*

1. *Quin col·lectiu es troba millor pagat en general?*

2. *Quin col·lectiu és més homogeni?*

3. *És cert que més del 75% de becaris de Conselleria es troba per davall dels 550 euros mensuals?*

Exercici 3.5.15 *Les donacions anuals, en euros, de 573 socis a una ONG vénen resumides, en euros, a la Taula 3.20.*

Taula 3.20: Taula de dades de l'Exercici 3.5.15

DONACIONS	0–10	10–25	25–50	50–100	100–250	250–500
Nº SOCIS	276	156	111	23	6	1

*Si cada soci pot deduir en la declaració de la renda el 15% de la seua donació anual, calculeu la **deducció mitjana** per soci.*

3.6 Pràctica R: 4. Descripció de mostres univariants

Objectius

Usar les capacitats de R per explorar i descriure mostres univariants a partir dels principals gràfics i estadístics implementats.

Tipus de variables

Les variables més senzilles que s'analitzen poden ser:

1. Qualitatives: les dades són etiquetes i només es poden considerar com iguals o distintes. Les dades també poden expressar graus d'una qualitat (com el nivell de satisfacció) que tenen un ordre, de menor a major.

2. Quantitatives: les dades expressen quantitats sobre...

   - Discretes: ...un concepte comptable per unitats (0, 1, 2...)
   - Contínues: ...un concepte mesurable sobre una escala comuna de nombres reals (com les llargàries, temps, o altres unitats que es podrien mesurar amb major o menor exactitud, usant més o menys nombres decimals)

Per algunes tasques, com ara alguns gràfics i la taula de freqüències, les variables quantitatives discretes s'assemblen més a les variables qualitatives que a les quantitatives discretes, ja que les repeticions són més habituals als dos primers tipus.

Descripció de mostres univariants qualitatives

Una mostra de dades univariants qualitatives pot estar emmagatzemada a un vector de tipus `character` o a una columna concreta d'un `data.frame`. Anem a treballar amb les dades creades pel codi:

```
set.seed(123456789)
d1vql <- sample(x=c('A', 'B', 'C', 'D', 'E', 'F'),
                        size=rpois(n=1, lambda=500),
                        replace=TRUE, prob=c(1,5,4,9,1,2))
```

Taula de freqüències: `table()`

La funció `table()` crea una taula de freqüències: torna un vector amb les freqüències absolutes (repeticions) de cada valor existent en el vector de dades que es passa com a argument.

En realitat, torna un vector les coordenades del qual estan etiquetades amb els valors corresponents. A l'exemple:

```
table(d1vql)
```

torna per pantalla:

```
d1vql
  A   B   C   D   E   F
 16 113  95 209  18  60
```

El vector tornat és el vector $(16, 113, 95, 209, 18, 60)$ mentre que les lletres que apareixen al damunt són etiquetes que ajuden l'usuari a saber cada freqüència a quina dada pertany.

La funció `table()` ordena el vector de freqüències per l'ordre de les categories de les dades (alfabètic). Normalment interessa escriure les freqüencies de major a menor, mostrant primer les dades més rellevants. Per obtenir aquesta modificació cal usar les funcions `sort()` i `rev()` mostrades a la Secció 1.6. Per exemple:

```
rev(sort(table(d1vql)))
```

torna per pantalla:

```
d1vql
  D   B   C   F   E   A
209 113  95  60  18  16
```

Ara el vector tornat té les freqüències ordenades, i les seues etiquetes. Si es vol obtenir una taula de freqüències amb els percentatges, en lloc de les freqüències absolutes, s'ha de fabricar a mà. A mode d'exemple:

```
rev(sort(table(d1vql)/length(d1vql)*100))
```

torna per pantalla:

```
d1vql
        D         B         C         F         E         A
40.900196 22.113503 18.590998 11.741683  3.522505  3.131115
```

Gràfics

Diagrama de barres `barplot()`

La funció `barplot()` fa el diagrama de barres corresponent a la taula de freqüències que es passe com a argument. Una sintaxi més completa és:

```
barplot(height, col = NULL, main = NULL, sub = NULL,
     xlab = NULL, ylab = NULL, xlim = NULL, ylim = NULL )
```

on:

- `height`: és el vector de freqüències. Normalment és `heigth=table(...)`.
- `col`: vector de colors (optatiu).
- `main, sub`: títol principal i secundari del gràfic.
- `xlab, ylab`: etiquetes per als eixos horitzontal i vertical.
- `xlim, ylim`: límits inferior i superior dels eixos horitzontal i vertical. Normalment es calculen automàticament, però l'usuari pot manipular-los.

Els valors `NULL` que figuren són valors per defecte. No cal escriure'ls si no els volem canviar. Prova amb:

```
barplot(height=table(d1vql))           # orden. per categ.
barplot(height=rev(sort(table(d1vql))))# orden. per freq.
```

Diagrama de sectors (pastís) `pie()`

La funció `pie()` fa un diagrama de barres a partir d'un vector amb les freqüències absolutes:

```
pie(x, labels = names(x), col = NULL, main = NULL, ...)
```

on:

- `x`: és el vector de freqüències. Normalment és `x=table(...)`.
- `labels`: etiquetes de les categories. Es prenen automàticament de la taula, però l'usuari pot canviar-les.
- `col`: vector de colors (optatiu).
- `main`: títol principal del gràfic.

Com sempre, hi ha més opcions per a usuaris avançats. Prova amb:

```
pie( x=table(d1vql) )
pie( x=rev(sort(table(d1vql))) )
n <- length(unique(d1vql))
pie( x=table(d1vql), col=grey((1:n)/n))
```

Moda

La moda, com a dada més present en la mostra de dades, es pot trobar ràpidament mirant la taula de freqüències. Una forma per a que la torne R sense haver de mirar-la seria agafar la primera component de la taula de freqüències quan està ordenada. Per exemple:

```
rev(sort(table(d1vql)))[1]
```

que torna per pantalla:

```
D
209
```

la moda i la seua freqüència absoluta.

Descripció de mostres univariants quantitatives

Les mostres univariants quantitatives es poden emmagatzemar en un vector de tipus `numèric` o en una columna concreta d'un `data.frame`. Anem a treballar amb les dades creades pel codi:

```
set.seed(123456789)
d1vqtd <- rpois(n=rpois(n=1, lambda=500), lambda=50 )
d1vqtc <- rnorm(n=rpois(n=1, lambda=500), mean=50, sd=10  )
```

que creen una mostra de dades de variables discreta i contínua, respectivament.

Taula de freqüències: `table()`

En el cas de variables discretes, les repeticions són freqüents, aleshores la taula es construeix com a la secció anterior. En el cas de variables contínues, les repeticions són poques, i l'abundància de dades diferents fa la tècnica inútil, i es recorre a fer intervals. Prova:

```
table(d1vqtd)
table(d1vqtc)
```

Les variables contínues es transformen en discretes formant intervals amb ajuda de la funció `cut()`, amb arguments:

- `x`: el vector numèric amb les dades a classificar en intervals.
- `breaks`: vector que expressa la partició dels intervals. El nombre d'intervals serà un menys que la llargària del vector de `breaks`.
- `right`: valor lògic que indica si els intervals són tancats per la dreta (`TRUE`) o per l'esquerra (`FALSE`).
- `include.lowest`: valor lògic que indica si es tanca (`TRUE`) o no es tanca (`FALSE`) l'interval (primer o últim) que quedava amb un extrem obert.

Una vegada creada la variable amb intervals es pot fer la taula d'aquesta nova mostra, és a dir:

```
a <- d1vqtc
intervals <- seq(from=min(a), to=max(a), length=10)
d1vqtc2 <- cut(x=d1vqt, breaks=intervals, include.lowest=TRUE)
table(d1vqtc2)
```

Gràfics (I)

Per a mostres de variables contínues amb poques dades

La distribució de les dades d'una xicoteta mostra es pot visualitzar molt convenientment amb un diagrama de punts. Aquest s'obté amb la funció `stripchart()`. Per exemple:

```
d1vqtc3 <- d1vqtc[1:30]
stripchart(d1vqtc3)
```

Quan hi ha moltes dades els punts formen línies contínues i no deixen apreciar com es ditribueixen.

Per mostres de variables contínues amb moltes dades

L'histograma és el gràfic convenient en aquest cas, i es demana amb la funció `hist()`. Per exemple:

```
hist( x=d1vqtc )
```

Una sintaxi més completa de la funció és:

```
hist(x, breaks = "Sturges", freq = NULL, probability = !freq,
          include.lowest = TRUE, right = TRUE,
          density = NULL, angle = 45, col = NULL, border = NULL,
          main = paste("Histogram of" , xname),
          xlim = range(breaks), ylim = NULL,
          xlab = xname, ylab,
          axes = TRUE, plot = TRUE, labels = FALSE,
          nclass = NULL, ...)
```

on els principals arguments són:

- `x`: la variable que conté les dades qualitatives.
- `breaks`: per defecte pren un valor típic, però l'usuari pot variar-lo donant un vector de punts de tall dels intervals.
- `freq`: el valor per defecte fa que les freqüències siguen absolutes. Un valor igual a `TRUE` fa que l'histograma siga com una funció de densitat (amb àrea total igual a 1).

Els altres arguments són semblants als ja utilitzats en la funció `plot()`, o prou intuïtius.

Un histograma en R, a banda de dibuixar un gràfic amb barres, és un objecte més complex que pots examinar si l'emmagatzemes en una variable. Prova a averiguar què té l'histograma anterior (fes que una variable l'emmagatzeme i després mostra la variable).

Per mostres de variables discretes

Una forma de representar mostres de dades de variables discretes, degut a les repeticions, és tractar-les com si foren variables qualitatives, i fer doncs un diagrama de barres, aprofitant que les categories s'ordenen per ordre alfabètic (que és el mateix que el numèric quan les etiquetes són nombres). Una altra forma és usar l'histograma, forçant la creació d'intervals, perquè hi haja un valor a cada interval. Per exemple:

```
barplot( table(d1vqtd) )
br <- seq( fr=min(d1vqtd)-0.5, to=max(d1vqtd)+0.5, by=1 )
hist( x=d1vqtd, breaks=br )
```

Estadístics

Els estadístics principals estan programats i només cal usar-los. Si la variable `x` té la mostra:

1. De posició:

   Mitjana aritmètica: `mean(x)`

   Mediana: `median(x)`

   Mínim: `min(x)`

   Màxim: `max(x)`

   Quantil d'ordre p: `quantile( x, prob=p )`

2. De dispersió:

   Recorregut: `diff( range(x) )`

   Recorregut interquartíl·lic: `IQR(x)`

   Variància mostral: `var(x)`

   Desviació típica mostral: `sd(x)`

   Coef. de variació de Pearson: `sd(x)/mean(x)`

Gràfics (II)

El diagrama de caixa: `boxplot()`

El diagrama de caixa i bigots és molt útil per intuir la posició central i dispersió de les mostres quantitatives. La funció `boxplot()` torna el gràfic. Per exemple:

```
boxplot(x)
```

Sobretot és útil en comparacions de mostres relacionades, i poden representar el gràfic de vàries mostres (`boxplot(x1, x2,...)`).

El diagrama de quantils: `ecdf()`

Un altre gràfic que dóna informació ràpida sobre la distribució de les dades és el diagrama de quantils. Amb aquest és molt ràpid contestar preguntes com 'quin percentatge de les dades de la mostra són inferiors a un valor donat?'. El diagrama de quantils es calcula amb la funció `ecdf()` i es dibuixa amb un `plot`. Per exemple:

```
plot( ecdf(d1vqtd) )
plot( ecdf(d1vqtc) )
```

Exercicis d'ensinistrament

Usa la mostra de dades que figura a l'arxiu `s4-descriptiva-1v-dades.txt` i carrega'l a una variable de R, per exemple `mostra`.

1. Per a la mostra univariant formada per la variable SEXE, obtín amb R:
   (a) La taula de freqüències.
   (b) El diagrama de barres.
   (c) El diagrama de sectors.
   (d) Quants homes conformen la mostra? Sol.: 98
   (e) Quin percentatge dels individus de la mostra són dones? Sol.: 44.31
2. Per a la mostra univariant formada per la variable SISOPER, obtín amb R:
   (a) La taula de freqüències absolutes i relatives, ordenades de major a menor freqüència.
   (b) El diagrama de barres amb les barres ordenades de major a menor freqüència.
   (c) Quin percentatge dels individus de la mostra usa Linux? Sol.: 51.70
3. Per a la mostra univariant formada per la variable NOTAFINAL:
   (a) Dibuixa un histograma de les notes amb 10 intervals.
   (b) Quants individus han superat l'assignatura? Sol.: 104
   (c) Quina nota prendries per representar el nivell del grup? Sol.: 6.12 (o 5.78)
   (d) Quin valor descriuria la variabilitat de notes en el grup? Sol.: 2.99 (altres valors són possibles)
   (e) Quin sexe pots considerar que té un nivell superior de la nota final? Sol.: Les dones ($6.354 > 5.948$)
   (f) Quin sexe pots considerar que forma un grup més homogeni respecte al nivell de la nota final? Sol.: Les dones ($0.4623544 < 0.5108202$)
   (g) Compareu les notes dels alumnes agrupats per sexe, usant un gràfic que visualitze bé la posició i la dispersió.
   (h) Compareu les notes dels alumnes agrupats per sistema operatiu, usant un gràfic que visualitze bé la posició i la dispersió.
   (i) Si el professor vol tenir un percentatge de suspensos del 30%, quina nota de tall divisòria hauria d'establir? Sol.: 4.33

Capítol 4

Descripció de mostres de dades multivariants

4.1 Què són i com es representen

Normalment, les mostres que interessen als problemes reals són multivariants, perquè la realitat és complexa i hi ha molts aspectes interrelacionats en cada situació. L'Estadística pot ajudar molt a refutar, confirmar o fins i tot quantificar aquestes relacions que, en un principi, ni se sospiten.

Com exemple citem el de la pàgina 11, on el professor d'una assignatura recopilava dades dels seus alumnes sobre 6 variables (vegeu la Taula 1.1 a la pàgina 11).

En aquest curs introductori només estudiem les mostres bivariants, a les quals només s'analitzen dues variables. Una mostra bivariant és una llista de dades dobles. Si anomenem X i Y a les variables, la mostra es pot escriure com figura a la Taula 4.1.

Taula 4.1: Notació per a una mostra de dades bivariants de mida n

ID	1	2	3	$\cdots$	n
X	x_1	x_2	x_3	$\cdots$	x_n
Y	y_1	y_2	y_3	$\cdots$	y_n

Cada dada és una parella (x_i, y_i). La representació eficaç de mostres bivariants inclou dues versions.

- Numèricament: es fa recompte de les repeticions, si n'hi ha, de les dades trobades de cada tipus **(Taula de freqüències)**.
- Gràficament:
  - **Diagrama de barres** (si les dues variables són qualitatives o quantitatives discretes).
  - **Diagrames de punts** o **boxplots** comparatius, (si una és qualitativa i l'altra quantitativa).

– **Diagrama de dispersió** o **núvol de punts** (si les dues variables són quantitatives contínues).

4.1.1 Taula de freqüències

Quan hi ha poques categories de cada variable i moltes repeticions, la taula de freqüències és útil, encara que, com en el cas univariant, no tant ara que els ordinadors poden fer càlculs immediats. L'estructura de la taula és com la mostrada a la Taula 4.2

Taula 4.2: Taula de freqüències absolutes i relatives d'una mostra bivariant (X, Y).

$X/\quad Y$	y_1	y_2	$\cdots$	y_l	Fr.abs.(rel.)
x_1	$n_{11}(f_{11})$	$n_{12}(f_{12})$	$\cdots$	$n_{1l}(f_{1l})$	$n_{1\cdot}(f_{1\cdot})$
x_2	$n_{21}(f_{12})$	$n_{22}(f_{22})$	$\cdots$	$n_{2l}(f_{21})$	$n_{2\cdot}(f_{2\cdot})$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	$n_{k1}(f_{k1})$	$n_{k2}(f_{k2})$	$\cdots$	$n_{kl}(f_{kl})$	$n_{k\cdot}(f_{k\cdot})$
Fr.abs.(rel.)	$n_{\cdot 1}(f_{\cdot 1})$	$n_{\cdot 2}(f_{\cdot 2})$	$\cdots$	$n_{\cdot l}(f_{\cdot l})$	$n(1.000)$

La taula de freqüències sempre presenta les freqüències absolutes de cada parella de categories. No obstant, segons la intenció de l'autor, les freqüències relatives es poden calcular: (1) respecte de tota la mostra, (2) respecte de les dades de cada fila, o (3) respecte de les dades de cada columna, donant lloc a taules alternatives.

La notació usada a la taula es detalla a continuació:

- x_i: categoria i-èsima de la variable X.
- y_j: categoria j-èsima de la variable Y.
- n_{ij}: freqüència absoluta de la categoria (x_i, y_j). Indica el nombre de repeticions de la dada (x_i, y_j) dins la mostra.
- n: mida de la mostra. Per tant, $n = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}$
- f_{ij}: freqüència relativa de la categoria (x_i, y_j) dins la mostra. Indica la proporció de dades de la mostra coincidents amb la categoria (x_i, y_j). Per tant,

$$f_{ij} = \frac{n_{ij}}{n}, \qquad \% = f_{ij} \times 100, \qquad \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} = 1.000$$

- $n_{i\cdot}$ (i $f_{i\cdot}$): freqüència absoluta (i relativa) de la categoria x_i. S'anomena freqüència marginal de la variable X, perquè deixa al marge l'altra variable.

- $n_{\cdot j}$ (i $f_{\cdot j}$): freqüència absoluta (i relativa) de la categoria y_j. S'anomena freqüència marginal de la variable Y, perquè deixa al marge l'altra variable.

Existeix el concepte de **distribució condicionada**, que mostrem breument. Per exemple, la distribució de X condicionada a $Y = y_3$ seria aquella que a cada x_i correspon la freqüència absoluta n_{i3} i la freqüència relativa $\frac{n_{i3}}{n_{\cdot 3}}$ (ja que només es tenen en compte les dades de $Y = y_3$). La notació particular és:

$$n_{i|3} = n_{i3}, \qquad f_{i|3} = \frac{n_{i3}}{n_{\cdot 3}} = \frac{f_{i3}}{f_{\cdot 3}}$$

Exercici 4.1.1 *Una enquesta sobre SEXE i SISTEMA OPERATIU (SO) dóna resultat a la Taula 4.1*

- *Quin percentatge de persones de la mostra és dona i usa Linux?*
- *Quin percentatge de persones de la mostra usa Linux? I Windows?*
- *Quin percentatge de persones de la mostra és dona? I home?*
- *Quin percentatge de persones de la mostra és dona o usa Linux?*
- *Quin percentatge de persones que usen Linux és dona?*
- *Quin percentatge de les dones usa Linux? I dels homes?*

SEXE \ SO	Linux	MacOS	Windows	Fr.abs.(rel.)
Dona	42(0.210)	18(0.090)	37(0.185)	97(0.485)
Home	47(0.235)	23(0.115)	33(0.165)	103(0.515)
Fr.abs.(rel.)	89(0.445)	41(0.205)	70(0.350)	200(1.000)

Figura 4.1: Taula de freqüències de l'Exercici 4.1.1

Quan les variables no són qualitatives, la taula de freqüències s'ha de fer com al cas univariant, usant intervals com categories.

4.1.2 Representació gràfica

En el cas de tractar dues variables qualitatives, una representació útil és el **diagrama de barres conjunt** (es fa el diagrama de barres d'una variable, i es coloreja dins cada barra segons la distribució de dades de l'altra variable) i una altra encara millor, la representació de les freqüències condicionades (on es divideix per l'alçada de cada barra, vegeu la Figura 4.2).

Quan tenim el cas d'una variable qualitativa i una altra quantitativa, els gràfics convenients són els diagrames de punts i els de caixes alineats (vegeu la Figura 4.3).

Per últim, a l'estudi conjunt de dues variables quantitatives, el gràfic a utilitzar és el **diagrama de dispersió** o **núvol de punts** (vegeu la Figura 4.4).

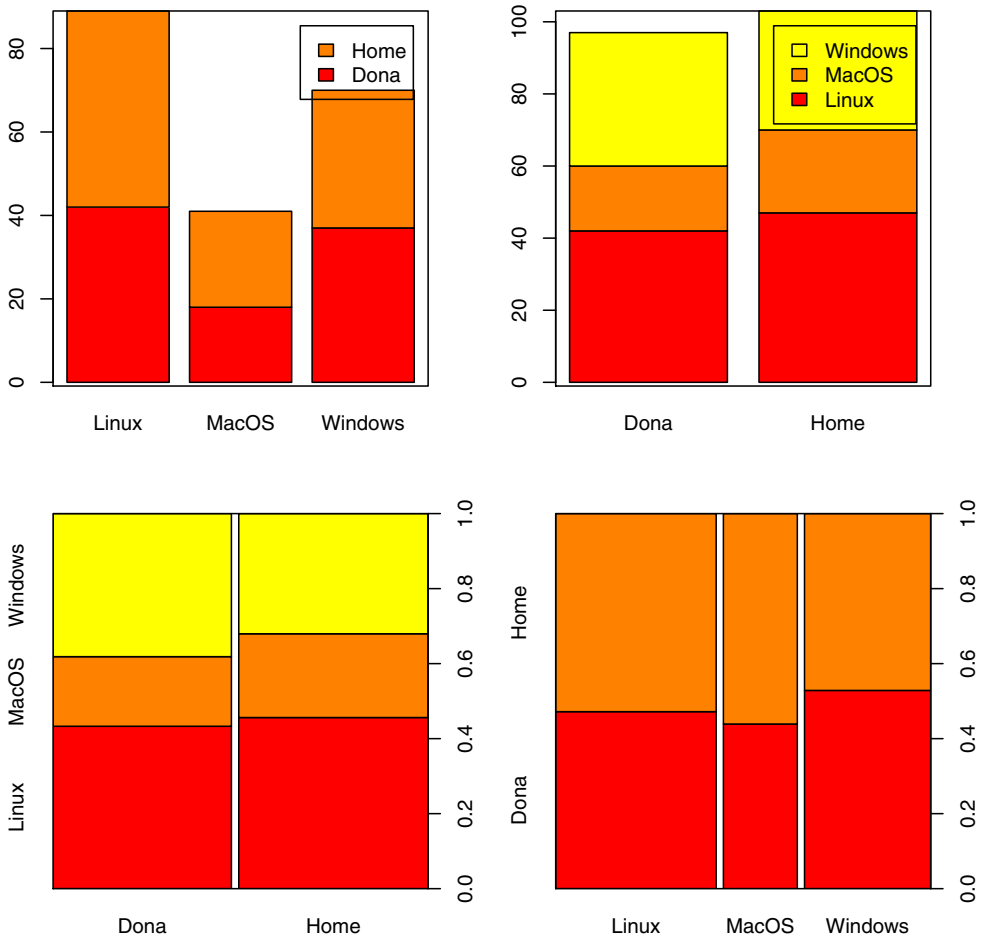


Figura 4.2: Dues representacions gràfiques equivalents per a l'estudi conjunt de dues variables qualitatives. La línia superior té diagrames de barres, la inferior mostra les distribucions condicionades (més convenients per a l'avaluació intuïtiva de la independència entre les variables). En aquest cas corresponents a la mostra de l'Exercici 4.1.1

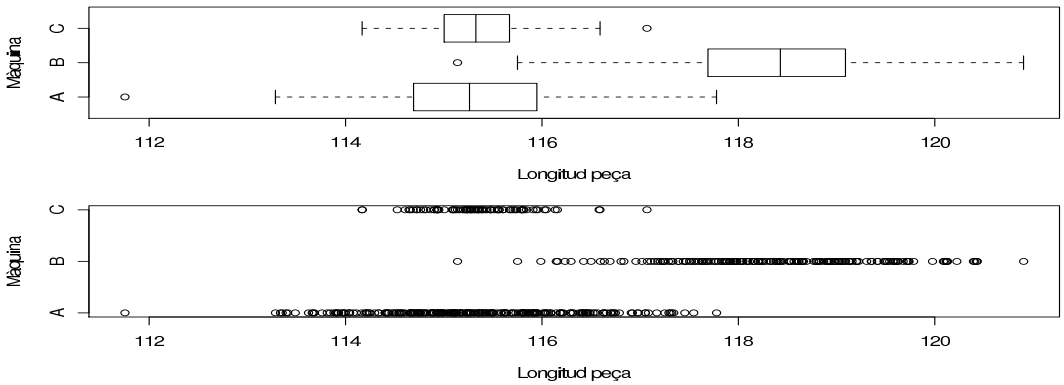


Figura 4.3: Dues representacions gràfiques equivalents per a l'estudi conjunt d'una variable qualitativa amb una quantitativa. En aquest cas corresponents a una mostra de peces fabricades, de les quals s'investiga la seua llargària i la màquina (A, B o C) que les ha fabricades

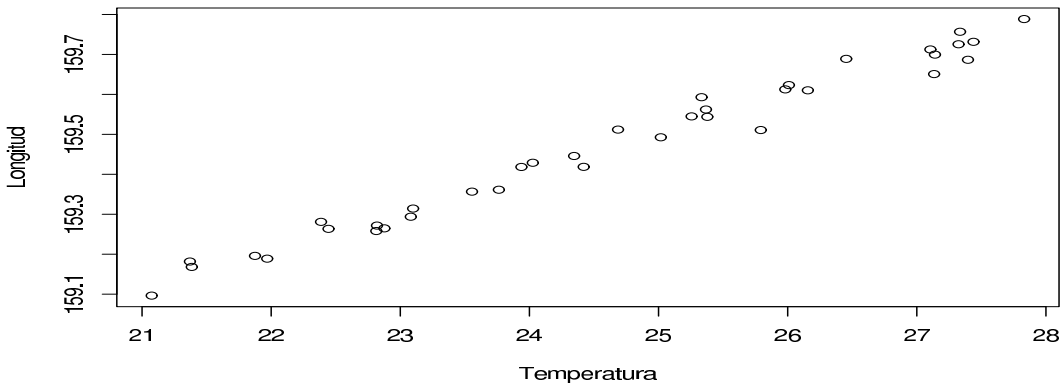


Figura 4.4: Diagrama de dispersió (també anomenat núvol de punts) corresponents a l'estudi conjunt de dues variables quantitatives. En aquest cas corresponents a una mostra d'objectes mesurats a diferents temperatures

4.2 Independència estadística entre variables

El concepte de dependència/independència estadística entre dues variables és senzill.

Definició 4.2.1 (Independència estadística entre variables) *Quan la distribució d'una variable Y no canvia en condicionar-la als distints valors de l'altra variable X, es diu que Y és* ***independent*** *de X.*

En el cas que Y siga independent de X, es pot provar matemàticament que també X és independent de Y, i aleshores es diu directament que X i Y són independents.

Exercici 4.2.1 *Usant la taula de freqüències de l'Exercici 4.1.1, escriviu les distribucions de la variable SEXE condicionades als distints valors de la variable SO, i comproveu si són similars o molt diferents.*

Quan no hi ha independència estadística, una de les variables sol venir afectada pel valor de l'altra, en una direcció concreta. Segons el context:

- La variable que **rep la influència** es diu variable **explicada**, **resposta**, **dependent**...

- La variable que **exerceix la influència** es diu variable **explicativa**, **de control**, **independent**...

Exemples:

- La velocitat de connexió depén del nombre d'usuaris connectats.
- El temps de caiguda lliure d'un cos depén de l'alçada.
- La longitud depén de la temperatura (dilatació).
- La salut d'una persona depén (entre altres) del consum de tabac (OMS).
- La nota en l'examen dependrà del temps emprat en la preparació de l'assignatura.

Descartar la independència entre dues variables usant mostres és més laboriós si es fa numèricament.

La definició d'independència es pot verificar més ràpidament usant la representació gràfica conjunta de les variables. Les representacions gràfiques de les mostres bivariants són poderoses per descartar la independència estadística o, al contrari, per detectar intuïtivament possibles relacions entre variables, siguen del tipus que siguen.

Per exemple, en el cas de l'estudi conjunt de dues variables qualitatives (vegeu la Figura 4.2), la independència entre les variables es "capta" quan la distribució de colors dins cada barra (d'un dels gràfics) és sempre la mateixa o

molt similar. És a dir, quan les barres semblen totes proporcionals entre si. En un altre cas caldria pensar que una variable influeix sobre l'altra. Et sembla que, a la Figura 4.2, el valor de la variable SEXE afecta la distribució de la variable SO?

En el cas de l'estudi conjunt d'una variable qualitativa amb una quantitativa, la independència entre les variables involucrades es materialitza en uns diagrames molt similars (en posició i dispersió). Si no es dóna el cas que tots els diagrames són similars, aleshores s'ha de rebutjar la possibilitat d'independència, i pensar que una de les variables afecta la distribució de l'altra. Et sembla que, a la Figura 4.3, la màquina escollida afecta la distribució de les longituds de les peces?

Per últim, en el cas de l'estudi conjunt de dues variables quantitatives (vegeu la Figura 4.4), la independència entre les variables es "capta" quan el núvol de punts no té cap forma definida (encara que ací s'hauria de matisar amb les distribucions de les variables per separat). Et sembla que, a la Figura 4.4, el valor de la temperatura afecta la longitud de la peça observada?

4.3 Estadístics de posició i dispersió

Quan les dues variables (X, Y) són quantitatives podem resumir la mostra en una posicio central i un grau de dispersió:

- Per a la posició, usem la mitjana, ara de dues variables:

$$\overline{(X,Y)} = \frac{\sum_i (x_i, y_i)}{n} = (\frac{\sum_i x_i}{n}, \frac{\sum_i y_i}{n}) = (\overline{x}, \overline{y})$$

- Per a la dispersió, si usem la variància tenim una dispersió:

$$\frac{\sum_i \left\| (x_i, y_i) - \overline{(x,y)} \right\|^2}{n-1}$$

  Treballant amb vectors, el quadrat es transforma en el producte per la transposta, amb el qual es té com a resultat una matriu, que s'anomena de variàncies-covariàncies:

$$\begin{pmatrix} \frac{\sum_i (x_i-\overline{x})^2}{n-1} & \frac{\sum_i (x_i-\overline{x})(y_i-\overline{y})}{n-1} \\ \frac{\sum_i (y_i-\overline{y})(x_i-\overline{x})}{n-1} & \frac{\sum_i (y_i-\overline{y})^2}{n-1} \end{pmatrix} = \begin{pmatrix} \mathrm{s}_X^2 & \mathrm{s}_{XY} \\ \mathrm{s}_{YX} & \mathrm{s}_Y^2 \end{pmatrix}$$

Definició 4.3.1 (Covariància i Correlació) *Es defineix la covariància mostral entre les variables X i Y com:*

$$s_{XY} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

Per altra banda es defineix el coeficient de correlació entre les variables X i Y com:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

La covariància canvia si canviem les unitats de les variables, el coeficient de correlació, no (en valor absolut). A més $r_{XY} \in [-1.0, 1.0]$ sempre!

La fórmula de la covariància és tant intuïtiva que el seu signe (positiu o negatiu) es pot intuir en molts casos a partir del núvol de punts (vegeu la Figura 4.5), mirant els quadrants que determinen els valors de les mitjanes d'ambdues variables.

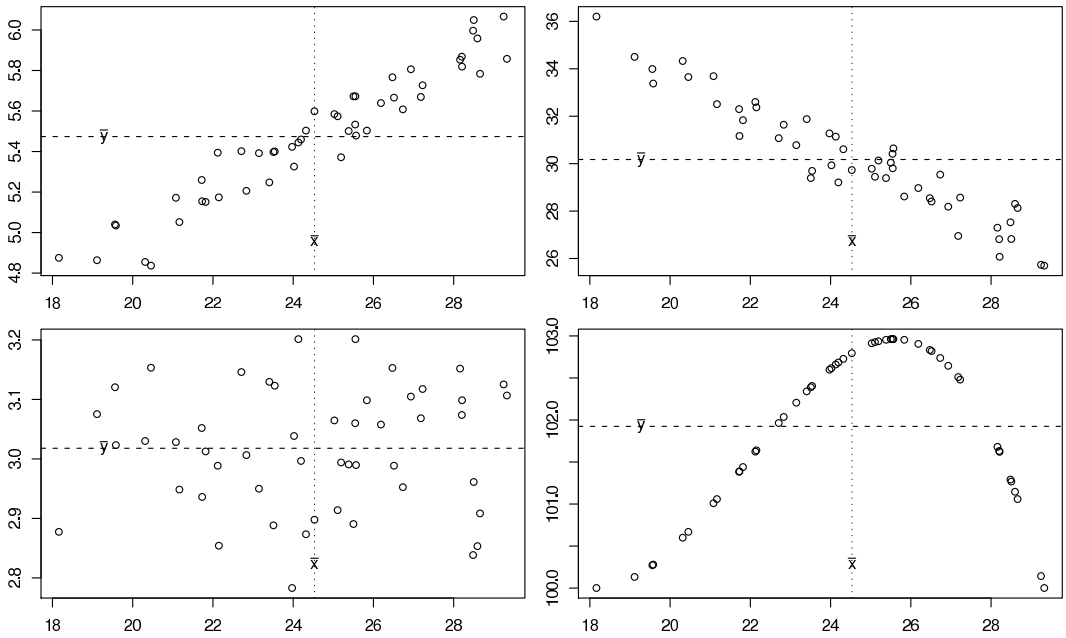


Figura 4.5: Quatre núvols de punts. Al superior esquerre es pot deduir que la covariància (i per tant la correlació) serà clarament positiva. Al superior dret es té el cas contrari, amb signe clarament negatiu. A l'inferior esquerre, el signe és impredictible. A l'inferior dret, es pot intuir que serà lleugerament positiva

El coeficient de correlació, com es comenta en la secció següent, té la sorprenent capacitat d'indicar el grau d'aproximació del núvol de punts que és la mostra a una hipotètica línia recta que passa "el més prop possible" del núvol.

4.4 Anàlisi de regressió: cas lineal

4.4.1 Càlcul de la funció

Quan s'intueix certa dependència entre dues variables, interessa saber si aquesta dependència és casualitat o obeeix a alguna relació entre elles (causalitat: causa-efecte), encara que puga ser a través d'una tercera variable intermediària (per exemple, les celebracions que es poden programar se solen fer durant la primavera o l'estiu, i no és la temperatura la que governa la decisió de la data de la celebració, sino una sèrie de factors per gaudir-la, entre els quals entra el bon oratge).

Encara que la regressió es pot fer sempre, perquè siga un procediment científic és necessari que:

1. Vinga motivada a priori per sospites fondamentades en la teoria on s'emmarca l'estudi.

2. El resultat es contraste amb les dades mitjançant un gràfic o un coeficient de "bondat d'ajustament".

3. Les conclusions que es traguen no se n'isquen dels marges de les dades de la mostra.

Si Y és la variable resposta i X és la de control, trobar la relació entre les variables consisteix en trobar una equació $Y^* = f(X)$ que siga el més consistent possible amb les dades de la mostra.

Suposem que la mostra de dades bivariants ve expressada a les dues primeres files de la Taula 4.3, i que disposem d'una funció f que calcula possibles valors de Y a partir de valors de X.

Taula 4.3: Taula amb la mostra de dades (files primera i segona), els valors calculats per la funció de regressió (fila tercera), i els errors o residus de la funció de regressió (fila quarta)

X	x_1	x_2	x_3	$\cdots$	x_n
Y	y_1	y_2	y_3	$\cdots$	y_n
Y^*	$y_1^* = f(x_1)$	$y_2^* = f(x_2)$	$y_3^* = f(x_3)$	$\cdots$	$y_n^* = f(x_n)$
E	$e_1 = y_1 - y_1^*$	$e_2 = y_2 - y_2^*$	$e_3 = y_3 - y_3^*$	$\cdots$	$e_n = y_n - y_n^*$

La "millor" funció f es tria amb els criteris:

1. Que siga senzilla.

2. Que tinga mínima discrepància amb les dades, és a dir, minimitzant, per exemple, la suma de quadrats dels errors.

En molts casos el núvol de punts informa l'investigador del tipus de funció que convé per fer la regressió, encara que per conéixer-la siga necessari fer càlculs. Per exemple:

1. **Tipus lineal**: $Y^* = a + bX$.

2. Tipus exponencial: $Y^* = a\mathrm{e}^{bX}$.

3. Tipus potencial ($Y^* = aX^b$).

4. Tipus parabòlic ($Y^* = a + bX + cX^2$).

5. ...

on a i b (i c si hi ha) són nombres per determinar amb la condició de mínima suma d'errors al quadrat.

Si acceptem de partida que la funció de regressió adequada és la lineal, usant derivades s'obté que **la recta que millor s'aproxima a un núvol de punts** (x_i, y_i) **és** aquella que:

$$b = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum_i x_i)^2} = \frac{\mathrm{s}_{XY}}{\mathrm{s}_X^2} \qquad a = \overline{y} - b\overline{x}$$

4.4.2 Bondat d'ajustament

Una volta calculada la millor funció explicativa de Y en funció de X, cal comprovar si és prou bona, o encara que sent la millor, no serveix per a molt.

Una forma de mesurar la bondat de la regressió és crear un estadístic que siga fàcilment interpretable. Per exemple, la suma de quadrats de la variable Y es pot descompondre com figura a continuació:

$$\underbrace{\sum_i (y_i - \overline{y})^2}_{\mathrm{SQ}_y} = \underbrace{\sum_i (y_i - y_i^*)^2}_{\mathrm{SQ}_{\mathrm{error}}} + \underbrace{\sum_i (y_i^* - \overline{y})^2}_{\mathrm{SQ}_{\mathrm{regres}}}$$

D'aquesta expressió es pot interpretar que la variància de Y està causada, per una part, per la variància dels errors de la regressió, i per una altra, per la variància de la pròpia funció de regressió Y^*, és a dir, per la variància causada per la variable X.

Definició 4.4.1 (Bondat d'ajustament) *Es defineix el* ***coeficient de determinació*** *com:*

$$R^2 = 1 - \frac{\mathrm{SQ}_{\mathrm{error}}}{\mathrm{SQ}_y}$$

El coeficient de determinació lineal R^2 s'interpreta com la part de variància de Y que sí que pot explicar-se amb la funció de regressió. Per tant té un valor màxim d'1 i un valor mínim de 0. A major valor de R^2, major bondat de la regressió.

Propietat 4.4.1 $R^2 \in [0.0 - 1.0]$ *sempre!*

Propietat 4.4.2 *En el cas de regressió lineal,* $R^2 = r_{XY}^2$.

A banda d'aquest estadístic, és convenient no deixar la tasca de decidir sobre la bondat de l'ajustament a un únic valor. Un gràfic dels errors (residus) en funció del valor de X pot donar pistes sobre la conveniència del model triat, o la consideració de punts que influeixen molt al resultat, i que potser caldria qüestionar. La Figura 4.6 està extreta de [1], on quatre mostres de dades especialment triades donen resultats sorprenents.

4.4.3 Prediccions

Una anàlisi de regressió que acaba amb un R^2 pròxim a 1.0 indica que les dades de Y vénen ben ajustades per la funció $Y^* = f(X)$. Gràficament, el núvol de punts s'ajusta prou a la gràfica de la funció $Y^* = f(X)$.

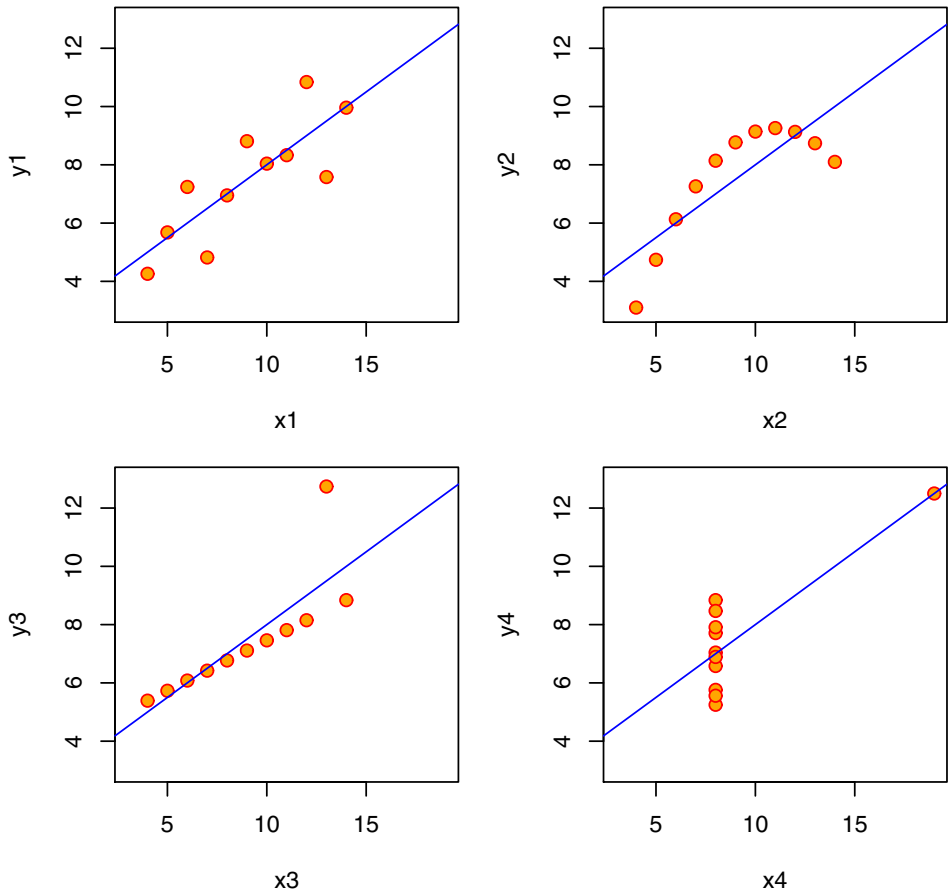


Figura 4.6: Exemple de [1], on relacions molt diferents entre dues variables, com es poden apreciar als gràfics, donen lloc als mateixos estadístics, i per tant al mateix ajustament amb la mateixa bondat, el qual abunda en la necessitat de no fiar-se només dels estadístics.

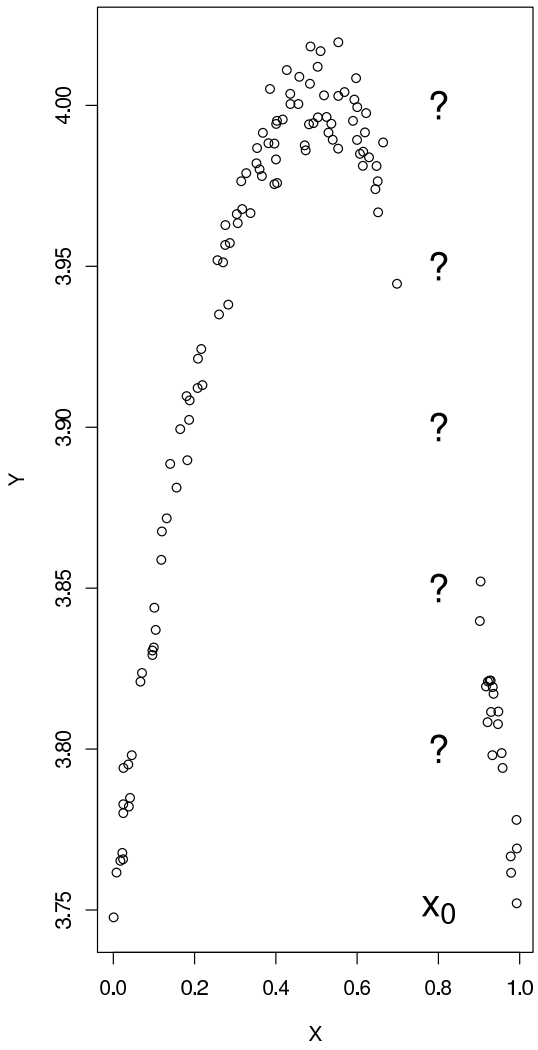


Figura 4.7: Exemple de núvol de punts de mostra de dades bivariants. No hi ha cap dada amb $X = 0.8$. Si fóra important estimar el valor de Y corresponent a $X = 0.8$, però no es puguera obtenir més mostra, com es podria resoldre?

Per a un nou valor de $X = x_0$ podem fer una **predicció** de valor de Y, calculant $y_0^* = f(x_0)$ (vegeu les Figures 4.7 i 4.8).

Podem definir la fiabilitat o qualitat de la predicció com al nivell de credibilitat que aqueixa predicció Y^* siga similar al valor real de Y corresponent al valor de $X = x_0$ i obtingut ampliant la mostra.

Així definida, la fiabilitat o qualitat de la predicció es pot mesurar amb el coeficient R^2, ja que si un núvol de punts s'ajusta molt al gràfic d'una funció, aleshores els valors Y de la mostra són similars als valors obtinguts amb la funció.

Hi ha estudis on és molt car o costós obtenir una gran mostra. En aquests casos la tècnica de regressio és útil: es poden considerar més dades sense cost addicional. Encara que sempre hi ha el risc que les prediccions no s'ajusten a la realitat, aquest risc és menor quan major siga el coeficient R^2.

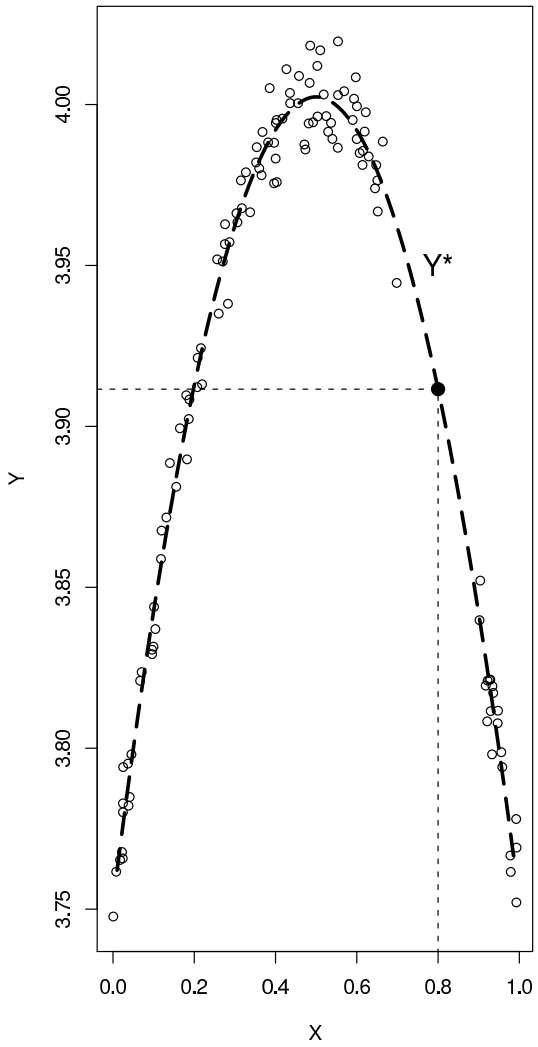


Figura 4.8: Solució al problema plantejat a la Figura 4.7. Es calcula la funció de regressió més raonable, i se substitueix el valor $X = 0.8$ per a obtenir una predicció del valor real de Y

Per últim, és important advertir que la interpretació del coeficient de deter-

minació com a qualitat de la funció de regressió i de les prediccions fetes amb aquesta, està limitada a l'interval de valors on se situa la mostra de dades.

A la Figura 4.9 es mostra el cas on es fa un estudi consistent a avaluar els resultats d'una dieta. S'arrepleguen 10 dades les primeres setmanes, i la recta de regressió calculada és de molta qualitat predictiva (ja que $R^2 = 0.98$). No obstant, usar aquesta recta per fer prediccions a llarg termini pot portar a informacions molt equivocades, i per tant a decisions incorrectes.

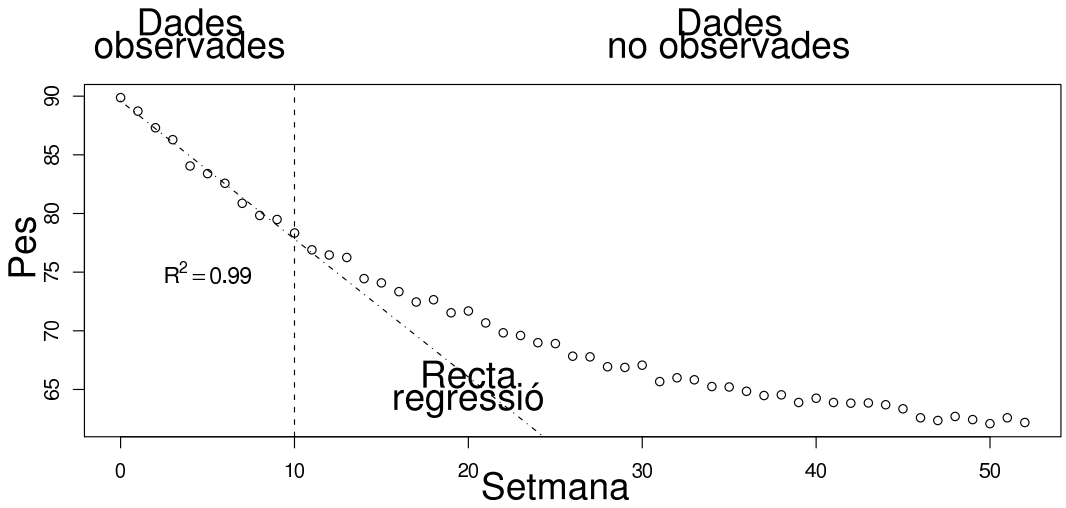


Figura 4.9: Perills d'usar l'anàlisi de regressió fora del marc on es té la mostra de dades. La mostra es restringeix als valors de X fins al 10. Les prediccions fetes dins d'aquest interval tenen una alta fiabilitat. Les prediccions fetes fora de l'interval són més incertes ja que no és possible saber com evolucionarien les dades reals

4.5 Exercicis proposats

Exercici 4.5.1 *Una acadèmia que prepara per fer oposicions fa simulacres d'examen al final del curs. També aconsegueix registrar la nota dels seus alumnes quan passen per l'oposició. Així, tenen les dades del curs passat (que van seguir el curs 8 alumnes):*

Alumne	1	2	3	4	5	6	7	8
Nota acadèmia	8.8	8.8	10.0	8.1	9.6	9.6	8.4	7.7
Nota oposició	9.0	8.5	9.6	8.2	9.6	9.7	8.7	7.9

En base a aquestes dades, un estudiant que aquest any trau una nota de 7 punts al simulacre de l'acadèmia, quina nota podria estimar que obtindria a l'oposició i quina fiabilitat (alta o baixa) tindria la predicció?

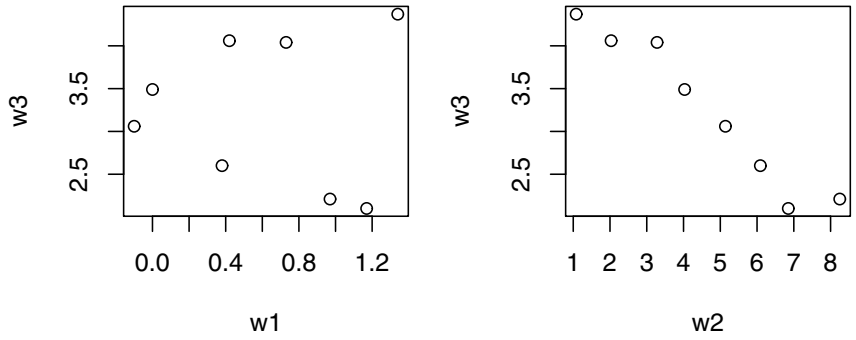
Exercici 4.5.2 *Un laboratori investiga la fabricació d'un adhesiu, basat en una disolució de cianocrilat. Usant una concentració de cianocrilat del 5%, es mesuren amb un aparell les resistències, en Nw, de l'adhesiu per investigar la influència de la temperatura sobre el poder d'adherència (vegeu la Taula 4.4).*

Taula 4.4: Dades de l'Exercici 4.5.2

Temp	20	22	24	26	28	30
Resist	1759.31	1757.75	1763.32	1753.21	1752.94	1752.88
Temp	32	34	36	38	40	
Resist	1750.92	1742.62	1735.02	1739.82	1742.13	

Un fabricant d'una localitat d'Alaska (on hi ha una temperatura habitual de $-15°C$*) necessita usar un adhesiu per a assemblar les peces que conformen els seus productes. Demana informació sobre l'adhesiu al laboratori i el laboratori li facilita les dades de la tabla anterior. El fabricant considera important conéixer la resistència de l'adhesiu abans de decidir-se a utilitzar-lo. En base a alguna tècnica estadística, quina predicció de resistència pot suposar que tindrà l'adhesiu? Es pot fiar molt o poc d'aquesta predicció?*

Exercici 4.5.3 *Un estudi teòric revela que una variable* ***w3*** *podria estar relacionada amb una de les variables* ***w1*** *o* ***w2****. Per tal de comprovar aquesta relació es presenten els gràfics*



Si les dades són les que figuren a la taula:

w1	1.34	0.42	0.73	0.00	−0.10	0.38	1.17	0.97
w2	1.08	2.03	3.28	4.03	5.14	6.09	6.85	8.25
w3	4.37	4.06	4.04	3.49	3.06	2.60	2.10	2.21

1. *Calculeu la recta de regressió que millor aproxima els valors de* ***w3*** *en funció dels valors de l'altra variable més convenient (****w1*** *o* ***w2****).*

2. *Al laboratori l'interessa pronosticar el valor de* ***w3*** *per un valor de* 15.5 *(de la variable que has usat en la regressió,* ***w1*** *o* ***w2****). Doneu una resposta professional al laboratori.*

Exercici 4.5.4 *Es planteja estudiar l'efecte que té la distància entre dos servidors de correu electrònic sobre el temps que empra un missatge a arribar d'un a altre servidor. Les 8 proves recollides són:*

Missatge	1	2	3	4	5	6	7	8
Distància (km)	365	389	534	125	350	890	1008	1167
Temps (s)	0.07	0.54	0.09	0.11	0.23	0.91	0.18	0.33

1. *Calculeu una estimació del temps que tardaria en arribar un correu electrònic enviat entre servidors que disten* 750 *km, usant alguna tècnica estadística inclosa en el programa de l'assignatura.*

2. *Valoreu la qualitat d'aquesta estimació justificant en què et bases.*

Exercici 4.5.5 *Una enquesta sobre el nivell de satisfacció dels clients per un servei (valorat des de "Molt desfavorable" fins a "Molt favorable") analitza les respostes conjuntament amb el sexe (home o dona) del client. Segons el gràfic resultant, mostrat a la Figura 4.10. Influeix el sexe en la distribució del nivell de satisfacció dels clients?*

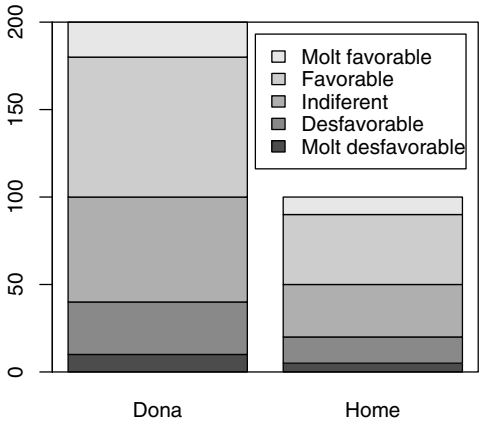


Figura 4.10: Gràfic de l'Exercici 4.5.5

Exercici 4.5.6 *Una mostra, recollida per un estudi sobre el descens de la concentració d'una substància en la sang amb el pas del temps, dóna lloc a les dades que es presenten a la Taula 4.5.*

Taula 4.5: Dades de l'Exercici 4.5.6

Temps (h)	1	2	3	4	5	6
Concentració (mg/l)	122.80	100.54	82.32	67.39	55.18	45.17

Realitzeu prediccions de la concentració que hi haurà a les 3 hores i mitja d'haver consumit la substància, i a les 10 hores, i raoneu sobre la qualitat de les prediccions realitzades.

Exercici 4.5.7 *Un psicòleg afirma, en base a una mostra obtinguda, que el nombre de respostes inadequades que dóna un xiquet en el transcurs d'una situació experimental disminueix a mesura que el xiquet creix. La Taula 4.6 recull les dades de què disposa el psicòleg.*

1. *Escriviu la taula de freqüències conjuntes.*

2. *Calculeu el percentatge de xiquets que cometen entre 10 i 13 errors, i el percentatge, d'entre aquests, que tenen entre 2 i 4 anys.*

3. *Quin és el nombre de respostes inadequades que es pot predir per a un xiquet de 10.5 anys? És fiable aquesta predicció?*

Exercici 4.5.8 *Per comprovar l'existència d'una relació lineal entre la temperatura (X) a què treballa un microprocessador (mesurada en graus Kelvin, $^\circ K$) i el rendiment (Y) mesurat en bilions d'operacions per segons (bops), es pren una sèrie de dades, amb els resultats que es mostren a la Taula 4.7.*

Responeu amb raonaments basats en les xifres de la taula:

1. *En més de la meitat de les observacions, el rendiment del microprocessador és superior a 47 bops, vertader o fals?*

2. *No disposem d'un gràfic, però, segons les dades que es mostren de l'estudi, a mesura que la temperatura augmenta, el rendiment del microprocessador... creix o decreix?*

3. *Si un sistema en el qual es va a usar el microprocessador va a treballar a $230^\circ K$, quina quantitat de bops se suposa que realitzarà en base a les dades?*

4.6 Pràctica R: 5. Descripció de mostres bivariants

Objectius

Usar les capacitats de R per explorar i descriure mostres bivariants a partir dels principals gràfics i estadístics implementats, amb especial atenció al descobriment de relacions de dependència entre les variables.

Tipus de variables

En estudiar mostres bivariants, els tractaments que es poden fer a les dades depenen de la seua natura (tipus), i per tant és necesari distingir els tres casos: (1) qualitativa vs qualitativa, (2) qualitativa vs quantitativa i (1) quantitativa vs quantitativa, tenint en compte, que la variable quantitativa discreta, encara que quantitativa, per allò de les repeticions de les dades, es pot considerar en ocasions com qualitativa.

Les dades que anem a usar com exemple són les emmagatzemades a l'arxiu `s5-descriptiva-2v-dades.txt`

Taula 4.6: Dades de l'Exercici 4.5.7

Edat	2	3	4	4	5	5	6	7	7	9	9	10	11	11	12
Resp.	11	12	10	13	11	9	10	7	12	8	7	3	6	5	5

Taula 4.7: Estadístics de l'Exercici 4.5.8

$n = 157$	$\overline{x}$	s	Mín	P_{25}	Med	P_{75}	Màx
X	273.0	16.8	223.0	259.6	271.5	292.7	333.0
Y	45.8	3.6	37.9	41.4	48.1	51.1	57.3
		$s_{XY} = 59.27$					

```
m <- read.table(file='s5-descriptiva-2v-dades.txt', header=T)
```

Per a analitzar les dades d'un full de dades en una o unes variables concretes, aquestes s'han de triar amb l'operador `[ ]` (o usant l'operador `$` si només és una variable). Per exemple:

```
# per analitzar la var. sexe treballarem amb
m$sexe                  #  o m["sexe"] o m[1]
# per analitzar conjuntament sexe i notafinal
m[ c("sexe", "notafinal") ]      # o m[c(1,4)]
```

Taula de freqüències `table()`

La funció `table()` també funciona amb mostres multivariants, i crea una taula de contingència (cal recordar que si s'usa sobre alguna variable quantitativa contínua —sense dades repetides—, seria necessari transformar aquesta variable en una més convenient, tallant en intervals com s'indicava en la pàgina 61).

La taula de freqüències té tantes dimensions com variables, per tant només es visualitzen bé les taules de mostres bivariants, que tindran l'aspecte de matrius. Prova amb:

```
table(m[c(1,2)])
table(m[c(2,1)])
```

i observa la diferència.

Gràfics

La funció `plot()` fa diagrames de punts de tots els emparellaments de variables presents a la mostra. Prova amb:

```
plot(m)
```

Observeu com les categories de la variable qualitativa s'han codificat numèricament en el gràfic. Pensant només en mostres bivariants, tenim 3 situacions possibles on els gràfics **poden ser molt informatius sobre la relació de dependència entre les variables que conformen la mostra**:

- **Qualitativa vs qualitativa:** El millor gràfic és el diagrama de barres creuat (un per cada variable).

```
barplot( table(m[c(1,2)]), legend=T ) # només fa falta
barplot( table(m[c(2,1)]), legend=T ) # u dels dos
```

Caldria ajustar la llegenda per a obtenir un resultat bonic, i també es podria afegir un `box()` per emmarcar el gràfic. La independència entre les variables s'aprecia quan la composició percentual de cada barra és "similar". En el cas contrari hi hauria indicis de dependència entre les dues variables.

- **Qualitativa vs quantitativa:** Es presenta, per a cada nivell de la variable qualitativa, un gràfic de la quantitativa associat. Hi ha dues opcions:

  - **Diagrama de punts** (si hi ha poques dades)

    ```
    stripchart(m$notafinal ~ m$sexe)
    ```

    Fa una comparativa de diagrames de punts de la variable 'notafinal' agrupats segons els valors la variable 'sexe'.

  - **Diagrama de caixa** (si hi ha massa dades per fer un diagrama de punts)

    ```
    boxplot(notafinal ~ sexe, data=m)
    ```

    Fa una comparativa de diagrames de caixa de la variable 'notafinal' agrupats segons els valors la variable 'sexe'.

  En ambdós casos, la independència entre les variables s'aprecia quan la distribució de punts (o mida de caixa i bigots) és "similar" en tots els casos. Si no, hi hauria indicis de dependència entre les dues variables.

- **Quantitativa vs quantitativa:** Cada dada bivariant forma un punt en el pla X-Y, i es forma un núvol de punts amb tota la mostra.

  ```
  plot( m[c(3,4)] )
  ```

  Fa un nuvol de punts amb les variables 'nivelmat' i 'notafinal'. La independència entre les variables s'aprecia quan el núvol de punts no té cap forma definida. En el cas contrari hi hauria indicis de dependència entre les dues variables.

Obtenir el gràfic és senzill, però hi ha una sèrie d'arguments que R calcula per defecte i que no sempre són els desitjats, com els límits i etiquetes dels eixos (`xlim`, `ylim`, `xlab`, `ylab`), colors de les barres i punts i forma d'aquests (`col`, `pch`), títol del gràfic (`main`, `sub`), etc. Podeu consultar l'ajuda per a aquests casos per a obtenir uns resultats més estètics.

Estadístics

Els estadístics conjunts que es poden calcular amb mostres bivariants quantitatives són:

- Covariància mostral: `cov( m[c(3,4)] )`

  Ens dóna les covariàncies entre cada parella de variables en forma de matriu. La covariància entre una variable i si mateixa es diu més pròpiament variància.

- Coeficient de correlació lineal mostral: `cor( m[c(3,4)] )`

  Ens dóna els coeficients de correlació entre cada parella de variables en forma de matriu. El coeficient de correlació entre una variable i si mateixa sempre val 1.000.

Es poden calcular la resta d'estadístics univariants (p. 63) sobre cada variable quantitativa sencera, o seleccionant només els valors segons criteris que impliquen els valors de les altres variables, segons interesse.

Exercicis d'ensinistrament

Usa la mostra de dades que figura a l'arxiu `s5-descriptiva-2v-dades.txt` i emmagatzemada a la variable `m`.

1. Considerant les variables SEXE i SISOPER:

   (a) Mostra en una taula de freqüències conjuntes la distribució dels individus.

   (b) Emet un judici sobre el grau d'independència entre les dues variables de forma intuïtiva, ajudant-te d'algun gràfic (teoria).

   (c) Quin percentatge dels individus de la mostra són homes que usen MacOS? Sol.: 12.5%

   (d) Quin percentatge dels homes de la mostra usa MacOS? Sol.: 22.44%

2. Considerant les variables SEXE i NIVELMAT:

   (a) Mostra gràficament l'efecte del sexe sobre la distribució de puntuacions a la prova inicial de nivell matemàtic, amb tres diagrames de punts que es puguen comparar.

   (b) Mostra gràficament l'efecte del sexe sobre la distribució de puntuacions a la prova inicial de nivell matemàtic, amb diagrames de caixa que es puguen comparar.

   (c) Consideres que el sexe influencia substancialment la distribució de NIVELMAT? (Sí o no)

   (d) Quin sexe té un major nivell... ...(i) segons el diagrama de caixa? ...(ii) segons algun estadístic convenient?

   (e) Quin sexe és més homogeni (respecte al nivell)... ...(i) segons el diagrama de caixa? ...(ii) segons algun estadístic convenient?

3. Considerant les variables NIVELMAT i NOTAFINAL:

(a) A priori, penses que les dues variables haurien de tenir relació? En cas afirmatiu, quina seria la variable independent o explicativa, i quina la variable dependent o explicada?

(b) Representeu gràficament les dades segons la lògica de l'apartat anterior, i contrasta si aquesta mostra incideix en l'apreciació que has fet de l'apartat anterior.

(c) Calculeu la covariància i el coeficient de correlació lineal entre les variables. Sol.: $s_{XY} = 5.175220$, $r_{XY} = 0.890512$

4.7 Pràctica R: 6. Recta de regressió

Objectius

A l'estudi conjunt de dues variables quantitatives és possible trobar indicis que una de les variables (X, independent) té una influència sobre l'altra (Y, dependent). El coneixement a priori de l'investigador, junt a un gràfic que mostre una tendència clara, són elements suficients per investigar la manera concreta (numèrica) en què se relacionen les variables.

La forma concreta que relaciona dues variables és una funció matemàtica $Y = f(X)$ senzilla que demostre que les dades y_i de la mostra es puguen aproximar bé amb els valors $f(x_i)$. I el cas més senzill de funció $f(X)$ és la funció lineal, és a dir $Y = a + bX$ per a algun valor concret de a i b, que és la tasca que realitzem en aquesta pràctica.

Les dades

Les dades estan emmagatzemades en dues columnes d'un full de dades. És molt important saber quina és la variable independent (o de control) i quina és la variable dependent (o de resposta).

En aquest cas treballarem amb les dades que figuren a l'arxiu `s6-regressio-dades-1.txt`.

Diagrama de dispersió

El diagrama de dispersió (o núvol de punts) és el gràfic que informa l'investigador de la possible relació entre les variables i la seua forma.

De vegades, les dades estan expressades en una escala en la qual no mostren una relació lineal clara. És possible que en altres escales sí es puga apreciar la relació lineal. Per això és recomanable crear variables transformades de les originals. Per exemple, la transformació logarítmica o l'arrel quadrada (per dades positives) són prou habituals:

```
mostra <- read.table(file='s6-regressio-dades-1.txt',
                     header=T)
plot(mostra)
# mirar les variables amb relacio sospitosa
```

```
# x1 vs x3: relacionades, no sembla lineal
# x2 vs x4: relacionades, no sembla lineal
# x2 vs x5: relacionades, si sembla lineal
# x4 vs x5: relacionades, no sembla lineal
plot( mostra$x1, log(mostra$x3) )
plot( log(mostra$x2), log(mostra$x4) )
# entre x4 i x5 no es troba la transformacio
```

Després de la inspecció dels gràfics s'haurien d'usar les dades més convenients, siguen les originals o les resultants de fer una transformació, per fer l'anàlisi de regressió lineal.

La funció `lm()`

La funció `lm()` (de *Linear Models*) està ja programada per a l'estudi de la regressió lineal. Una sintaxi prou completa és:

```
lm(formula, data, subset)
```

on els arguments indiquen:

- `data`: (opcional) el full de dades que conté les columnes a analitzar. Potser les dades estan emmagatzemades en dos vectors que no formen un full de dades.
- `formula`: expressió que indica la forma de dependència. Normalment és `resposta ~ control`, on `resposta` simbolitza la variable dependent o de resposta, i `control` simbolitza la variable independent o de control, i són dues columnes del full de dades `data`.
- `subset`: (opcional) vector especificant un subconjunt d'observacions, si no interessen totes les que hi ha a la mostra.

L'objecte tornat per la funció és de tipus llista (`list`, no estudiat en aquest manual) que té una sèrie de components a les quals es pot accedir amb l'operador `$`. Les components més usades són:

- `coefficients`: vector amb els coeficients a (Intercept) i b (Slope) de la recta de regressió $Y = a + bX$ (Resposta $= a +$ bControl)
- `residuals`: vector amb els valors dels residus (és a dir, les diferències entre els valors reals de la variable resposta i els valors atribuïts per la recta de regressió).

Per tant, si posem el resultat de la funció `lm()` a la variable `reg`, aleshores, podem recuperar els coeficients de la recta de regressió i els residus amb, per exemple:

```
reg <- lm( formula=x5 ~ x2, data=mostra)
reg$coefficients   # els coef. a i b de la recta
reg$residuals      # els errors de la recta
```

Gràfic de la recta de regressió

Una volta feta l'anàlisi, és molt informativa la representació gràfica simultània de les dades i la recta de regressió. A l'exemple que estem mostrant:

```
reg <- lm(x5 ~ x2, mostra)
recta.x <- seq(from=min(mostra$x2),
               to=max(mostra$x2), length=100)
recta.y <- reg$coeff[1] + reg$coeff[2]*recta.x
plot(mostra[c('x2', 'x5')])
points(x=recta.x, y=recta.y, type='l')
```

Bondat de l'ajustament

L'avaluació sobre la bondat de l'ajustament de les dades a la funció de regressió es pot fer: (1) visualment, mirant els gràfics de les dades i de la funció superposats, i (2) calculant un estadístic interpretable que descriga aquesta característica.

L'estadístic R^2, que en el cas de la regressió és $R^2 = \mathrm{r}^2_{XY}$ (o $R^2 = \mathrm{r}^2_{XY} \times 100\%$ si s'expressa en tant per cent), serveix per avaluar la bondat de l'ajustament. Quan més pròxim a 1.00 (o a 100%), millor és la qualitat de l'ajustament de les dades a la recta de regressió. Així, amb:

```
cor(mostra)^2*100
```

podem comprovar que la R^2 entre $x2$ i $x5$ és molt alta.

Exercicis d'ensinistrament

Usant les dades de l'arxiu `s6-regressio-dades-2.txt` contesta les següents preguntes:

1. Considerant les variables NIVELMAT i NOTAFINAL:
   (a) Representeu gràficament les dades i raoneu si el nivell matemàtic inicial sembla influir en la nota final de l'assignatura en aquesta mostra.
   (b) Quin percentatge de persones de la mostra ha "aprovat" les dues proves inicial i final? Sol.: 53.97%
   (c) Calculeu la fórmula de la recta de regressió que calcularia la nota final a partir del nivell inicial. Sol.: NOTAFINAL $= -1.810 + 1.372 * \text{NIVELMAT}$
   (d) Dibuixeu l'esmentada recta acompanyada de les dades.
   (e) Emeteu un judici sobre la qualitat de la regressió feta ajudant-vos dels càlculs que siguen necessaris. Sol.: $R^2 = 0.793$ (o 79.30%)
   (f) Quina nota es pronosticaria per a un alumne amb un valor de NIVELMAT igual a 7.5 i quina seria la seua fiabilitat? Sol.: 8.48 amb un 79.30% de fiabilitat.

PART III
POBLACIONS DE DADES (MODELS DE PROBABILITAT)

Capítol 5

Probabilitats

5.1 Experiments aleatoris

Definició 5.1.1 (Experiment) *Un **experiment** és un procès que, a partir d'unes condicions inicials, dóna lloc a un resultat objectiu observable. Es poden considerar de dos tipus:*

- **Determinista:** Les condicions inicials són repetibles i donen lloc a un mateix resultat a cada repetició.
- **Aleatori:** Les condicions inicials són difícilment repetibles i la repetició pot donar lloc, eventualment, a diferents resultats.

Per tant, tot experiment aleatori ve descrit per un conjunt de resultats possibles (anomenat **espai mostral**) i la incertesa de saber quin resultat eixirà la pròxima vegada que es realitze.

5.1.1 Resultat i esdeveniment

Dins d'un conjunt com E es poden considerar subconjunts, des del més trivial, que no té cap element, i s'anomena conjunt buit (representat amb $\varnothing$) fins a un altre, també trivial i extrem, que té tots els elements, és a dir, el mateix E.

Amb els subconjunts de E es poden fer 3 operacions bàsiques:

- **Unió:** Donats dos conjunts A i B, es representa amb $A \cup B$ (A unió B) el conjunt que reuneix tots els elements existents siga en A siga en B (siga en ambdós).
- **Intersecció:** Donats dos conjunts A i B, es representa amb $A \cap B$ (A intersecció B) el conjunt que reuneix tots els elements existents simultàniament en A i en B.
- **Complementari:** Donat un conjunt A, subconjunt de E, es representa amb $\overline{A}$ (complementari de A, en E) el conjunt que reuneix tots els elements de E que no pertanyen a A.

Amb aquesta introducció podem passar a la següent definició.

Definició 5.1.2 (Espai d'esdeveniments i esdeveniment) *Donat un espai mostral E, un espai d'esdeveniments de E és una col·lecció de subconjunts de E que:*

1. *Té al conjunt E com a membre.*
2. *Sempre que té un membre A, té també com a membre el seu complementari $\overline{A}$.*
3. *Sempre que té dos membres A i B, té també com a membre la unió $A \cup B$.*

*Un **esdeveniment** és un membre de l'espai d'esdeveniments.*

Exemple 5.1.1 *Els resultats dels partits de futbol són un exemple típic d'experiment aleatori. Considerem un partit **València vs Vila-real**.*

- *L'espai mostral inclouria tots els resultats intel·lectualment possibles, és a dir,*
$$E = \{(0,0), (1,0), (0,1), (2,0), \ldots, (7,5), \ldots\}$$
(cada parella de valors indica el nombre de gols marcat per l'equip mencionat en el lloc respectiu).
- *Un exemple de resultat és:* $(3,3)$
- *Un exemple d'esdeveniment és:* $\{(0,0), (1,1), (2,2), (3,3), (4,4), \ldots\}$
- *Un altre exemple de resultat:* $(1,2)$
- *I un altre exemple d'esdeveniment és:* $\{(3,0), (0,5)\}$

Al potencial usuari de la teoria que modelitza l'incertesa l'interesa que l'objecte de treball siguen els esdeveniments, més que els resultats, perquè en ocasions són diversos el resultats convenients.

Els esdeveniments se solen denotar amb lletres majúscules i es poden expressar de dues formes:

1. **Per extensió**: donant la llista de resultats. Per exemple:
   $A = \{(0,0)\}$
   $B = \{(3,0), (2,1), (1,2), (0,3)\}$
   $C =$

2. **Per comprensió**: donant una proposició que fa al·lusió als resultats d'interès. Per exemple:
   $A =$ "empat sense gols"
   $B =$
   $C =$ "guanya el Vila-Real"

Exercici 5.1.1 *Completa els esdeveniments anteriors incomplets, tenint en compte que són els mateixos en cada llista.*

Amb els esdeveniments, com són conjunts, es poden fer unions ($A \cup B =$ "A" o "B" en la versió "per comprensió"), interseccions ($A \cap B =$ "A" i "B") i complementari ($\overline{A} =$ no "A"), entre altres operacions.

Quan es va a realitzar un experiment aleatori, i s'observa un esdeveniment concret A, es diu que l'esdeveniment A ocorre, quan el resultat de l'experiment és un element de A, i que A no ocorre en el cas contrari.

Exemple 5.1.2 *Continuant amb l'Exemple 5.1.1 i considerant l'esdeveniment* $D =$ *"empatar"* $= \{(0,0),(1,1),(2,2),\ldots\}$*, si el resultat del partit és* $(3,3)$ *direm que D ha ocorregut, i si el resultat fóra* $(1,2)$ *diríem que D no ha occoregut.*

Hi ha dos esdeveniments especials o trivials:

- L'esdeveniment de tots els resultats possibles: és el propi espai mostral i se sol denotar amb la lletra E. S'anomena també **esdeveniment segur**, perquè ocorre amb tota seguretat.
- L'esdeveniment de no cap resultat: es diu conjunt buit i es representa amb el signe $\varnothing$. S'anomena també **esdeveniment impossible**, perquè no ocorre mai (l'experiment dóna sempre algun dels resultats de E).

Donat que alguns esdeveniments poden ser resultat d'unions, interseccions i/o complementaris d'altres esdeveniments, és interessant conéixer algunes propietats bàsiques d'aquests operadors ($\cup, \cap$ i $\overline{\ }$).

- $A \cup B = B \cup A$, $A \cap B = B \cap A$, $A \cup \varnothing = A$
- $A \cup \overline{A} = E$, $A \cap \overline{A} = \varnothing$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

5.1.2 Freqüència relativa a llarg termini vs probabilitat subjectiva

Una manera de quantificar l'incertesa d'un esdeveniment particular d'un experiment aleatori és repetir l'experiment indefinidament i observar en quina proporció ocorre l'esmentat esdeveniment.

Exemple 5.1.3 *A Internet es pot trobar la informació sobre el signe (1='victòria', X='empat' i 2='derrota') dels partits "València CF - Reial Madrid" desde l'inici de la competició de lliga espanyola. Amb les dades es poden fer la taula i el gràfic d'evolució temporal de proporció de cada signe (vegeu la Taula 5.1 i Figura 5.1).*

A partir d'un estudi d'aquest tipus, una manera d'assignar graus de certesa a resultats és assignar els valors de proporció obtinguts. La filosofia és: si ha ocorregut molt en el passat, és raonable que ocórrega molt en el futur. Aquesta versió s'anomena **freqüencialista**.

Taula 5.1: Evolució temporal de la proporció de cada signe del partit. A cada partit es calcula, de la mostra de signes acumulats fins a aqueix partit, la proporció de cada signe (llegir la taula columna per columna)

Partit	1	2	3	4	5	$\cdots$
Resultat	2	X	1	1	2	$\cdots$
Prop. acum. d'1	$\frac{0}{1}=0$	$\frac{0}{2}=0.0$	$\frac{1}{3}=0.3333$	$\frac{2}{4}=0.5$	$\frac{2}{5}=0.4$	$\cdots$
Prop. acum. d'X	$\frac{0}{1}=0$	$\frac{1}{2}=0.5$	$\frac{1}{3}=0.3333$	$\frac{1}{4}=0.25$	$\frac{1}{5}=0.2$	$\cdots$
Prop. acum. de 2	$\frac{1}{1}=1$	$\frac{1}{2}=0.5$	$\frac{1}{3}=0.3333$	$\frac{1}{4}=0.25$	$\frac{2}{5}=0.4$	$\cdots$

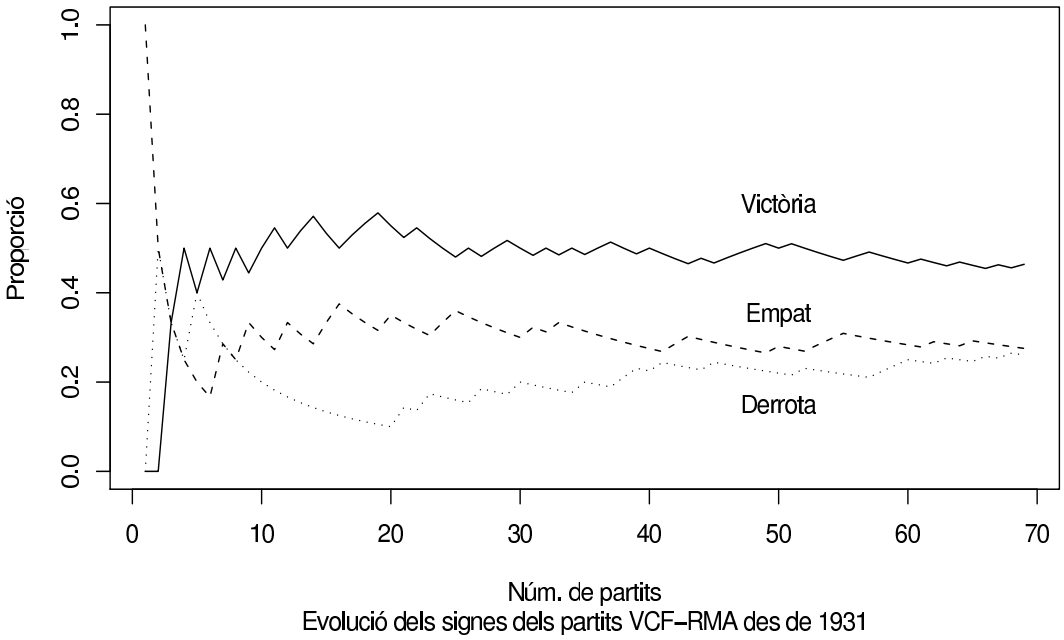


Figura 5.1: Evolució de la proporció de cada signe (1='victòria', X='empat' i 2='derrota') desde l'inici de la competició lliguera espanyola

Observant la Figura 5.1 podriem assignar unes probabilitats de 0.5 a la victòria i 0.25 a derrota i empat (o altres valors molt similars inspirats en la mateixa gràfica).

$$\text{Probabilitat} \equiv \% \text{ (o proporció) a llarg termini}$$

L'alternativa a aquesta metodologia és assignar **subjectivament** els graus de certesa dels resultats possibles. Un argument a favor d'aquesta versió és la impossibilitat de repetir l'experiment un gran nombre d'ocasions. Això sí, si s'assignen graus de certesa als resultats posibles, no es pot fer amb absoluta llibertat, sinó que és important que es complisquen les mateixes restriccions que quan s'assignen amb la versió freqüencialista.

5.2 Probabilitat

La definició matemàtica de probabilitat respecta les dues possibles filosofies de la secció anterior. Només adopta com axiomes les propietats que, intuïtivament, una assignació de probabilitats ha de respectar.

5.2.1 Definició axiomàtica i propietats

Definició 5.2.1 (Probabilitat) *Una probabilitat P és un criteri que associa un valor numèric a cada esdeveniment, de manera que:*

1. $P(A) \geq 0$*, per a qualsevol esdeveniment* A*.*
2. $P(E) = 1$ *on* E*, és l'esdeveniment de tot l'espai mostral.*
3. $P(A \cup B) = P(A) + P(B)$*, si* A *i* B *són esdeveniments disjunts (és a dir,* $A \cap B = \varnothing$*, no tenen resultats coincidents).*

Usant raonaments lògics es pot demostrar la següent llista de propietats bàsiques.

Propietat 5.2.1 *Si* P *és una probabilitat definida seguint la Definició 5.2.1, aleshores:*

1. $P(\varnothing) = 0$*.*
2. $P(\overline{A}) = 1 - P(A)$ *per a qualsevol esdeveniment* A*.*
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ *per a esdeveniments* A *i* B *qualsevol.*

5.2.2 Equiprobabilitat

Pensem en un experiment amb espai mostral $E = \{r_1, r_2, \ldots, r_n\}$ on tots els resultats són intercanviables en el sentit que l'experiment és el mateix si permutem els resultats.

Exemple 5.2.1 *A l'experiment de llançar un dau (sacsejant-lo prou) i observar la cara mostrada en caure sobre la taula, si disposem d'un dau de* 6 *cares perfectament construït, la posició inicial del dau no sembla afectar ni la mecànica de l'experiment ni les possibilitats dels resultats.*

A l'experiment de barrejar conscienciosament una baralla de cartes i extraure una carta a l'atzar, la configuració inicial de la baralla no sembla afectar ni la mecànica de l'experiment ni les possibilitats dels resultats.

En aquests exemples, una permutació en l'ordre dels resultats no afecta la natura de l'experiment, que continua sent el mateix.

La lògica indica que l'*incertesa* hauria de ser la mateixa per a tots els resultats.

Per tant, una probabilitat P lògica seria aquella que:

- $P(\{r_i\}) = \dfrac{1}{n}$, per qualsevol resultat r_i.
- $P(A) = \dfrac{\text{Nombre d'elements de } A}{\text{Nombre d'elements de } E}$, per qualsevol esdeveniment A.

En aquests tipus d'experiments és fonamental saber comptar elements de conjunts, i hi ha ocasions en què les **tècniques combinatòries** són de gran ajuda.

Exercici 5.2.1 *Una urna té una bola blanca i una altra negra. Anem a fer l'experiment de traure una bola, anotar el seu color (B o N) i tornar-la a l'urna, i repertir-ho 7 vegades, obtenint una llista ordenada de 7 lletres. Calculeu les probabilitats de:*

(a) Obtenir la llista BBNNNNN
(b) Obtenir la llista BNNNNNB
(c) Obtenir una llista qualsevol amb 2 boles blanques i 5 negres

Exercici 5.2.2 *Calculeu la probabilitat d'obtenir, en un sorteig de l'ONCE:*

(a) les 5 xifres... *(b) les 4 últimes xifres...*
(c) les 3 últimes xifres... *(d) les 2 últimes xifres...*
(e) l' última xifra...

...del número premiat.

Exercici 5.2.3 *Si tots els resultats del futbol foren equiprobables, calculeu la probabilitat d'obtenir:*

(a) 14 encerts *(b) 13 encerts* *(c) 12 encerts* *(c) 11 encerts*

Exercici 5.2.4 *Calculeu la probabilitat d'obtenir, a la loteria primitiva:*

(a) 6 encerts *(b) 5 encerts més complementari*
(c) 4 encerts *(d) 3 encerts*

Exercici 5.2.5 *Suposant que les dates de naixement foren absolutament atzaroses al llarg de l'any (cosa que no és veritat, però que passarem per alt), calculeu la probabilitat que, a una classe amb* $\boxed{}$ *alumnes, hi haja alguna (una o més) coincidència en la data de naixement.*

5.2.3 Probabilitat condicionada i independència

Les probabilitats inicials dels esdeveniments es modifiquen (s'actualitzen) quan es té una informació parcial sobre el resultat de l'experiment aleatori.

Exemple 5.2.2 *D'un grup d'alumnes dels quals sabem les dades sobre el seu sexe i el sistema operatiu de preferència (vegeu la Taula 5.2), l'experiment consisteix en triar-ne completament a l'atzar un d'aquests, i observar el seu sexe i sistema operatiu.*

Taula 5.2: Dades de l'Exemple 5.2.2

	SISOPER		
SEXE	Win	Lin	Mac
Home	78	45	13
Dona	29	23	5

Aleshores, si denotem l'esdeveniment que la persona triada pertanga a cada sexe o sistema operatiu usant la lletra inicial corresponent:

- *Calculeu* $P(H) = \frac{\quad}{\quad} =$
- *Calculeu* $P(H \text{ sabent } L) = \frac{\quad}{\quad} =$
- *Calculeu* $\dfrac{P(H \cap L)}{P(L)} = \frac{\quad}{\quad} =$
- *Comprova que* $P(H \text{ sabent } L) = \dfrac{P(H \cap L)}{P(L)}$.

Aquest exemple motiva la lògica de la següent definició.

Definició 5.2.2 (Probabilitat condicionada) *Si P és una probabilitat i A és un esdeveniment amb $P(A) > 0$, aleshores es defineix una nova probabilitat, condicionada a A, que es denota per $P(\cdot|A)$, i que per qualsevol esdeveniment B val:*

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

L'expressió $P(B|A)$ pot llegir-se com "probabilitat que ocòrrega l'esdeveniment B sabent que ha ocorregut l'esdeveniment A".

Exercici 5.2.6 *Tenim 50 memòries RAM: 15 de la marca XXX i 35 de la marca YYY, i sabem que de cada marca n'hi ha 3 de defectuoses. Estan totes barrejades a la mateixa caixa. Si triem una a l'atzar, sense mirar...*

- *...i no es pot identificar de quina marca és, ni si funciona correctament o no, calculeu separadament la probabilitat que siga de la marca XXX i la probabilitat que siga de la marca YYY.*

- *...i no es pot identificar de quina marca és, ni si funciona correctament o no, calculeu separadament la probabilitat que siga defectuosa i la probabilitat que funcione correctament.*

- *...i es pot identificar en la placa que és de la marca XXX, però no pots comprovar si funciona correctament, calculeu la probabilitat que siga defectuosa.*

- *...i no es pot identificar de quina marca és, però en comprovar-la resulta defectuosa, calculeu la probabilitat que siga de la marca XXX si saps que és defectuosa.*

En moltes ocasions és important conéixer si certa informació sobre el resultat d'un experiment aleatori modifica les probabilitats que es tenen sobre esdeveniments del nostre interès. Quan això no ocorre, es té el fenomen de la independència.

Definició 5.2.3 (Independència d'esdeveniments) *Un esdeveniment B és independent d'un altre A si la probabilitat (inicial) de B no varia en condicionar-la a A, és a dir:*

$$P(B) = P(B|A)$$

A partir d'aquesta definició, es pot demostrar que la independència és un fenomen recíproc, és a dir, que si B és independent de A, aleshores A ho és també de B. A més es pot demostrar que:

Propietat 5.2.2 *A i B són esdeveniments independents si i només si:*

$$P(A \cap B) = P(A)P(B)$$

5.2.4 Teoremes de la Probabilitat i de Bayes

Gràcies a la formalització del concepte de probabilitat condicional, es poden calcular les probabilitats d'esdeveniments en experiments que tenen una estructura seqüencial, i on es coneixen de manera natural les probabilitats condicionades, però no les probabilitats "tal qual" d'alguns esdeveniments.

Teorema 5.2.1 (Probabilitat total) *Si P és una probabilitat i E l'espai mostral, que es pot particionar en dos esdeveniments A_1 i A_2 (és a dir, $E = A_1 \cup A_2$ on $A_1 \cap A_2 = \varnothing$) amb $P(A_1) > 0$ i $P(A_2) > 0$.*

Si per a qualsevol esdeveniment B només es coneixen les probabilitats condicionades $P(B|A_1)$ i $P(B|A_2)$, aleshores la probabilitat de B es pot calcular com:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2)$$

Aquest resultat és fàcilment generalitzable a una quantitat qualsevol d'esdeveniments $A_1, A_2, A_3, \ldots$ que formen una partició de E (és a dir, que siguen disjunts per parelles i la unió dels quals done l'espai mostral E).

Exercici 5.2.7 *Suposem que el 30% de les memòries RAM que tenim són de la marca XXX i que la resta són de la marca YYY. Si sabem que són defectuoses l'1% de les memòries de marca XXX i el 2% de les de marca YYY, i s'agafa una memòria a l'atzar, calculeu la probabilitat que siga defectuosa.*

Als experiments seqüencials, o en experiments on es coneixen de manera natural certes probabilitats condicionades, hi ha ocasions on és important conéixer les probabilitats condicionades "contràries" (és a dir, on està canviat l'ordre dels esdeveniments). Una observació senzilla porta al següent resultat, importantíssim en la teoria de les Probabilitats.

Teorema 5.2.2 (Bayes) *Si P és una probabilitat i E l'espai mostral, que es pot particionar en dos esdeveniments A_1 i A_2 (és a dir, $E = A_1 \cup A_2$ on $A_1 \cap A_2 = \varnothing$) amb $P(A_1) > 0$ i $P(A_2) > 0$.*

Si a més se sap que ha ocorregut l'esdeveniment B, les probabilitats de A_1 i A_2 queden actualitzades com:

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)}$$
$$P(A_2|B) = \frac{P(B|A_2)P(A_2)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)}$$

Com ocorre amb el Teorema 5.2.1, aquest resultat també és vàlid per a qualsevol quantitat d'esdeveniments $A_1, A_2, A_3, \ldots$

Exercici 5.2.8 *Si s'agafa una memòria RAM (de les usades a l'Exercici 5.2.7) a l'atzar i es comprova que és defectuosa, però no es pot identificar la marca, calculeu la probabilitat que siga de cadascuna de les marques possibles. Com s'han actualitzat les probabilitats en saber que era defectuosa? Podries intuir abans de fer els càlculs si les probabilitats s'actualitzen a l'alça o a la baixa?*

El Teorema de Bayes és molt important i s'utilitza en la base de moltes tècniques, com per exemple en l'anàlisi d'imatges. Quan es dissenya un mètode de reconeixement d'imatges, d'una banda, aquest es pot avaluar al laboratori, observant l'efectivitat del mètode usant una bateria d'imatges i comptabilitzant la taxa d'èxits (és a dir, calculant les probabilitats de les imatges captades condicionades a les imatges originals). Quan el dispositiu funciona autònomament, i no controlem les imatges originals, el dispositiu ha d'observar-les i classificar-les. Aleshores el que es controla és la imatge captada, però no l'original, i per tant, conéixer les probabilitats de les imatges originals condicionades a les imatges captades és fonamental per a la presa de decisions. Usem com exemple concret i molt senzill l'exercici següent.

Exercici 5.2.9 *Un sensor d'imatge es fabrica per a reconéixer automàticament els colors del semàfor. En fer-li proves al laboratori, s'ha vist que, quan la llum del semàfor estava en verd, el sensor ho detectava com groc en un 5% de les ocasions i com roig en un 1%. Quan estava en groc, el sensor ho detectava com verd en un 8% de les ocasions i com roig en el 30% de les ocasions,*

mentre que quan estava en roig, el sensor ho detectava com verd en un 10% de les ocasions i com groc en un 25% de les ocasions.

Si aquest sensor es col·loca a un cotxe automàtic, que es fa circular per un circuit i, en un moment donat, es troba amb un semàfor, i el sensor detecta que està en verd, quina és la probabilitat que el semàfor estiga realment en cadascun dels colors que pot estar? (suposem que el semàfor és tal que, en el moment que el sensor detecta la imatge, les probabilitats d'estar realment a cada color són 0.6 (verd), 0.02 (groc) i 0.38 (roig).

5.3 Exercicis proposats

Exercici 5.3.1 *Una empresa d'enquestes contractada pel gremi d'acadèmies d'una ciutat demana als opositors, en eixir de la prova, si han preparat l'oposició en una acadèmia o no. Tots els opositors contesten amb sinceritat. Després se'ls demanen les dades personals (nom i cognoms) per poder creuar les dades amb les puntuacions finals i estudiar conjuntament el resultat de l'oposició i l'ús d'acadèmies. A aquesta part no contesten tots. De fet, les dades arreplegades són les següents:*

- *Usa acadèmia? Sí (94), no (72) (tots els opositors han contestat amb sinceritat)*
- *Relació 'Ús acadèmia' vs 'Resultat oposició' (no tots els opositors han contestat)*

	Aprova opos.	*No aprova opos.*
Acadèmia	*5*	*17*
No acadèmia	*2*	*10*

Usant les dades del quadre per obtenir les probabilitats d'aprovar (o no) condicionades a anar a l'acadèmia (o no), calculeu la probabilitat que, triat un opositor dels que ha aprovat l'oposició (independentment de si va contestar a les dues preguntes o no), aquest haja sigut client d'alguna acadèmia.

Exercici 5.3.2 *Si l'estat del tràfic de la xarxa és fluit, un missatge tarda menys d'un segon a arribar al destinatari en el 95% de les ocasions. Si l'estat és dens, aleshores açò ocorre només el 15% de les ocasions.*

Si l'estat de la xarxa és fluit durant el 55% de la jornada i dens la resta del temps:

1. *Quina probabilitat tenim que un missatge que anem a enviar arribe en menys d'un segon al destinatari?*
2. *Si m'envie un missatge a mi mateix i comprove que tarda a arribar-me més d'un segon, quina és la probabilitat que la xarxa estiga fluida? (*Nota: si no has fet l'apartat anterior i et fa falta la seua solució, agafa com a solució el valor 0.75*)*

Exercici 5.3.3 *Un dispositiu té una peça al seu interior, de manera que el 80% dels dispositius que tenen la peça defectuosa no funcionen, mentre que només el 10% dels dispositius que tenen la peça correcta són defectuosos.*

Suposem que es coneix que el 5% de les peces internes fabricades són defectuoses. Aleshores, si adquirim un dispositiu d'aquests:

1. *Quina probabilitat tenim que funcione correctament, abans de comprovar-ho?*

2. *Si comprovem que el dispositiu NO funciona, quina és la probabilitat que la peça interna siga defectuosa? (*Nota: si no has fet l'apartat anterior i et fa falta la seua solució, agafa com a solució el valor 0.97*)*

Exercici 5.3.4 *Cada dia, una alarma falla (i* ***sona sense haver-hi perill****) amb probabilitat de* 0.05 *mentre que falla (i* ***no sona havent-hi perill****) amb probabilitat* 0.001*. S'estima que, cada dia, la probabilitat que s'ataque el lloc protegit per l'alarma, és de* 0.025*. Quina és la probabilitat que, en rebre's l'avís d'alarma en la central, no hi haja un perill real?*

Exercici 5.3.5 *Jurassic Petroleum ha classificat els sòls en tres tipus (A, B i C), segons les possibilitats de descobrir-hi petroli. La companyia perfora un pou en un lloc, que té probabilitats 0.35, 0.55 i 0.10 de pertànyer a cadascun dels tres tipus de sòl, respectivament. D'acord amb l'experiència, hi ha petroli en un 40% de perforacions en sòl A, en un 25% de perforacions en sòl B i en un 30% de perforacions en sòl C.*

Si no hi ha petroli al pou perforat, quina és la probabilitat que el pou es trobe en un sòl B?

Exercici 5.3.6 *L'administració d'un país colonitzat fa un referèndum sobre la instal·lació de míssils estrangers al seu territori. A l'eixida d'un col·legi electoral, una empresa de sondeigs demana (anònimament) als votants què han votat. Se sap d'altres referèndums semblants que si un votant ha votat "no", contesta la veritat en un 40% dels casos, mentre que si ha votat "sí", contesta la veritat en un 95% dels casos.*

En fer l'escrutini del col·legi, es veu que hi ha un 70% de vots negatius.

- *Si un votant ha contestat "no" a l'empresa de sondeigs, quina és la probabilitat que haja votat "no"?*

- *Quina és la probabilitat que almenys 3 de les primeres 10 paperetes escrutades siguen afirmatives?*

Exercici 5.3.7 *En un sorteig de l'ONCE es tria un número qualsevol del 00000 al 99999 (a banda d'un número de sèrie que ara no ens interessa). Si compres un únic número, calculeu la probabilitat d'obtenir:*

(a) les 5 xifres, (b) les 4 últimes xifres, (c) les 3 últimes xifres, (d) les 2 últimes xifres, (e) l'última xifra

del número premiat.

Exercici 5.3.8 *Una travessa es pot veure com una llista de 14 signes del tipus 1, X, 2, on l'ordre dels signes segueix els dels partits de la setmana. Si cada partit pot acabar en qualsevol signe amb la mateixa probabilitat, i fem una travessa senzilla, calculeu la probabilitat d'obtenir:*

(a) 15 encerts, (b) 14 encerts, (c) 13 encerts, (d) 12 encerts

Exercici 5.3.9 *Calculeu la probabilitat d'obtenir, a la loteria primitiva:*

(a) 6 encerts, (b) 5 encerts més complementari, (c) 4 encerts, (d) 3 encerts

Exercici 5.3.10 *Calculeu la probabilitat que, en una classe amb 40 alumnes, hi haja almenys 2 persones amb la mateixa data de naixement. Encara que no és veritat, suposarem que la gent naix un dia qualsevol de l'any totalment a l'atzar (i que els anys tots tenen 365 dies, que tampoc és cert).*

Exercici 5.3.11 *Un missatge es transmet usant el codi binari de 0's i 1's. Cada bit transmés (0 o 1) ha de passar per tres "relays" abans d'arribar al receptor. A cada "relay" el bit pot sofrir una inversió amb probabilitat 0.20. Assumint que els relays funcionen de manera independent:*
Transmissor —> Relay 1 —> Relay 2 —> Relay 3 —> Receptor

- *Si s'emet un 1 des del transmissor, quina és la probabilitat que aqueix 1 siga transmés correctament pels tres "relays"?*
- *Si s'emet un 1 des del transmissor, quina és la probabilitat que arribe un 1 al receptor?*
- *Suposem que per norma general el 70% dels bits emesos són 1. Si el receptor rep un 1, quina és la probabilitat que haja sigut un 1 realment el bit enviat pel transmissor?*

Exercici 5.3.12 *En la ciència forense, la probabilitat que dues persones coincidisquen en una característica (color de cabell, tipus de sang, etc.) s'anomena "probabilitat de coincidència". Suposa que les freqüències dels fenotipus de sang en la població són:*

A	B	AB	O
0.42	0.10	0.04	0.44

- *Quina és la probabilitat que dues persones triades a l'atzar siguen ambdues de sang tipus A?*
- *Fes el mateix càlcul amb tots els altres tipus de sang.*
- *Troba la probabilitat que dues persones triades a l'atzar tinguen tipus de sang coincident.*
- *La probabilitat que dues persones no coicidisquen en un tret determinat s'anomena "poder discriminant". Quin és el poder discriminant per a la comparació del tipus de sang en l'apartat (c)?*

Exercici 5.3.13 *Un sistema de seguretat s'usa per a detectar els atacs informàtics. El sistema té la característica que dóna l'alarma el 90% de les ocasions on hi ha un atac, però també dóna l'alarma l'1% de les ocasions on no hi ha cap atac. Segons dades de criminalitat actual, la probabilitat que es produisca un atac en un moment donat és 0.001.*

- *Si sona l'alarma, quina és la probabilitat que hi haja un atac real en eixe moment?*
- *Si no sona l'alarma, quina és la probabilitat que hi haja un atac real en eixe moment?*

Un client preocupat, per augmentar la seguretat, pensa a instal·lar dos sistemes que funcionen independentment. Per saber si val la pena gastar-se els diners, demana calcular:

- *Si hi ha un atac real, quina és la probabilitat que algun sistema (!) done l'alarma?*
- *Si no hi ha cap atac real, quina és la probabilitat que algun sistema (!) done l'alarma?*
- *Si no sona cap alarma, quina és la probabilitat que hi haja un atac real en aqueix moment?*
- *Si sona alguna alarma, quina és la probabilitat que hi haja un atac real en aqueix moment?*

Capítol 6

Variable aleatòria

6.1 Definició i tipus

Definició 6.1.1 (Variable aleatòria) *Partint d'un experiment aleatori amb espai mostral E i una probabilitat P definida, una variable aleatòria (real) X és un criteri que associa a cada resultat de l'experiment (element de E) un valor numèric (real, element de $\mathbb{R}$).*

Exemple 6.1.1 *Si llancem 3 monedes tenim*

$$E = \{CCC, CC+, C+C, C++, +CC, +C+, ++C, +++\}.$$

La variable aleatòria X_1 = "nombre de cares" compleix que $X_1 \in \{0, 1, 2, 3\}$

Per altra banda, si un joc consisteix a apostar 1€, guanyant 6€ només si obtens les 3 cares, es podria definir la variable X_2 = "benefici final del joc". En aquest cas $X_2 \in \{-1, +5\}$.

La probabilitat P es trasmet des de l'espai d'esdeveniments de E a l'espai d'esdeveniments de $\mathbb{R}$. De manera natural es defineix la nova probabilitat Pr:

$$\Pr(A) = P(X^{-1}(A))$$

per qualsevol esdeveniment A de $\mathbb{R}$, es té que Pr és una probabilitat. Hi ha dos casos a distingir per a una variable aleatòria X:

- **Discreta** ($X(E)$ és un conjunt discret, tipus $\mathbb{N}$ o $\mathbb{Z}$).

  $$X(E) = \{x_1, x_2, x_3, \ldots\}$$

  Els esdeveniments són tots els subconjunts de $X(E)$

- **Contínua** ($X(E)$ és un conjunt continu, tipus interval, (a, b) o $\mathbb{R}$).

  $$X(E) = (a, b)$$

  (eventualment, podrien ser $a = -\infty$ i/o $b = +\infty$). Ací, els esdeveniments no són tots els subconjunts de $X(E)$, però els que es formen a partir d'unions i complementaris d'intervals oberts i tancats.

Exemple 6.1.2 *Els experiments on es compta la quantitat d'unitats o de vegades que occorre un esdeveniment donen lloc a variables discretes, on l'espai mostral és* $\mathbb{N}$ *o un subconjunt d'aquest.*

Els experiments on es pren una mesura sobre una escala de mesures, com és el temps o les distàncies, donen lloc a variables contínues, on l'espai mostral és $\mathbb{R}$ *o subintervals com* $(0, +\infty)$*. Encara que el temps i l'espai es mesuren amb un nivell de precisió finita (milisegons, micres, etc.), conceptualment poden agafar valors qualsevols de l'interval.*

Si x és un nombre real qualsevol (potser enter), tenim dos esdeveniments molt simples (si no els que més), la manera estàndard de denotar-se i la seua interpretació respecte a l'experiment original (vegeu la Taula 6.1).

Taula 6.1: Donat un valor x qualsevol, els dos esdeveniments més bàsics que estan relacionats amb aquest valor x

Esdeveniment	Notació	Interpretació
$\{x\}$	$\{X = x\}$	Resultats de l'experiment associats al valor x
$(-\infty, x]$	$\{X \leq x\}$	Resultats de l'experiment associats a valors inferiors o iguals a x

Una segona llista d'esdeveniments també molt simples o intuïtius és la que figura a la Taula 6.2 (amb la notació usada i la seua interpretació).

Taula 6.2: Altres esdeveniments bàsics que estan relacionats amb valors x o x_1 i x_2

Esdeveniment	Notació	Interpretació
$(-\infty, x)$	$\{X < x\}$	Resultats de l'experiment associats a valors estrictament inferiors a x
$(x, +\infty)$	$\{X > x\}$	Resultats de l'experiment associats a valors estrictament superiors a x
$(x_1, x_2]$	$\{x_1 < X \leq x_2\}$	Resultats de l'experiment associats a valors entre x_1 i x_2 amb la possibilitat del valor x_2 però no del valor x_1
$[x, +\infty)$	$\{X \geq x\}$	Resultats de l'experiment associats a valors superiors o iguals a x
$(-\infty, x) \cup (x, +\infty)$	$\{X \neq x\}$	Resultats de l'experiment associats a valors diferents a x
etc.	etc.	etc.

La raó de donar-los en segon terme és que aquests esdeveniments es poden deduir dels dos primers, i a partir d'aquest, usant unions i complementaris

(vegeu la Taula 6.3).

Taula 6.3: Esdeveniments senzills i la seua relació amb els dos més elementals (vegeu la Taula 6.1), i amb els quals es van construint a partir d'aquests

Esdeveniment	Relació amb els precedents
$\{X < x\}$	$\{X < x\} \cup \{X = x\} = \{X \leq x\}$
$\{X > x\}$	$\{X > x\} = \overline{\{X \leq x\}}$
$\{X \geq x\}$	$\{X \geq x\} = \{X > x\} \cup \{X = x\}$
$\{X \neq x\}$	$\{X \neq x\} = \overline{\{X = x\}}$
$\{x_1 < X \leq x_2\}$	$\{X \leq x_1\} \cup \{x_1 < X \leq x_2\} = \{X \leq x_2\}$

Des del punt de vista matemàtic, gràcies a les relacions mostrades a la Taula 6.3 i a les propietats de cada probabilitat (vegeu la Definició 5.2.1, punt 3, i Propietat 5.2.1, punt 2), les probabilitats dels esdeveniments de la Taula 6.2 es poden obtenir a partir de les probabilitats dels dos esdeviniments més simples (Taula 6.1). Aquesta és la raó per la qual hi ha 2 (i només 2) funcions de probabilitat, que es presenten a la secció següent.

6.2 Funcions associades a les probabilitats de variables aleatòries

La principal utilitat de definir el concepte abstracte de variable aleatòria és aprofitar la traducció de les probabilitats (de l'espai mostral original) com funcions matemàtiques (de l'espai mostral numèric).

6.2.1 Funcions f i F a la variable discreta

Si X és una variable aleatòria discreta, podem definir dues funcions associades a les probabilitats dels dos esdeveniments bàsics:

- Funció de probabilitat (massa o quantia) f:

$$f(x) = \Pr(X = x)$$

- Funció de distribució acumulada F:

$$F(x) = \Pr(X \leq x)$$

Aquestes funcions estan molt relacionades, i sempre es pot calcular una d'aquestes a partir de l'altra:

- $F(x) = \sum_{y \leq x} f(y)$, ja que $\{X \leq x\} = \bigcup_{y \leq x} \{X = y\}$ és una unió disjunta, la probabilitat de la unió passa a ser la suma de les probabilitats de cada conjunt.

- $f(x) = F(x) - F(x^-)$, on $x \in X(E)$ i x^- simbolitza el valor immediatament inferior a x en el conjunt $X(E)$ (si no n'hi ha cap, $F(x^-) = P(X \leq x^-) = 0$). Aquesta relació és conseqüència de la relació $\{X \leq x\} = \{X \leq x^-\} \cup \{X = x\}$ (que és també una unió disjunta). Obviament, $f(x) = 0$ si $x \notin X(E)$.

A la Figura 6.1 es pot veure un exemple concret de la parella de funcions f i F per a una variable discreta.

Exercici 6.2.1 *Siga una probabilitat amb funció de massa* $f(x) = \dfrac{1}{2^x}$ *per* $x \in \{1, 2, 3, \ldots\}$. *Calculeu* $\Pr(X = 1)$, $\Pr(X = 5)$, $\Pr(X \leq 3)$, $\Pr(X \geq 5)$. *Aposteu pel resultat amb major probabilitat.*

Exercici 6.2.2 *Siga una probabilitat amb funció de distr. acum.* $F(x) = \dfrac{x}{10}$ *per* $x \in \{1, 2, \ldots, 10\}$. *Calculeu* $\Pr(X = 1)$, $\Pr(X = 5)$, $\Pr(X \leq 3)$, $\Pr(X \geq 5)$. *Aposteu pel resultat amb major probabilitat.*

6.2.2 Funcions F i f a la variable contínua

Si X és una variable aleatòria contínua, donat que $X(E)$ és un interval, la infinitud de valors possibles fa que l'únic model matemàtic de probabilitat que es pot definir tinga certes restriccions. Per una banda, la família d'esdeveniments no és qualsevol família de subconjunts. No obstant, eixa família (els subconjunts Borelians) cobreix quasi la totalitat de subconjunts imaginables i raonables. A més, la probabilitat ha d'assignar probabilitat 0 a "quasi tots" els valors aïllats $x \in X(E)$, per a poder complir la condició $P(X(E)) = 1$. Els intervals són els protagonistes.

El concepte de "suma" que existeix en la variable discreta es converteix en "integral" en la variable contínua. Per tant, només podem mantenir una funció (de les dues de la secció anterior) que representa una probabilitat. L'altra funció serà interpretable, però no expressarà cap probabilitat.

- Funció de distribució acumulada F:

$$F(x) = \Pr(X \leq x)$$

  Té la mateixa definició que en el cas discret, i la característica de ser contínua.

- Funció de densitat de probabilitat f:

$$f(x) = \lim_{h \to 0} \frac{\Pr(X \in [x, x+h])}{h}$$

  En dividir la probabilitat de l'interval $[x, x+h]$ entre la seua llargària, h (i prendre límit quan $h \to 0$), ja no es pot parlar de probabilitat (que és, des de la seua definició, una paraula reservada a un criteri que compleix certes propietats), però no deixa de ser un indicador de la "certesa", "credibilitat" o "versemblança" del resultat $x \in X(E)$. Per açò s'anomena funció de "densitat" de probabilitat.

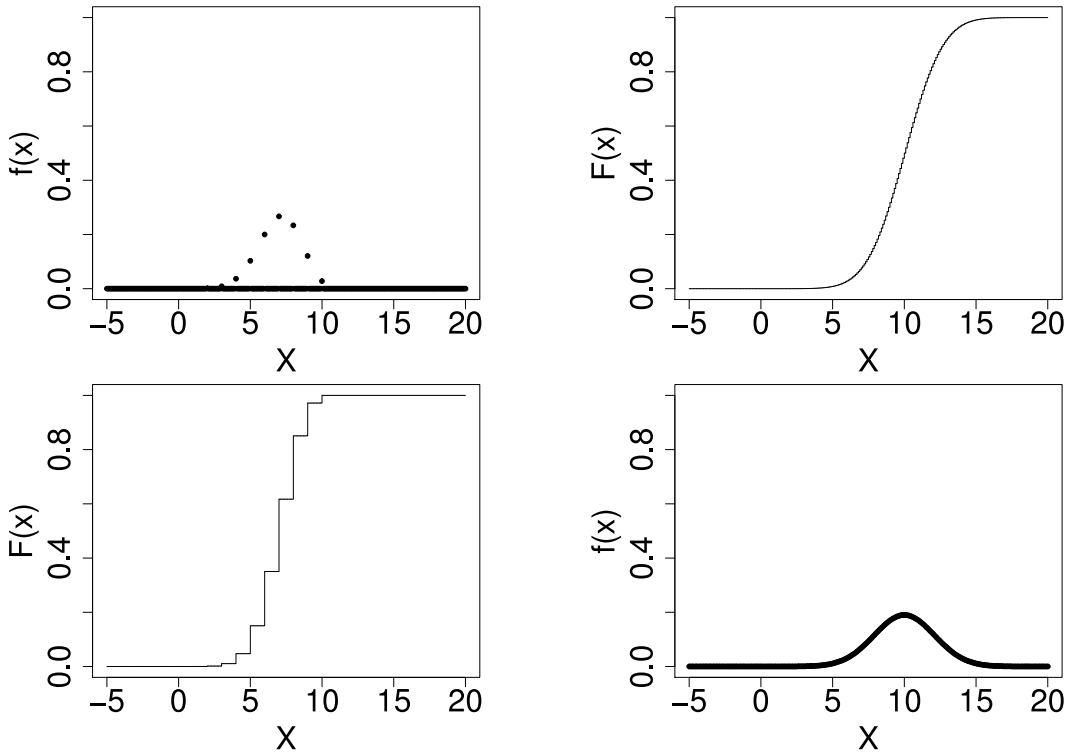


Figura 6.1: Exemple de funcions f i F per a una variable aleatòria X discreta (esquerra) i contínua (dreta)

Aquestes funcions continuen estant molt relacionades, i es poden calcular sempre una d'aquestes a partir de l'altra, en virtud de la definició de f:

- $F(x) = \int_{-\infty}^{x} f(t)dt$
- $f(x) = F'(x)$

En realitat, la igualtat entre $f(x)$ i $F'(x)$ només es verifica als valors x per als quals f és contínua (habitualment tots, excepte, com a màxim, una quantitat finita). A la Figura 6.1 es pot veure un exemple concret de la parella de funcions f i F per a una variable contínua.

De la definició de F es dedueix:

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

(vegeu la Figura 6.2) i de la relació de F amb f es té que:

$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)\mathrm{d}x.$$

(vegeu la Figura 6.3).

Com es comentava a l'inici de la secció, l'únic model matemàtic de probabilitat que gestiona una variable contínua assigna probabilitat nul·la a qualsevol valor aïllat $x \in X(E)$. Per açò, es té que:

$$F(x) = P(X \leq x) = P(X < x).$$

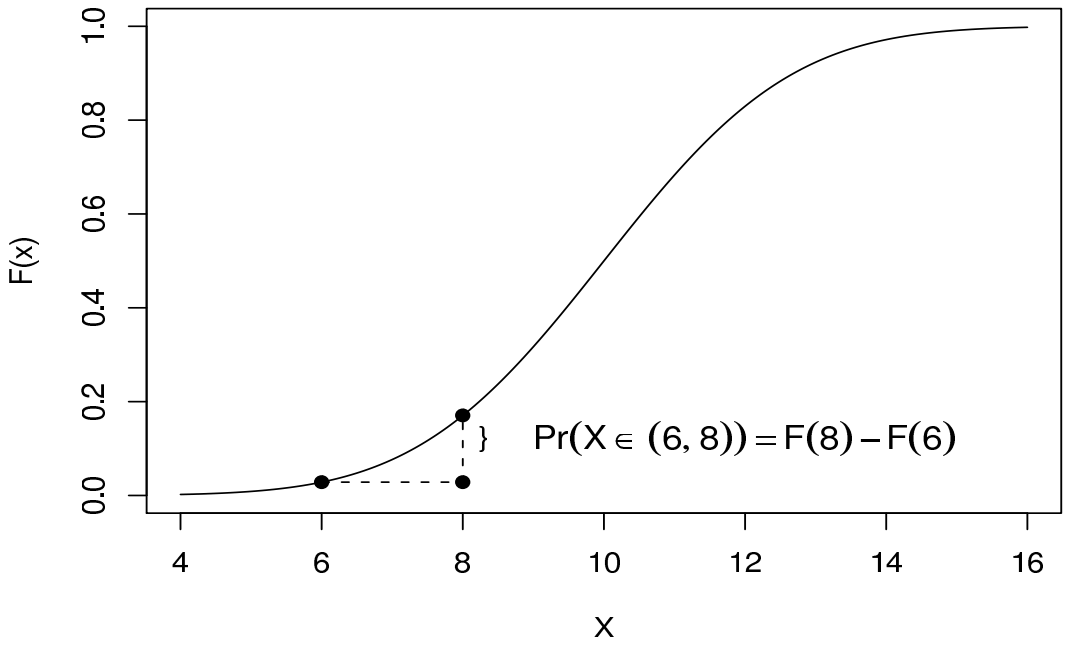


Figura 6.2: Exemple de càlcul de la probabilitat d'un interval usant la funció F i la seua interpretació gràfica

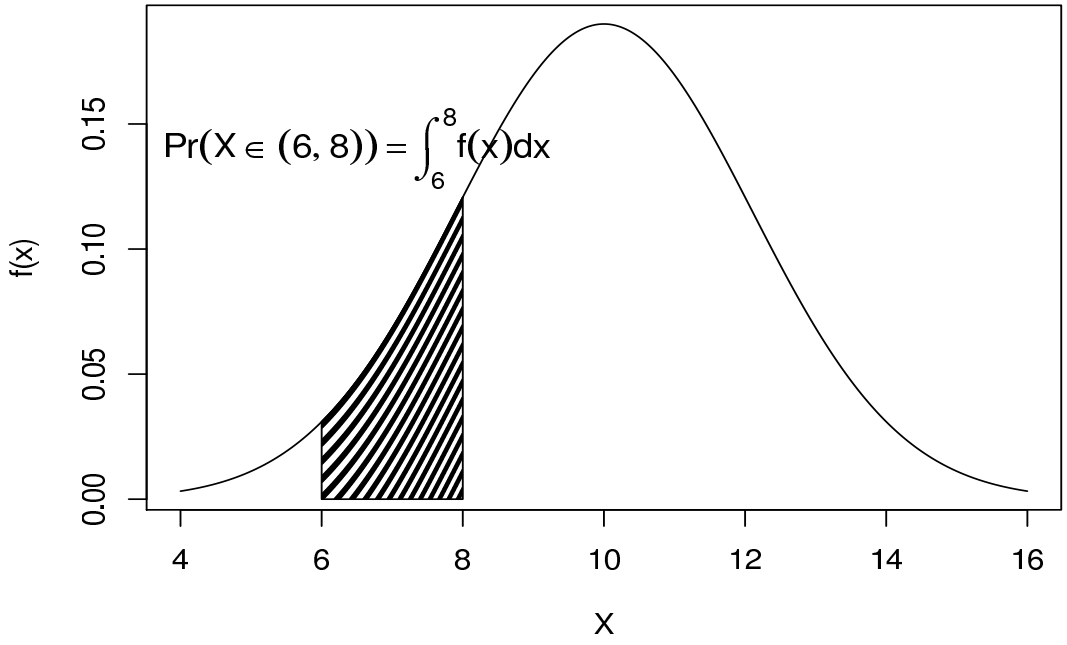


Figura 6.3: Exemple de càlcul de la probabilitat d'un interval usant la funció f i la seua interpretació gràfica

i que:

$$P(x_1 \leq X \leq x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2).$$

En resum, en la variable contínua no cal preocupar-se de si els extrems dels intervals compten o no compten a l'hora de calcular les probabilitats.

Exercici 6.2.3 *Considerem una variable aleatòria X amb funció de distribució acumulada*

$$F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^2}{100}, & x \in (0, 10) \\ 1, & x \geq 10 \end{cases}$$

(representada a la Figura 6.4, esquerra). Calculeu $\Pr(X = 1)$, $\Pr(X = 5)$, $\Pr(X \leq 3)$, $\Pr(X \geq 5)$. *Aposteu per l'interval de llargària* 2 *amb major probabilitat de tots els possibles.*

Exercici 6.2.4 *Considerem una variable aleatòria X amb funció de densitat*

$$f(x) = \begin{cases} 0, & x \leq 0 \\ 0.03x, & x \in (0, 5) \\ 0.15, & x \in (5, 10) \\ 0, & x \geq 10 \end{cases}$$

(vegeu la Figura 6.4, dreta). Calculeu $\Pr(X = 1)$, $\Pr(X = 5)$, $\Pr(X \leq 3)$, $\Pr(X \geq 5)$. *Aposteu per l'interval de llargària* 2 *amb major probabilitat.*

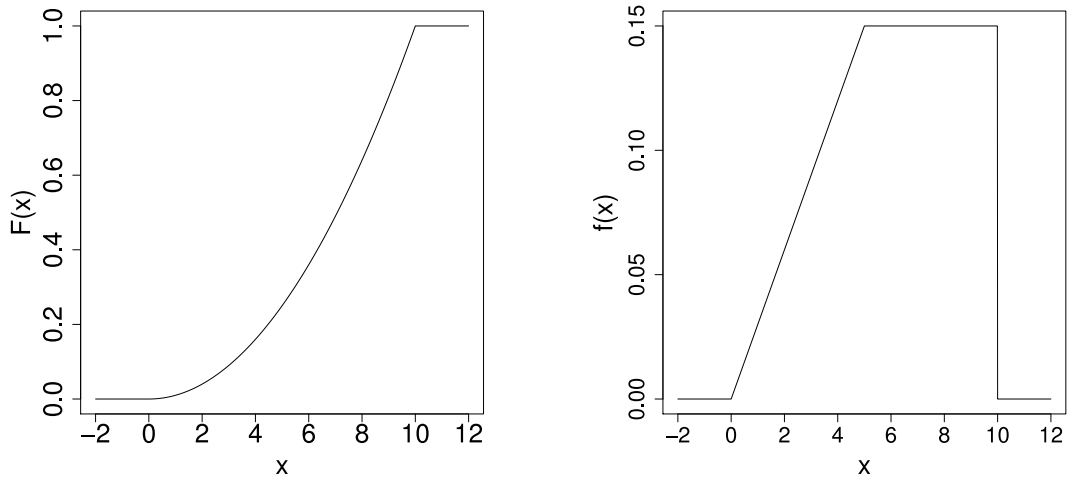


Figura 6.4: Gràfica de la funció F de l'Exercici 6.2.3 (esquerra) i f de l'Exercici 6.2.4 (dreta)

6.2.3 Propietats de les funcions f i F

Propietat 6.2.1 *Si* Pr *és una probabilitat definida per una variable aleatòria X, i es defineixen les funcions f i/o F a partir de* Pr *tal com s'explica a les Seccions 6.2.1 i 6.2.2, aleshores, la funció F verifica:*

- $F(-\infty) = 0$, $F(+\infty) = 1$.
- *És creixent.*
- *En variables aleatòries discretes, és discontínua (encara que contínua per la dreta).*
- *En variables aleatòries contínues, és contínua.*

Mentre que la funció f compleix:

- *És no negativa.*
- *En variables aleatòries discretes,* $\sum_x f(x) = 1$.
- *En variables aleatòries contínues,* $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Reciprocament, qualsevol funció, F o f, definida complint les condicions anteriors és perfectament vàlida per definir una probabilitat Pr.

Exercici 6.2.5 *Calculeu la funció de probabilitat proporcional a $g(x) = x$ per a $x \in \{1, \ldots, 5\}$.*

Exercici 6.2.6 *Calculeu la funció de densitat proporcional a $g(x) = x$ per a $x \in (0, 5)$.*

Exercici 6.2.7 *Per què no pot ser $f(x) = 0.2[(x-2)^2 - 1]$ per a $x \in \{0, 1, 2, 3, 4\}$ una funció de probabilitat vàlida?*

6.3 Variable aleatòria multidimensional

També és possible associar al resultat d'un experiment més d'un valor.

Exemple 6.3.1 *Si llancem 1 dau i 2 monedes i anotem conjuntament el resultat del dau i la posició de les monedes, trobem que l'espai mostral es pot representar com $E = \{1CC, 1C+, 1+C, 1++, 2CC, 2C+, 2+C, 2++, \ldots, 6CC, 6C+, 6+C, 6++\}$.*

Es podria considerar la variable aleatòria X = "número del dau", amb $X \in \{1, 2, 3, 4, 5, 6\}$. També es podria pensar a la variable aleatòria Y = "número de cares", amb $Y \in \{0, 1, 2\}$.

Però una altra variable aleatòria podria ser l'observació conjunta Z = "número del dau i nombre de cares". Es tindria que $Z \in \{(1,0), (1,1), (1,2), (2,0), (2,1), (2,2), \ldots\} = \{1, 2, 3, 4, 5, 6\} \times \{0, 1, 2\}$ i podriem escriure $Z = (X, Y)$. Es tractaria d'una variable bidimensional.

Les variables aleatòries multidimensionals també admeten funcions de distribució, etc. En el cas bidimensional, si $Z = (X, Y)$ es defineixen amb la mateixa filosofia:

- Variable aleatòria discreta:

  $f_Z(x,y) = \Pr(X = x, Y = y)$ i

  $F_Z(x,y) = \Pr(X \leq x, Y \leq y)$

- Variable aleatòria contínua:

  $F_Z(x,y) = \Pr(X \leq x, Y \leq y)$

Les distribucions marginals de cada variable aleatòria unidimensional s'obtenen directament des de la definició, notant que quan només observem una variable, acceptem tots els valors possibles de l'altra:

$$f_X(x) = \begin{cases} \text{(discr.)} \sum f_Z(x,y) \\ \text{(cont.)} \int_{-\infty}^{y=+\infty} f_Z(x,y)\mathrm{d}y \end{cases} \qquad F_X(x) = F_Z(x, +\infty)$$

$$f_Y(y) = \begin{cases} \text{(discr.)} \sum f_Z(x,y) \\ \text{(cont.)} \int_{-\infty}^{x=+\infty} f_Z(x,y)\mathrm{d}x \end{cases} \qquad F_Y(y) = F_Z(+\infty, y)$$

Les distribucions condicionals, encara que d'aparença més complexa, també s'obtenen després de l'aplicació directa de la definició de probabilitat condicional:

$$f_{X|Y=y}(x) = \begin{cases} \text{(discr.)} \dfrac{f_Z(x,y)}{\sum_x f_Z(x,y)} \\ \text{(cont.)} \dfrac{f_Z(x,y)}{\int_{-\infty}^{+\infty} f_Z(x,y)dx} \end{cases} \qquad F_{X|Y=y}(x) = \frac{F_Z(x,y)}{F_Z(+\infty, y)}$$

$$f_{Y|X=x}(y) = \begin{cases} \text{(discr.)} \dfrac{f_Z(x,y)}{\sum_y f_Z(x,y)} \\ \text{(cont.)} \dfrac{f_Z(x,y)}{\int_{-\infty}^{+\infty} f_Z(x,y)dy} \end{cases} \qquad F_{Y|X=x}(y) = \frac{F_Z(x,y)}{F_Z(x, +\infty)}$$

Noteu que només tenen sentit aquestes probabilitats quan els denominadors respectius no s'anul·len.

Definició 6.3.1 (Independència de variables aleatòries) *La variable aleatòria X és independent de la variable aleatòria Y si la distribució de probabilitat condicionada de X no és afectada pels valors que prenga Y, és a dir, si*

$$f_{X|Y=y}(x) = f_X(x) \text{ per tot } (x,y)$$

(o equivalentment amb la funció F).

Es pot demostrar que la noció d'independència és recíproca, raó per la qual es dirà que les variables X i Y són (o no són) independents.

Propietat 6.3.1 *Dues variables aleatòries X i Y són independents si i només si la distribució de probabilitat de la variable aleatòria conjunta $Z = (X, Y)$ verifica que:*

$$f_Z(x, y) = f_X(x) \cdot f_Y(y) \text{ per tot } (x, y)$$

Corol·lari 6.3.1 *Si $X_1, X_2, \ldots, X_n$ són variables aleatòries independents entre si, i totes amb la mateixa funció f, aleshores la distribució conjunta de $X = (X_1, \ldots, X_n)$ té com a funció f_X:*

$$f_X(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

Exercici 6.3.1 *Tenim un dau de 6 cares i un altre de 12 cares. Denotem f_6 com la funció de probabilitat de la variable aleatòria "resultat del dau de 6 cares" i f_{12}, respectivament, pel "resultat del dau de 12 cares". Es llança un dels dos daus 4 vegades...*

1. *...obtenint els valors: $2, 4, 1, 2$. Calculeu la probabilitat amb cada dau:*

$$f_6(2, 4, 1, 2) = \qquad\qquad f_{12}(2, 4, 1, 2) =$$

2. *...obtenint els valors: $3, 7, 5, 5$. Calculeu la probabilitat amb cada dau:*

$$f_6(3, 7, 5, 5) = \qquad\qquad f_{12}(3, 7, 5, 5) =$$

Si estiguereu obligats a endevinar quin dau s'ha utilitzat en cada cas, per quin es decantarieu i amb quin criteri?

En els capítols d'inferència es trauran conclusions a partir de les mostres, tal com s'intueix en aquest exercici.

6.4 Mitjana i variància d'una variable aleatòria

Una distribució de probabilitat d'una variable aleatòria es pot interpretar com una distribució de freqüències d'una població. Els descriptors més importants de la població són la **posició central** i la **dispersió** i es defineixen principalment mitjançant:

- Mitjana o esperança matemàtica (denotat indistintament per $\mathbb{E}(X)$, μ o μ_X):

$$\text{(discr.)} \sum_x xf(x) \qquad\qquad \text{(cont.)} \int_{-\infty}^{+\infty} xf(x)dx$$

- Variància (denotat indistintament per $\mathbb{V}(X)$, σ^2 o σ_X^2):

$$\text{(discr.)} \sum_x (x-\mu_X)^2 f(x) \qquad\qquad \text{(cont.)} \int_{-\infty}^{+\infty} (x-\mu_X)^2 f(x)dx$$

  També es pot parlar de la desviació típica (denotada per $SD(X)$, o σ, o σ_X) com l'arrel quadrada de la variància.

Exercici 6.4.1 *Calculeu la mitjana i la variància per a les distribucions que apareixen en els exercicis d'aquest tema.*

- $f(x) = \frac{1}{2^x}$ *per* $x \in \{1, 2, 3, \ldots\}$
- $F(x) = \frac{x}{10}$ *per* $x \in \{1, 2, \ldots, 10\}$
- $f(x) = \begin{cases} 0, & x \leq 0 \\ 0.03x, & x \in (0, 5) \\ 0.15, & x \in (5, 10) \\ 0, & x \geq 10 \end{cases}$
- $F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{x^2}{100}, & x \in (0, 10) \\ 1, & x \geq 10 \end{cases}$

Amb una variable aleatòria bidimensional $Z = (X, Y)$ també es pot definir la covariància entre les seues components X i Y com:

$$\mathbb{C}\text{ov}(X, Y) = \begin{cases} \text{(d.)} \displaystyle\sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \\ \text{(c.)} \displaystyle\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy \end{cases}$$

És facil provar matemàticament que si X i Y són variables aleatòries independents, aleshores $\mathbb{C}\text{ov}(X, Y) = 0$ (però no necessàriament el contrari).

Propietat 6.4.1 *Si X i Y són variables aleatòries i a, b són nombres reals, aleshores:*

- $\mathbb{E}(a + bX) = a + b\mathbb{E}(X)$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\mathbb{V}(a + bX) = b^2\mathbb{V}(X)$
- $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathbb{C}\text{ov}(X, Y)$

6.5 Exercicis proposats

Exercici 6.5.1 *Suposant que X és una variable aleatòria discreta, escriviu les següents probabilitats de dues maneres diferents: (1) únicament utilitzant la funció de probabilitat f, i (2) únicament utilitzant la funció de distribució F:*

- $P(X = 3) =$
- $P(X \neq 3) =$
- $P(X \leq 3) =$

- $P(X < 3) =$
- $P(X > 3) =$
- $P(X \geq 3) =$
- $P(3 < X \leq 7) =$
- $P(3 < X < 7) =$
- $P(3 \leq X < 7) =$
- $P(3 \leq X \leq 7) =$
- $P(X = 3|X \leq 5) =$
- $P(X = 3|X \geq 5) =$

Exercici 6.5.2 *Suposant que X és una variable aleatòria contínua, escriviu les següents probabilitats únicament utilitzant la funció de distribució F:*

- $P(X = 3) =$
- $P(X \neq 3) =$
- $P(X \leq 3) =$
- $P(X < 3) =$
- $P(X > 3) =$
- $P(X \geq 3) =$
- $P(3 < X \leq 7) =$
- $P(3 < X < 7) =$
- $P(3 \leq X < 7) =$
- $P(3 \leq X \leq 7) =$
- $P(X = 3|X \leq 5) =$
- $P(X = 3|X \geq 5) =$

Exercici 6.5.3 *Cada matí teniu l'opció d'agafar dos trajectes diferents per arribar a la universitat. La duració de cada trajecte depén de l'estat del trànsit que, per simplificar, direm que està fluid (el 10% de les ocasions), normal (el 60% de les voltes) o dens (30% de les ocasions). Segons l'estat del tràfic s'obtenen els temps dels trajectes segons la Taula 6.4. Calculeu el trajecte que haurieu de prendre si no coneixeu l'estat del trànsit, usant com a criteri triar el de menor temps esperat.*

Taula 6.4: Taula de l'Exercici 6.5.3

Estat	Probabilitat	Trajecte 1	Trajecte 2
F = Fluid	0.10	15 m	30 m
N = Normal	0.60	35 m	40 m
D = Dens	0.30	70 m	50 m

Exercici 6.5.4 *Imagineu que se us proposa triar entre dos jocs d'atzar. En ambdós jocs s'ha de llançar una moneda, i el premi guanyat (en cada joc) ve indicat (en euros) a la Taula 6.5. Nombres negatius indiquen pèrdues. En l'opció A pots tenir pèrdues, mentre que en la B sempre guanyes.*

- *Si només podeu triar una volta, quina opció triarieu i per què?*
- *Si podeu repetir el procediment de triar una opció, diguem 1000 voltes, quina opció triarieu en cada ocasió i per què?*

Taula 6.5: Taula de l'Exercici 6.5.4

Resultat	Probabilitat	Joc A	Joc B
Cara	0.5	350	120
Creu	0.5	-100	80

Exercici 6.5.5 *Si X i Y són variables aleatòries independents de manera que $\mathbb{V}(X) = 3$ i $\mathbb{V}(Y) = 4$, quant val $\mathbb{V}(X + Y)$?*

Capítol 7

Models de poblacions de dades numèriques

7.1 Introducció

7.1.1 Objectius

L'objectiu fonamental d'aquest capítol és l'exposició del concepte de model de variable aleatòria i la presentació d'una sèrie de models senzills.

El concepte de variable aleatòria significa l'abstracció d'un experiment concret en un espai de resultats numèrics. A partir d'aquesta abstracció, dos experiments molt diferents es poden considerar "iguals" si les probabilitats dels esdeveniments que ens interessen són iguals. Per exemple, si un jugador estima que té una probabilitat de 0.001 de guanyar un joc i, d'altra banda, un metge estima que la probabilitat de curació del seu pacient és 0.001, les dues situacions són comunes i poden analitzar-se amb un mateix plantejament abstracte, que no involucre ni jocs ni malalties, simplement valors numèrics. Aquest plantejament es converteix en una entitat matemàtica que s'anomena model.

Els models matemàtics que es creen tracten de descriure situacions reals, raó per la qual es plantegen seguint els criteris de:

- Lògica
- Simplicitat
- Explicació de la realitat

Els dos últims criteris poden ser contradictoris, si la situació real que es tracta de modelitzar és realment complexa, però la idea que cal tenir és la següent: *si dos models s'ajusten igualment bé a la realitat, aleshores es prendrà com a vàlid el més simple dels dos.*

Una hipòtesi raonable (lògica) en multitud de situacions és la de l'equiprobabilitat: en un espai mostral de resultats "indistingibles", tots haurien de tenir, lògicament, les mateixes oportunitats d'eixir com a resultat, no hi ha raons per decantar-se en atribuir més versemblança a uns que a altres. Podem veure aquest tipus de situació en dos exemples.

Exemple 7.1.1 *Una urna conté n boles idèntiques excepte pel color. Una mà innocent meneja, sense mirar, amb força i molt de temps, les boles, fins que es decideix per traure'n una. Si no hi ha indicacions addicionals, un model lògic implica unes probabilitats iguals (a $1/n$) per totes i cadascuna de les boles.*

Si una porció p de les boles són d'un color determinat, aleshores la probabilitat d'obtenir una bola d'aquell color és p.

Exemple 7.1.2 *Una fletxa pot rodar al voltant del centre d'un cercle (com una ruleta). Una mà innocent fa girar amb molta força la fletxa (sense poder controlar on pararà) de manera que pot acabar apuntant a qualsevol direcció possible. Si no hi ha indicacions addicionals, un model lògic implica que totes i cadascuna de les direccions són igualment versemblants.*

Si una porció p de la circumferència es marca de manera especial, aleshores la probabilitat que la fletxa apunte a aqueixa part de la circumferència és p.

7.1.2 Simulació d'experiments

Si una situació real involucra esdeveniments amb probabilitats iguals a les probabilitats d'esdeveniments d'un experiment més senzill, es pot realitzar un paral·lelisme entre els dos, i *simular* la situació real mitjançant l'experiment senzill.

Exemple 7.1.3 *Si estimes que el fet que aproves una assignatura és un fenomen aleatori, perquè hi ha molts factors que no podràs controlar (dificultat de l'examen, disponibilitat de temps d'estudi, patiment de malalties inesperades, etc.), pots fixar subjectivament la probabilitat d'aprovar en un valor p (major o menor, segons la teua confiança).*

Si vols fer una simulació de la situació real de presentar-te a l'examen, en lloc d'esperar a la fi del curs, pots usar una urna o una ruleta, de manera que pugues crear un esdeveniment amb la mateixa probabilitat p. Aleshores, jugues amb l'urna o la ruleta, i segons el resultat obtingut, observaràs l'esdeveniment, i podràs simular si has aprovat o no l'examen.

És clar que el resultat que isca en la simulació no vincula el resultat que obtindràs si et presentes a l'examen.

La simulació d'experiments és una eina molt útil en camps on els experiments són cars, o perillosos. Per exemple, quan es dissenyen magatzems de residus nuclears, s'estudia per simulacions el temps transcorregut fins que ocorre una fuita. Quan les simulacions mostren que aquest temps és *quasi sempre* superior a milers d'anys, s'accepta que el magatzen és segur i es contrueix segons s'ha dissenyat.

7.1.3 Poblacions de dades

Un experiment aleatori té associat un conjunt de resultats possibles i impredictibles. Conéixer-lo bé implica poder avaluar les opcions o possibilitats que té cada resultat possible. Realitzar l'experiment i observar el seu resultat dóna lloc a una *dada*. Repetir el procès un nombre de voltes dóna lloc a una *mostra*

de dades. Intuïtivament, quan major és la mostra, millor es coneix la natura de l'experiment (valors obtinguts, variabilitat d'aquests, etc.). Si es poguera repetir l'experiment una infinitat de vegades s'obtindria la població de resultats, que és al cap i a la fi l'objecte que es vol conéixer, però la infinitat no és abastable al món real.

A continuació s'exposen, esquemàticament, els principals models de variable aleatòria discreta i contínua, amb la presentació de les propietats més destacables. A nivell pràctic, és important saber reconéixer en casdacuna de les situacions concretes que se'ns presenten:

- La presència d'una quantitat aleatòria (variable aleatòria).
- El model que segueix la variable aleatòria (si és que segueix algun dels ací citats).

Amb això, es poden estimar probabilitats d'esdeveniments concrets, que poden ajudar a prendre decisions més encertades.

7.2 Prova de Bernoulli de paràmetre p

El model de variable aleatòria més senzill és aquell que només té dos resultats (0 o 1), i valora si s'observa (1) o no s'observa (0) un esdeveniment particular. La descripció del model és la següent.

Definició 7.2.1 *El model* ***prova de Bernoulli de paràmetre*** p *ve definit per l'experiment i té les propietats que es mostren a continuació:*

- ***Experiment:*** *Un esdeveniment etiquetat com* ***èxit*** *ocorre amb una probabilitat p coneguda. Es va a realitzar l'experiment i es va a observar si l'èxit ocorre o no.*
- ***Variable aleatòria:*** $X = \begin{cases} 0 & , \textbf{\textit{no}} \text{ ocorre l'èxit} \\ 1 & , \text{ ocorre l'èxit} \end{cases}$
- ***Espai mostral:*** $X \in \{0, 1\}$
- ***Notació:*** $X \sim \text{Be}(p)$
- $\boldsymbol{f(x)} = \begin{cases} 1-p & x = 0 \\ p, & x = 1 \\ 0, & \text{altre cas} \end{cases}$
- $\boldsymbol{F(x)}$ *No la gastarem al llarg del llibre (vegeu com s'obté a partir de f en la p. 107)*
- ***Esperança i Variància:*** $\mathbb{E}(X) = p$, $\mathbb{V}(X) = p(1-p)$
- ***Nota:*** *X pot ser interpretada com el* ***nombre d'èxits obtinguts*** *(0 o 1).*

L'experiment descrit s'anomena prova de Bernoulli de paràmetre p.

Exercici 7.2.1 *Un estudi de transmissió informàtica en cert canal arriba a la conclusió que un bit és transmés erròniament (és a dir, el seu valor canvia de l'emissor al receptor) en* 1 *de cada* 20 *enviaments, independentment del valor del bit emés i de la correcció en l'emissió del bit anterior. Es transmet un bit:*

- *Si pensem que "èxit" és la transmissió correcta, quina és la probabilitat que ho faça correctament?*

  $X_1 =$ "....." $\rightarrow$ $X_1 \sim \text{Be}(p = \quad)$

  $\Pr(''\text{correcte}'') = \Pr(X_1 = \quad) =$

- *Si pensem que "èxit" és la transmissió incorrecta, quina és la probabilitat que ho faça correctament?*

  $X_2 =$ "....." $\rightarrow$ $X_2 \sim \text{Be}(p = \quad)$

  $\Pr(''\text{correcte}'') = \Pr(X_2 = \quad) =$

Exercici 7.2.2 *Deduïu les fórmules d'esperança i variància a partir de les fórmules de la pàgina 114.*

7.3 Binomial de paràmetres n i p

El model binomial apareix en situacions on es fa un nombre concret de proves de Bernoulli, i l'interès de l'usuari està a avaluar les probabilitats sobre quantes de les proves donaran un resultat reeixits (totes, no cap, algunes...). Per exemple, si un alumne fa un examen tipus test amb 20 preguntes, i contesta aleatòriament, cada pregunta es converteix en una prova de Bernoulli (bé/malament), i la quantitat aleatòria d'interès és el nombre de preguntes contestades correctament (ja que l'aprovat depén d'aqueix nombre).

Introduïm el model precedit d'un parell d'exercicis assequibles.

Exercici 7.3.1 *En realitzar un experiment aleatori, un esdeveniment A té probabilitat* 0.17 *d'ocórrer. Si l'experiment es va a realitzar* 3 *vegades, cadascuna independentment de l'anterior, quina és la probabilitat que ocórrega l'esdeveniment A en...*

- *...totes les ocasions?*
- *...cap ocasió?*
- *...x ocasions? (x és qualsevol valor possible)*

(Ajuda: usar la Proposició 5.2.2 de la pàgina 98).

Després contesteu les mateixes preguntes si l'esdeveniment A té una probabilitat p d'ocórrer, i es realitza un nombre n de vegades.

Exercici 7.3.2 *Quants bytes (de* 8 *bits) diferents es poden escriure amb* 3 *uns i* 5 *zeros?*

(Ajuda: és un exercici de combinatòria. Penseu que cada byte queda determinat quan decideixes les tres posicions que ocupen els uns. Les posicions són de la 1 a la 8, i en triar les tres posicions, l'ordre en què les tries no és rellevant).

I en el cas més general, quantes cadenes de 'n' bits diferents es poden escriure amb 'x' uns i la resta de zeros? (x pot ser qualsevol valor possible).

Amb la resolució dels exercicis anteriors, el model binomial que es presenta ací no necessita més explicació.

Definició 7.3.1 *El model* ***binomial de paràmetres*** n ***i*** p *ve definit per l'experiment i té les propietats que es mostren a continuació:*

- ***Experiment:*** *Es realitzen n proves de Bernoulli de paràmetre p independents i s'observa en quantes d'aquestes ha ocorregut l'èxit*
- ***Variable aleatòria:*** $X =$ *"****nombre d'èxits****".*
- ***Espai mostral:*** $X \in \{0, 1, 2, \ldots, n\}$
- ***Notació:*** $X \sim \text{Bin}(n, p)$
- $\boldsymbol{f(x)} = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n \\ 0, & \textit{altre cas} \end{cases}$
- $\boldsymbol{F(x)}$ *No la gastarem al llarg del llibre (vegeu com s'obté a partir de f en la p. 107)*
- ***Esperança i Variància:*** $\mathbb{E}(X) = np$, $\mathbb{V}(X) = np(1-p)$
- ***Nota:*** $X = \sum_{i=1}^{n} X_i$, *on cada* $X_i \sim Be(p)$ *i són totes independents.*

Exercici 7.3.3 *Estem a l'estudi de transmissió informàtica de l'Exercici 7.2.1 (un bit és transmés erròniament en* 1 *de cada* 20 *enviaments). Ara l'experiment consisteix a transmetre una cadena de 8 bits i observar la qualitat de la transmissió pel nombre de bits correctes o incorrectes.*

- *Definiu la variable aleatòria* $X_1 =$*".....".....* *de manera que s'adeqüe al perfil d'una variable aleatòria amb distribució Binomial de paràmetres* $n=$ ☐ *i* $p=$ ☐.
- *Definiu la variable aleatòria* $X_2 =$*".....".....* *d'altra manera que s'adeqüe al perfil d'una variable aleatòria amb distribució Binomial de paràmetres* $n=$ ☐ *i* $p=$ ☐.
- *Nombre esperat de bits correctes?*
- *Probabilitat que es transmeta la cadena perfectament?*

  $\Pr(X_1 \quad) =$ $\Pr(X_2 \quad) =$
- *Probabilitat que es transmeta amb 2 o menys errors?*

  $\Pr(X_1 \quad) =$ $\Pr(X_2 \quad) =$

Propietat 7.3.1 *Si* $X_1 \sim \text{Bin}(n_1, p)$ *i* $X_2 \sim \text{Bin}(n_2, p)$ *són independents i* $Y = X_1 + X_2$*, aleshores* $Y \sim \text{Bin}(n_1 + n_2, p)$*.*

Exercici 7.3.4 *Demostreu la propietat anterior usant un raonament que involucre només la Nota que apareix a la Definició 7.3.1 del model binomial.*

Exercici 7.3.5 *Deduïu les fórmules d'esperança i variància a partir de les fórmules de la pàgina 114, o millor, a partir de la Nota de la Definició 7.3.1 i de la Propietat 6.4.1 de la pàgina 115.*

7.4 Binomial negativa de paràmetres r i p

Encara basant-se en les proves de Bernoulli, un altre plantejament és aquell que compta les proves que es fan quan s'intenta arribar a obtenir un nombre concret d'èxits. Com és justament l'enfocament contrari que el de la binomial, per això la raó del seu nom.

Per exemple, el nombre de convocatòries que necessita un alumne per aprovar, és el nombre de proves de Bernoulli (cada convocatòria) que es fan quan s'intenta arribar a un èxit (que és aprovar).

Definició 7.4.1 *El model* ***binomial negatiu de paràmetres*** r ***i*** p *ve definit per l'experiment i té les propietats que es mostren a continuació:*

- ***Experiment:*** *Prefixat un nombre enter d'èxits* r*, es fan proves de Bernoulli de paràmetre* p *independents i indefinidament fins que s'observa el* r*-èssim èxit. Després s'atura el procés i es compten el nombre de proves que s'han realitzat.*
- ***Variable aleatòria:*** $X =$ *"**nombre de proves de Bernoulli realitzades (incloent-hi els*** r ***èxits)**".*
- ***Espai mostral:*** $X \in \{r, r+1, r+2, r+3, \ldots\}$.
- ***Notació:*** $X \sim \text{BinNeg}(n, p)$
- $\boldsymbol{f(x)} = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r}, & x = r, r+1, \ldots \\ 0, & \textit{altre cas} \end{cases}$
- $\boldsymbol{F(x)}$ *No la gastarem al llarg del llibre (vegeu com s'obté a partir de* f *en pàgina 107)*
- ***Esperança i Variància:*** $\mathbb{E}(X) = \dfrac{r}{p}$, $\mathbb{V}(X) = \dfrac{r(1-p)}{p^2}$
- ***Nota:*** *Quan* $r = 1$ *es coneix com* ***Geomètrica*** *i en aquest cas* $F(x) = 1 - (1-p)^x$*.*

Exercici 7.4.1 *Una (mala) impressora té una probabilitat d'espatllar-se de 0.05 a cada pàgina que imprimeix. S'estudia el nombre de pàgines impresses fins avariar-se (incloent-hi la de l'avaria).*

- *Feu un paral·lelisme entre l'experiment i l'extracció de boles d'una urna.*
- *Definiu la variable aleatòria $X =$"....." de manera que s'adeqüe al perfil d'una variable aleatòria amb distribució Binomial negativa de paràmetres $r =$ ☐ i $p =$ ☐.*
- *Expectativa del nombre de pàgines impresses fins la pròxima avaria.*
- *Probabilitat d'avariar-se a la 15a pàgina.*

  $\Pr(X \quad) =$
- *Probabilitat d'imprimir més de 15 pàgines abans de l'avaria.*

  $\Pr(X \quad) =$
- *Probabilitat d'imprimir més de 50 pàgines abans de l'avaria.*

  $\Pr(X \quad) =$

Exercici 7.4.2 *En l'estudi de la transmissió de bits de l'Exercici 7.2.1 (5% erronis) es va rebent i comparant la cadena de bits rebuda amb l'original, analitzant la llargària de la cadena quan es compten 10 errors.*

- *Definiu la variable aleatòria $X =$"....." de manera que s'adeqüe al perfil d'una variable aleatòria amb distribució Binomial negativa de paràmetres $r =$ ☐ i $p =$ ☐.*
- *Expectativa de la llargària de la cadena.*
- *Probabilitat d'haver transmés més de 100 bits?*

  $\Pr(X \quad) =$
- *Probabilitat d'haver transmés més de 1000 bits?*

  $\Pr(X \quad) =$

Exercici 7.4.3 *Deduïu la fórmula de F a partir de f per al cas $r = 1$, i les fórmules d'esperança i variància a partir de les fórmules de la pàgina 114.*

Exercici 7.4.4 *Deduiu la fórmula de f de la definició del model a partir de l'observació següent:*

> *L'esdeveniment $X = x$ ocorre quan la x-ena prova és un èxit, i dins les $x - 1$ proves anteriors hi ha hagut $r - 1$ èxits.*

i usant, per tant, la definició del model binomial, i l'independència de les proves de Bernoulli.

Exercici 7.4.5 *Deduïu les fórmules d'esperança i variància a partir de les fórmules de la pàgina 114, o millor, a partir de la Nota de la Definició 7.3.1 i de la Propietat 6.4.1 de la pàgina 115.*

7.5 Hipergeomètrica de paràmetres N, K i n

Els models anteriors s'apliquen a proves de Bernoulli idèntiques (totes amb la mateixa p) i independents. El model que presentem ací no es pot considerar d'aquesta manera i necessita el seu espai propi, encara que en ocasions es podrà confondre amb el model binomial.

Definició 7.5.1 *El model* ***hipergeomètric de paràmetres*** N, K ***i*** n *ve definit per l'experiment i té les propietats que es mostren a continuació:*

- ***Experiment:*** *Una urna conté N boles equiprobables, de les quals només K són de tipus* ***èxit****. Es trauen n boles (sense reemplaçament!) i s'observa quantes són de tipus èxit.*
- ***Variable aleatòria:*** $X =$ *"****nombre d'èxits****".*
- ***Espai mostral:*** $X \in \{\max(0, n-(N-K)), \ldots, \min(n, K)\}$.
- ***Notació:*** $X \sim \text{Hyper}(n, N, K)$
- $\boldsymbol{f(x)} = \begin{cases} \dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, & x = \max(0, n-(N-K)), \ldots, \min(n, K) \\ 0, & \textit{altre cas} \end{cases}$
- $\boldsymbol{F(x)}$ *No la gastarem al llarg del llibre (vegeu com s'obté a partir de f en la p. 107)*
- ***Esperança i Variància:*** $\mathbb{E}(X) = n\dfrac{K}{N}$, $\mathbb{V}(X) = n\dfrac{K}{N}(1-\dfrac{K}{N})\dfrac{N-n}{N-1}$

Lògicament, l'ús d'una urna amb boles és un model. En les situacions reals les urnes són conjunts d'elements qualssevol igualment elegibles, i es poden distingir de dos tipus.

Exercici 7.5.1 *Un producte ve en lots de* 500 *unitats, de les quals* 35 *són defectuoses. El comprador pren* 7 *unitats per comprovar si compra el lot (el comprarà només si totes són correctes).*

- *Definiu la variable aleatòria $X =$"....." de manera que s'adeqüe al perfil d'una variable aleatòria amb distribució Hipergeomètrica de paràmetres $N = \square$, $K = \square$ i $n = \square$.*
- *Expectativa d'unitats defectuoses trobades als lots.*
- *Probabilitat de no triar cap producte defectuós (i que, per tant, el client compre el lot).*

  $\Pr(X \quad) =$
- *Idem, però si equivoquem la distribució de X per una binomial $Y \sim \text{Bin}(n = \quad, p = \quad)$.*

  $\Pr(Y \quad) =$

Exercici 7.5.2 *Com al model binomial, l'obtenció de la fórmula de f és un exercici de combinatòria: l'esdeveniment $X = x$ (és a dir, triar x boles de tipus èxit) implica triar x boles del grup de K boles de tipus èxit, i $n - x$ boles del grup de $N - K$ boles de tipus no èxit, independentment. El nombre d'eleccions possibles són totes les eleccions de n boles d'un grup de N boles, i són totes equiprobables. Podeu deduir ara la fórmula?*

Exercici 7.5.3 *Calculeu les probabilitats dels possibles resultats del model binomial de paràmetre $n = 5$ i $p = 0.3$ i del model hipergeomètric de paràmetres $N = 1000$, $K = 300$ i $n = 5$, i compareu si són similars. Per què penseu que són tant similars?*

7.6 Poisson de paràmetre λ

El model que presentem en aquesta secció és més complex de definir rigorosament. A grans trets, una situació segueix el model de Poisson quan segueix les condicions descrites com **(Po1)**–**(Po5)**:

- **(Po1):** Hi ha un fenomen observable (**èxit**) que ocorre puntualment al llarg del temps o de l'espai. Per simplificar l'explicació pensarem únicament en temps.
- **(Po2):** Es va a registrar la variable X, que representa el nombre d'ocasions que s'observa l'èxit al llarg d'un interval de llargària coneguda i prefixada.
- **(Po3):** La probabilitat de que s'observe exactament un únic èxit en un interval de temps infinitessimal (és a dir, de llargària quasi zero) és proporcional a la llargària de l'interval.
- **(Po4):** La probabilitat que ocòrrega l'èxit més d'una volta en intervals de temps infinitessimals és nul·la.
- **(Po5):** El nombre d'ocasions que ocorre l'èxit en un interval és independent del que ocorre en un altre interval disjunt a l'anterior.

Es pot demostrar que mitjana del nombre d'ocasions que s'observa eixe èxit en un interval de temps prefixat és proporcional a la llargària de l'interval. Denotarem per λ l'esmentada mitjana.

Exercici 7.6.1 *Usant les condicions que defineixen un procés de Poisson, anem a explorar una forma bastant natural d'arribar a la fórmula de la funció de probabilitat f del model de Poisson de mitjana λ. Contesteu cada apartat que demane un resultat:*

1. *Siga X = "nombre d'èxits ocorreguts a l'interval", de la qual coneixem la seua mitjana, λ. En principi, el nombre pot ser qualsevol enter no negatiu. Aleshores $X \in \{0, 1, 2, \ldots\}$.*

2. *Tallem l'interval inicial en n trossos d'igual llargària, agafant un valor $n \in \mathbb{N}$ molt gran, perquè els trossos de l'interval siguen molt xicotets.*

3. *Per a cadascú dels n trossos, podem definir la variable X_n^j = "Nombre d'èxits observats al j-èssim tros".*

4. *Abusant de **(Po4)**, ja que els trossos resultants són molt xicotets (encara que no de llargària zero), quins valors pot prendre la variable X_n^j?*

$$X_n^j \in \{ \quad \}$$

5. *Quina és la mitjana de X_n^j per comparació amb la mitjana a tot l'interval sencer?*

$$\mathbb{E}(X_n^j) =$$

6. *Usant els resultats de 3–5, quin model de variable aleatòria, ja introduït anteriorment, segueix la variable X_n^j?*

$$X_n^j \sim$$

7. *Si definim X_n = "nombre d'èxits ocorreguts a l'interval original", juntant els èxits observat a cada tros per separat, aleshores podrem escriure $X_n = \sum_{j=1}^{n} X_n^j$. Usant **(Po5)** (és a dir, la independència), quina distribució segueix la variable X_n?*

$$X_n \sim$$

8. *Calculeu doncs, per a $k = 1, 2, \ldots, n$,*

$$P(X_n = k) = f(k) =$$

9. *A l'apartat 4 hem abusat de la propietat **(Po4)**, ja que els intervals eren xicotets, però no infinitesimalment. Ara anem a fer justícia. Anem a prendre el límit quan $n \to \infty$ perquè els trossos siguen de llargària infinitesimal. Definim X_∞ = "nombre d'èxits ocorreguts a l'interval" com a límit en cert sentit de les variables X_n quan n creix a l'infinit. Calculeu, usant tècniques bàsiques,*

$$P(X_\infty = k) = \lim_{n\to\infty} P(X_n = k) =$$

La manera descrita a l'exercici anterior representa una forma d'arribar a la distribució de Poisson, encara que no l'única. Passem a la definició del model.

Definició 7.6.1 *El model de **Poisson de paràmetre** λ ve definit per l'experiment i té les propietats que es mostren a continuació:*

- ***Experiment:*** *Un esdeveniment etiquetat com* ***èxit*** *pot passar en cada punt d'un interval de temps o espai prèviament fixat, amb les propietats etiquetades com* ***(Po1)–(Po5)*** *de la pàgina 127. A més, es coneix que l'èxit té una mitjana de λ aparicions a l'interval. Es va a observar en quantes ocasions apareix l'èxit durant aqueix interval (o qualsevol altre de la mateixa llargària).*

- ***Variable aleatòria:*** *X = "**nombre d'èxits observats en un interval d'aqueixa llargària**"*

- ***Espai mostral:*** $X \in \{0, 1, 2, 3, \ldots\}$

- ***Notació:*** $X \sim \mathrm{Po}(\lambda)$

- $\boldsymbol{f(x)} = \begin{cases} \mathrm{e}^{-\lambda} \cdot \dfrac{\lambda^x}{x!}, & x = 0, 1, 2, 3, \ldots \\ 0, & \textit{altre cas} \end{cases}$

- $\boldsymbol{F(x)}$ *No la gastarem al llarg del llibre (vegeu com s'obté a partir de f en pàgina 107)*

- ***Esperança i Variància:*** $\mathbb{E}(X) = \lambda$, $\mathbb{V}(X) = \lambda$

En general, les propietats **(Po1)**–**(Po5)** són difícilment comprovables en cada situació real, i se suposaran certes generalment, com a l'exercici següent.

Exercici 7.6.2 *Un servidor ftp rep una mitjana de 3.5 peticions per minut. Calculeu:*

- *Expectativa de peticions per al pròxim minut.*

- *Probabilitat de cap petició al següent minut?*

  $\Pr(X \quad) =$

- *Probabilitat de 2 o més peticions al següent minut?*

  $\Pr(X \quad) =$

- *Probabilitat d'entre 2 i 5 peticions al següent minut?*

  $\Pr(\quad X \quad) =$

- *Probabilitat de rebre més de 200 peticions durant la següent hora?*

  $\Pr(X \quad) =$

Exercici 7.6.3 *Deduïu les fórmules d'esperança i variància a partir de les fórmules de la pàgina 114.*

Propietat 7.6.1 *Si $X_1 \sim \mathrm{Po}(\lambda_1)$ i $X_2 \sim \mathrm{Po}(\lambda_2)$ són independents i $Y = X_1 + X_2$, aleshores $Y \sim \mathrm{Po}(\lambda_1 + \lambda_2)$.*

Propietat 7.6.2 *Si* $X \sim \text{Bin}(n, p)$*,* n *és "gran" i* p *"menut" (per exemple* $n > 30$*,* $p < 0.1$ *i* $np < 10$*), i* $Y \sim \text{Po}(\lambda = np)$*, aleshores* X *i* Y *tenen distribucions molt semblants (raó per la qual es podria usar una per a fer càlculs aproximats de l'altra).*

La demostració de la Propietat 7.6.2 és justament l'Exercici 7.6.1

Exemple 7.6.1 *Aproximadament un de cada* 4500 *discos durs ix defectuós de fàbrica. Tenim un lot de* 10000 *discos durs, del qual interessa analitzar la possible quantitat d'unitats defectuoses. Usant la variable aleatòria*

$X =$ *"....."*, $X \sim$

calculeu:

- *Expectativa del nombre de discos defectuosos.*
- *Probabilitat de cap disc defectuós?*
  $\Pr(X \quad) =$
- *Probabilitat d'un sol disc defectuosos?*
  $\Pr(X \quad) =$
- *Probabilitat de* 2 *o més discos defectuosos?*
  $\Pr(X \quad) =$

7.7 Uniforme a l'interval (a, b)

En ocasions, el valor resultant d'un experiment no és una quantitat de comptar (0, 1, 2, ...), però un valor mesurat sobre una escala contínua (interval) de nombres com, per exemple, el temps o les distàncies.

En aquestes ocasions la variable aleatòria que es fa servir és contínua, i els esdeveniments del tipus $\{X = x\}$ deixen de tenir sentit, ja que tenen probabilitat nul·la segons el model matemàtic de probabilitat. Per tant, l'ús de la funció f queda relegat, i només es fa servir la funció F.

El model més senzill de variable aleatòria contínua és aquell relacionat amb un experiment on l'espai mostral (de resultats possibles) es correspon amb un interval de llargària finita i on tots els valors de l'interval són igualment elegibles.

Definició 7.7.1 *El model* ***uniforme a l'interval*** (a, b) *ve definit per l'experiment i té les propietats que es mostren a continuació:*

- ***Experiment:*** *Un nombre d'un interval* (a, b) *és triat completament a l'atzar, on tots els valors són igualment versemblants, i es va a observar el valor triat.*
- ***Variable aleatòria:*** $X =$ ***"valor triat"***

- ***Espai mostral:*** $X \in (a, b)$
- ***Notació:*** $X \sim \text{Unif}(a, b)$
- $\boldsymbol{f(x)} = \begin{cases} \dfrac{1}{b-a}, & x \in (a, b) \\ 0, & \textit{altre cas} \end{cases}$
- $\boldsymbol{F(x)} = \begin{cases} 0, & x \leq a \\ \dfrac{x-a}{b-a}, & x \in (a, b) \\ 1, & x \geq b \end{cases}$
- ***Esperança i Variància:*** $\mathbb{E}(X) = \dfrac{a+b}{2}$, $\mathbb{V}(X) = \dfrac{(b-a)^2}{12}$

Exercici 7.7.1 *Un virus informàtic tria completament a l'atzar una hora de cada dia (des de les 00:00:00 fins les 23:59:59.999...) i esborra un arxiu en aqueix moment.*

- *Definiu la variable aleatòria*

  $$X = \text{“.....”}$$

  de manera que s'adeqüe al perfil d'una variable aleatòria amb distribució uniforme a l'interval (,).

- *Calculeu la probabilitat (o el percentatge d'ocasions a llarg termini) que l'arxiu triat siga eliminat en horari d'oficina (entre les 9 h i les 17 h).*

- *Si no s'elimina el virus en tot un any, aproximadament quants arxius serien eliminats en la primera hora de cada dia?*

El model uniforme és tan senzill que normalment no són necessàries les fórmules per a calcular probabilitats: la probabilitat d'un interval (situat dins de l'interval on agafa els valors la uniforme) és igual a la porció que representa en l'interval total.

Exercici 7.7.2 *Deduïu les fórmules d'esperança i variància a partir de les fórmules de la p. 114.*

7.8 Exponencial de paràmetre λ

Un procés de Poisson és un mecanisme que transcorre al llarg del temps o l'espai, donant lloc a observacions puntuals sota unes condicions particulars (vegeu la pàgina 127).

Al model de Poisson de la Secció 7.6, es fixava un interval de temps o espai concret, i es registrava el nombre d'observacions ocorregudes dins de l'interval.

Un segon enfocament del mateix procés és fixar un punt (no un interval) en el temps o l'espai, esperar a trobar la següent observació de l'esdeveniment,

i registrar aleshores l'interval de temps o espai transcorregut. Aquest valor registrat ja no és una quantitat de comptar, però un valor d'un interval continu, per tant estem considerant una variable aleatòria contínua.

Definició 7.8.1 *El model* ***exponencial de paràmetre*** λ *ve definit per l'experiment i té les propietats que es mostren a continuació:*

- ***Experiment:*** *Un procés de Poisson de paràmetre* λ *ocorre en l'espai o el temps. Fixem un punt qualsevol de l'espai o un instant qualsevol del temps, o bé un punt on s'ha observat una aparició de l'èxit. Ara s'observa l'interval d'espai o temps transcorregut fins a una nova aparició de l'èxit.*
- ***Variable aleatòria:*** $X =$ *"****Espai o temps transcorregut fins a la nova aparició de l'èxit (o entre 2 aparicions consecutives)****".*
- ***Espai mostral:*** $X \in (0, \infty)$
- ***Notació:*** $X \sim \text{Exp}(\lambda)$
- $\boldsymbol{f(x)} = \begin{cases} 0, & x \leq 0 \\ \lambda \text{e}^{-\lambda x}, & x > 0 \end{cases}$
- $\boldsymbol{F(x)} = \begin{cases} 0, & x \leq 0 \\ 1 - \text{e}^{-\lambda x}, & x > 0 \end{cases}$
- ***Esperança i Variància:*** $\mathbb{E}(X) = \dfrac{1}{\lambda}$, $\mathbb{V}(X) = \dfrac{1}{\lambda^2}$
- ***Nota:*** *El paràmetre* λ *fa referència al procés de Poisson, i no és la mitjana de la variable aleatòria, que val* $1/\lambda$*. Per tant és molt important atribuir el valor correcte al paràmetre i no confondre'l amb la mitjana.*

Exercici 7.8.1 *Un servidor rep una mitjana de 6.5 peticions per minut. Calculeu:*

- *La variable* $X =$*"Temps fins a la pròxima petició (en minuts)" es distribueix com* $X \sim \text{Exp}(\lambda =$ $)$
- *Temps mitjà que transcorre entre dues peticions consecutives?*
- *Probabilitat que es reba una petició abans del 5 segons següents?*
- *Probabilitat que passe 1 minut sense rebre cap nova petició?*

Exercici 7.8.2 *En analitzar la durabilitat d'un dispositiu, es fa un estudi on s'observa que el temps mitjà de duració dels aparells és d'1.35 anys.*

- *La variable* $X =$*"durada del dispositiu (en anys)" pot considerar-se que es distribueix com* $X \sim \text{Exp}(\lambda =$ $)$*.*
- *Probabilitat que un dispositiu qualsevol dure més de 2 anys?*
- *Probabilitat que un dispositiu qualsevol dure menys d' un mes?*

- *Si la garantia dels dispositius és de 3 anys, quin percentatge aproximat d'aparells es tornaran en el període de garantia?*
- *Si el fabricant del dispositiu vol fixar un nou període de garantia, de manera que aproximadament el 5% dels dispositius fallen dins del període, com l'hauria de fixar?*

Exercici 7.8.3 *Si X segueix el model exponencial de paràmetre λ i indica el temps fins l'aparició d'un fenòmen observable, demostra que les probabilitats de duració futura són independents de les duracions mínimes constatades. Expressat tècnicament,*

$$P(X > x + h | X > x) = P(X > h)$$

per qualsevol $x, h > 0$. Es tracta de la propietat anomenada "manca de memòria" (a la distribució de X no li afecta saber que X ja val més que x).

Exercici 7.8.4 *Deduïu les fórmules d'esperança i variància a partir de les fórmules de la p. 114.*

7.9 Erlang de paràmetres λ i r

El model que es presenta en aquesta secció complementa el model exponencial.

Definició 7.9.1 *El model **Erlang de paràmetres** r **i** λ ve definit per l'experiment i té les propietats que es mostren a continuació:*

- ***Experiment:** Un procés de Poisson de paràmetre λ ocorre en l'espai o el temps. D'altra banda, prefixem un nombre d'èxits r. Fixem un punt qualsevol de l'espai o un instant qualsevol del temps, o bé un punt on s'ha observat una aparició de l'èxit. Ara s'observa l'interval d'espai o temps transcorregut fins a la nova aparició del r-èssim èxit.*
- ***Variable aleatòria:** X = "**Espai o temps transcorregut fins a la nova aparició del r-èssim èxit (o entre $r+1$ aparicions consecutives)**".*
- ***Espai mostral:** $X \in (0, \infty)$*
- ***Notació:** $X \sim \mathrm{Erl}(r, \lambda)$*
- $\boldsymbol{f(x)} = \begin{cases} 0, & x \leq 0 \\ \dfrac{\lambda^r x^{r-1} \mathrm{e}^{-\lambda x}}{(r-1)!}, & x > 0 \end{cases}$
- $\boldsymbol{F(x)}$ *No la gastarem al llarg del llibre (vegeu com s'obté a partir de f en pàgina 109)*
- ***Esperança i Variància:** $\mathbb{E}(X) = \dfrac{r}{\lambda}$, $\mathbb{V}(X) = \dfrac{r}{\lambda^2}$*

- ***Nota:*** *$X = \sum_{i=1}^{n} X_i$ on cada $X_i \sim \text{Exp}(\lambda)$ i són totes independents.*

Exercici 7.9.1 *Un node d'una xarxa informàtica rep i emet paquets d'informació, de manera que els van rebent a raó de* 100 *paquets cada* 2.76 *segons en mitjana, i retransmet els paquets en lots de 10 paquets (és a dir, espera a rebre 10 paquets per a fer un enviament amb tots junts). Calculeu:*

- *La variable aleatòria $X =$"Temps des d'una recepció a una altra" es distribueix com $X \sim \quad (\lambda = \quad)$.*
- *La variable aleatòria $Y =$"Temps des d'una emisió a una altra" es distribueix com $Y \sim \text{Erl}(r = \quad, \lambda = \quad)$.*
- *Temps mitjà des d'una emisió a una altra?*
- *Probabilitat que el temps des d'una emisió a la següent siga de menys de* 0.5 *segons?*
- *Probabilitat que passe més d'un segon sense enviar informació?*

Exercici 7.9.2 *Deduïu les fórmules d'esperança i variància a partir de les fórmules de la p. 114.*

7.10 Normal o Gaussiana de paràmetres μ i σ^2

El model que es presenta en aquesta secció és d'una importància màxima, i ocupa un lloc central, entre tots els altres models.

7.10.1 Definició

- **Experiment:** En la natura o en processos industrials existeix una variabilitat natural i inevitable (o un error inherent en els instruments de mesura). Ens fixem en una característica a mesurar.
- **Variable aleatòria:** $X =$ "**Valor observat o mesurat**".
- **Espai mostral:** $X \in (-\infty, +\infty)$
- **Notació:** $X \sim \text{N}(\mu, \sigma^2)$
- $\boldsymbol{f(x)} = \dfrac{1}{\sqrt{2\pi\sigma^2}} \text{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- $\boldsymbol{F(x)}$ No té expressió analítica (vegeu com s'obté a partir de f en la p. 109)
- **Esperança i Variància:** $\mathbb{E}(X) = \mu$, $\mathbb{V}(X) = \sigma^2$

La funció de densitat de la normal té un aspecte que depén lògicament dels paràmetres μ i σ^2, tal i com es pot observar a la Figura 7.1. Atorga més versemblança als valors pròxims a la mitjana μ, i la versemblança va baixant (més o menys ràpid segons la dispersió donada per σ) a mesura que els valors s'allunyen de la mitjana.

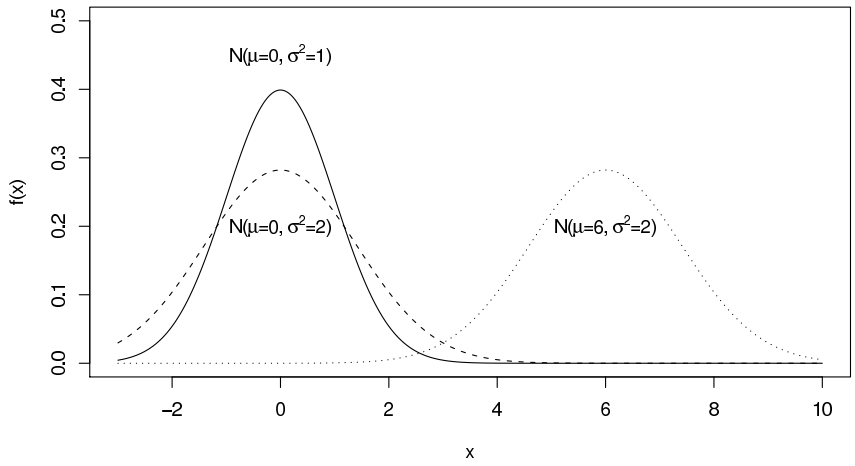


Figura 7.1: Funcions de densitat de probabilitat de la distribució normal per a les tres parelles de paràmetres indicades.

7.10.2 Propietats

Propietat 7.10.1 *Sobre la transformació lineal i combinacions es tenen les següents propietats:*

1. *Si* $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ *i definim una nova variable* $Y = a + bX$*, on* $a, b \in \mathbb{R}$*, aleshores* $Y \sim \mathrm{N}(\mu = a + b\mu_X, \sigma^2 = b^2\sigma_X^2)$*.*
2. *Si* $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ *i* $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ *i definim una nova variable* $W = X + Y$ *aleshores* $W \sim \mathrm{N}(\mu = \mu_X + \mu_Y, \sigma^2 = \sigma_X^2 + \sigma_Y^2)$*.*

Exercici 7.10.1 *Els temps dels* 4 *corredors de relleus* 4×100 *m. es distribueix normalment amb mitjanes* $11.3, 11.1, 10.6, 10.0$ *s i variàncies* 0.04*,* 0.02*,* 0.01*,* 0.01 s^2*, respectivament. Quina és la distribució exacta de* $X =$ *"temps total de la carrera"?*

Propietat 7.10.2 (Tipificació) *Si* $X \sim \mathrm{N}(\mu, \sigma^2)$ *i es defineix la variable* $Z = \dfrac{X - \mu}{\sigma}$*, aleshores* $Z \sim \mathrm{N}(\mu = 0, \sigma^2 = 1)$*, i es diu* ***normal tipificada*** *o* ***estándar****.*

Exemple 7.10.1 (Càlcul amb les taules) *Només hi ha taula de probabilitats per a* Z *normal tipificada. Per tant, si* $X \sim \mathrm{N}(\mu = 10, \sigma^2 = 2)$ *i es demana* $\Pr(X \leq 12)$*:*

$$\begin{aligned}\Pr(X \leq 12) &= \Pr\left(\frac{X - 10}{\sqrt{2}} \leq \frac{12 - 10}{\sqrt{2}}\right) \\ &= \Pr(Z \leq 1.41) = F_Z(1.41) = \textit{(taules)} = 0.9207\end{aligned}$$

7.10.3 Teorema del límit central

Propietat 7.10.3 (Teorema del límit central) *Si* $X_1, X_2, \ldots, X_n$ *són variables aleatòries independents, de qualsevol distribució, i amb variàncies fini-*

tes, aleshores, quan n és gran ($n \to \infty$)

$$\sum_{i=1}^{n} X_i \sim_{aprox.} \mathrm{N}(\mu = \sum_{i=1}^{n} \mathbb{E}(X_i), \sigma^2 = \sum_{i=1}^{n} \mathbb{V}(X_i))$$

Exercici 7.10.2 *Uns taulells de ceràmica es fabriquen d'una superficie mitjana de* $3.15\ cm^2$ *i variància* $0.1\ cm^4$. *Si comprem una caixa de* 50000 *taulells:*

- *Si X = "Superfície total dels taulells", $X \sim$ .....*
- *Superficie total esperada?*
- *Probabilitat que es puga cobrir una àrea de* $15.77\ m^2$*?*

Propietat 7.10.4 (Relació Normal amb Binomial i Poisson) *La distribució normal és molt útil per a aproximar altres distribucions en casos particulars:*

- *Si $X \sim \mathrm{Bin}(n, p)$ i n es prou gran, aleshores $X \sim_{aprox.} \mathrm{N}(\mu = np, \sigma^2 = np(1-p))$. Per exemple, una condició objectiva sobre n i p generalment acceptada és $0 < np - 3\sqrt{np(1-p)} < np + 3\sqrt{np(1-p)} < n$*
- *Si $X \sim \mathrm{Po}(\lambda)$, i λ es prou gran, aleshores $X \sim_{aprox.} \mathrm{N}(\mu = \lambda, \sigma^2 = \lambda)$. Per exemple, $\lambda > 10$ és un valor generalment acceptat.*

Exercici 7.10.3 *El* 10% *de les reserves de bitllets d'avió són cancel·lades abans dels* 5 *primers dies. Si avui han gestionat* 347 *reserves:*

- *Si X = "Nombre de cancel·lacions realitzades abans de* 5 *dies", aleshores $X \sim$ .....*
- *Nombre de cancel·lacions esperades durant els* 5 *dies?*
- *Probabilitat que no es produeixca cap cancel·lació?*
- *Probabilitat que es produïsquen més de 5 cancel·lacions?*

Nota 7.10.1 (Correcció per continuitat) *Si X és una variable aleatòria discreta (p. e. binomial o Poisson) i per qualsevol motiu es considera contínua (p. e., si aproximem amb la normal), recorda que:*

$$\Pr(X = x) = \begin{cases} f(x) & \text{si } X \text{ és variable aleatòria discreta} \\ 0 & \text{si } X \text{ és variable aleatòria contínua} \end{cases}$$

Per tant, per a calcular probabilitats cal identificar cada valor discret amb un interval (obert o tancat, és igual, vegeu la Taula 7.1).

Taula 7.1: Exemples d'identificacions entre esdeveniments en la variable discreta quan s'usa un model continu per calcular aproximadament les seues probabilitats

Variable aleatòria discreta	Variable aleatòria contínua
$\Pr(X = x)$	$\Pr(x - 0.5 < X < x + 0.5)$
$\Pr(X = 3)$	$\Pr(2.5 < X < 3.5)$
$\Pr(X = 0)$	$\Pr(-0.5 < X < 0.5)$
$\Pr(5 < X \leq 7)$	$\Pr(5.5 < X < 7.5)$

7.11 Exercicis proposats

Exercici 7.11.1 *Segons estudis astronòmics, cada dia arriba al planeta Terra una mitjana de* 7.5 *meteorits de gran calibre. Calculeu:*

1. *Quina és la probabilitat que un dia com demà arriben* 10 *metorits o més?*
2. *Quina és la probabilitat que dos metorits arriben a la Terra amb una separació inferior a 60 minuts?*

Exercici 7.11.2 *El temps que pren una impressora a imprimir una pàgina és variable i es pot suposar completament aleatori entre* 1.0 *i* 3.4 *segons. Calculeu:*

1. *Si s'envia un document d'una pàgina a l'impressora, amb quina probabilitat serà imprés en menys de* 2.0 *segons?*
2. *Si s'envia un document de* 53 *pàgines, amb quina probabilitat serà imprés en menys de* 2.0 *minuts? (*Atenció a les unitats!*)*

Exercici 7.11.3 *Un cable de fibra òptica té una mitjana de* 3.75 *defectes per quilòmetre. Calculeu:*

1. *Si compreu* 2*km de cable, quina és la probabilitat que hi trobeu* 10 *defectes o més?*
2. *Quina és la probabilitat que dos defectes consecutius estiguen separats per menys de* 83.3 *m? (*Atenció a les unitats!*)*

Exercici 7.11.4 *Un procés de fabricació està dividit en quatre subprocesos consecutius. El temps de fabricació que cada subprocés pren per finalitzar la seua part es pot modelitzar com una normal de mitjana* 13.6 *segons i desviació típica* 1.6 *segons.*

1. *Quina és la probabilitat que el tercer subprocés tarde més de* 15.0 *segons?*
2. *Quina és la probabilitat que el procés complet tarde més d'un minut sencer?*

Exercici 7.11.5 *Un venedor (que ven a través del telèfon) estima que la probabilitat que una telefonada acabe sent una venda és de* 0.05*. El seu objectiu diari és realitzar* 3 *vendes, i acaba la seua jornada de treball, siga l'hora que siga. Calculeu:*

1. *Nombre mitjà (a llarg termini) de telefonades fetes per dia.*

2. *Probabilitat d'acabar la jornada laboral amb exactament 50 telefonades.*

Exercici 7.11.6 *Una empresa registra les incidències ocorregudes diàriament. Atenent només el nombre d'incidències diàries es comprova que hi ha hagut una mitjana de* 3.71 *incidències diàries en la història de l'empresa. Calculeu:*

1. *La probabilitat que el dia següent no hi haja cap incidència.*

2. *Si la fàbrica du* 3927 *dies en funcionament, en quants d'aquests (aproximadament) no es va produir cap incident?*

Exercici 7.11.7 *Una oposició consta d'un examen escrit en el qual es desenvolupa una pregunta, d'una llista de* 71 *preguntes possibles que formen el temari.*

En arribar a l'examen, l'opositor ha de triar 3 *boles (numerades de l'*1 *al* 71*) d'una urna, i el seu examen consistirà en una de les* 3 *preguntes indicades per les boles (la que vulga).*

Suposant que l'opositor sap perfectament 30 *de les* 71 *preguntes, i desconeix absolutament les altres* 41 *preguntes, quina probabilitat exacta té de poder respondre l'examen?*

Exercici 7.11.8 *Una prova tipus test consta de* 10 *preguntes amb* 4 *alternatives de solució per pregunta. Una persona contesta a l'atzar totes les preguntes. Calculeu:*

1. *Nombre esperat de preguntes correctes.*

2. *Probabilitat d'aprovar la prova (és a dir, d'obtenir* 5 *o més respostes correctes).*

Exercici 7.11.9 *Un proveïdor de màquines ofereix dos models de màquina (M_1 i M_2) als seus clients, que les necessiten per a fabricar peces d'una longitud teòrica de 150 mm.*

El client C_1 admet com a bones peces amb longitud entre 148.5 *i* 151.5 *mm, mentre que el client C_2 és més exigent i només admet com a bones les peces la longitud de les quals estiga entre* 148.8 *i* 151.2 *mm.*

Un estudi realitzat sobre la qualitat de les màquines revela que la màquina M_1 fa peces amb longitud completament arbitrària dins de l'interval [148.3, 151.7]. *La màquina M_2 realitza les seues peces amb longitud normalment distribuïda de mitjana* 150 *mm i desviació típica* 1.0 *mm.*

Fes els càlculs necessaris perquè cada client trie la màquina que més li convinga (pot ser la mateixa o diferent).

Exercici 7.11.10 *El cor d'un sistema informàtic posseïx* 10 *arxius, dels quals* 3 *són fonamentals, i si resulta esborrat qualsevol d'aquests, el sistema no podria recuperar-se. Quan penetra un virus (que esborra arxius a l'atzar) en aquest cor, l'antivirus ho detecta i l'elimina quan el virus ha esborrat el segon arxiu. Calculeu la probabilitat que un virus que penetre, aconseguisca que el sistema siga irrecuperable.*

Exercici 7.11.11 *Quan cau la xarxa elèctrica en cert edifici, el temps que es roman sense subministrament és variable i es pot modelitzar com una distribució exponencial, de manera que el temps mitjà sense servei és d'aproximandament* 13.3 *minuts.*

Un treballador d'aquest edifici vol adquirir un UPS amb 30 minuts de bateria. Quina probabilitat té que un dia que falla el subministrament, l'UPS li siga insuficient i perda el treball que tinga en curs en el seu ordinador?

Exercici 7.11.12 *En una cua d'un servidor, el temps (en segons) emprat per cada client en rebre el servei i eixir de la cua es distribueix segons la llei de "Ujiprob" de paràmetres* $a = 1$ *i* $b = 2$*. Un client ocupa el lloc* 37 *de la cua, i no desitja esperar si la probabilitat que tarde més de* 140 *s a rebre el seu servei és superior o igual a* 0.1*. Quin càlcul aproximat li permet prendre una decisió i quina seria?*

(Ajuda: si X *segueix la distribució "Ujiprob" de paràmetres* a *i* b*, aleshores* $\mathbb{E}(X) = a + b$ *i* $\mathbb{V}(X) = a/b^2$*)*

Exercici 7.11.13 *El nombre de peticions que arriben a un servidor web se sol modelitzar com a variable aleatòria de Poisson. Suposant que hi ha una mitjana de 10 peticions per hora:*

- *Quina és la probabilitat que es reben exactament 5 peticions en la següent hora?*
- *Quina és la probabilitat que es reben 3 o menys peticions en la següent hora?*
- *Quina és la probabilitat que es reben exactament 5 peticions en les següents dues hores?*
- *Quina és la probabilitat que es reben exactament 5 peticions en els següents 30 minuts?*

Exercici 7.11.14 *La resistència de mostres de ciment pot ser modelitzada per una distribució normal de mitjana 6000* Kg/cm^2 *i desviació típica de 100* Kg/cm^2*.*

- *Amb quina probabilitat, una mostra tindrà una resistència inferior als 6250* Kg/cm^2*?*
- *Amb quina probabilitat, una mostra tindrà una resistència entre 5800 i 5900* Kg/cm^2*?*
- *El 95% de la població de mostres de ciment supera una resistència de ...* Kg/cm^2*?*

Exercici 7.11.15 *El primer semàfor que trobeu en eixir cap a la Universitat (quan agafeu el cotxe) fa cicles de 15 segons amb verd i 35 segons amb roig. Si cada matí eixiu a una hora independent de la resta dels dies, calculeu:*

- *Sobre els 5 pròxims dies, quina és la probabilitat de trobar-lo verd exactament un dia?*

- *Sobre els 20 pròxims dies, quina és la probabilitat de trobar-lo verd exactament 4 dies?*

- *Sobre els 20 pròxims dies, quina és la probabilitat de trobar-lo verd més de 4 dies?*

Exercici 7.11.16 *El temps transcorregut entre dues telefonades consecutives al servei "112" es distribueix com una variable aleatòria exponencial, amb un temps mitjà entre telefonades de 13 s.*

- *Quina és la probabilitat que passe més de mig minut sense rebre cap telefonada?*

- *Quina és la probabilitat que es reba almenys una telefonada en un interval de 10 s?*

- *Quina és la probabilitat que la primera telefonada del dia siga abans de les 00:00:15? (és un servei de 24 hores)*

Exercici 7.11.17 *Un dispositiu electrònic situat en una cadena d'ompliment deté la línia de producció quan detecta tres paquets d'un pes inferior al tolerat. Suposant que la probabilitat d'un paquet de pes inferior al tolerat és 0.01 i que cada ompliment és independent de la resta:*

- *Quina és la probabilitat que es detinga justament en acabar d'omplir el paquet número 100?*

- *Quina és la probabilitat que es detinga abans d'omplir 100 paquets?*

- *Quina és la probabilitat d'haver omplit més de 150 paquets abans de parar-se?*

- *Si cada vegada que es deté el procés es registra el nombre de paquets omplits, i es repeteix el procés un gran nombre de vegades, quin seria el valor mitjà dels valors registrats?*

Exercici 7.11.18 *El gruix d'un tipus de peça usada per construir avions és un valor completament aleatori entre 0.95 i 1.05 mm.*

- *Quin percentatge de les peces excedeix els 1.02 mm de gruix?*

- *Quin és el gruix màxim del 90% de les peces més fines?*

Exercici 7.11.19 *Un test de funcionament consisteix a comprovar 20 aparells diferents triats a l'atzar d'un lot de 140.*

- *Si el lot té 20 aparells defectuosos, quina és la probabilitat que aparega algun d'aquests en la mostra?*

- *Si el lot té 5 aparells defectuosos, quina és la probabilitat que aparega algun d'aquests en la mostra?*

Exercici 7.11.20 *El temps de reacció d'un conductor a un estímul visual es distribueix com una normal amb mitjana 0.4 s i desviació típica 0.05 s.*

- *Quin percentatge aproximat de vegades, el temps de reacció dels conductors és de més de 0.5 s?*
- *Quin l'interval conté el 95% dels temps de reacció més "normals" (és a dir, exceptuant el 2.5% dels temps més curts i l'altre 2.5% dels temps més llargs)?*

Exercici 7.11.21 *En un sistema de comunicació de dades, els missatges que arriben a un node es lliguen en un paquet (de 5 missatges) abans de ser transmesos a la xarxa. Si els missatges arriben al node segons una llei de Poisson de 30 missatges per minut:*

- *Quina és la probabilitat que un paquet es forme en menys de 10 s?*
- *Quina és la probabilitat que un paquet es forme en menys de 5 s?*

Exercici 7.11.22 *La vida d'un làser semiconductor funcionant a potència constant es distribueix normalment amb mitjana de 7000 hores i desviació típica de 600 hores.*

- *Quina és la probabilitat que el làser falle abans de 5000 hores d'ús?*
- *Quina és la durada (en hores) que supera el 95% dels làsers d'aquest tipus?*
- *Si tres làsers formen part d'un aparell, i se suposa que fallen independentment, quina és la probabilitat que després 7000 hores n'hi haja algun làser funcionant?*

Exercici 7.11.23 *Un canal, pel qual es transmeten bits d'un emissor a un receptor, té la propietat d'invertir el 10% dels bits que el travessen, de forma que són rebuts incorrectament. Si s'envia un missatge de només 8 bits, calculeu:*

- *Probabilitat que el missatge arribe íntegre (sense errors).*
- *Probabilitat que el missatge rebut porte menys de 4 errors.*
- *Si el missatge s'envia des de l'emissor al receptor repetidament, en quin percentatge aproximat de les ocasions el missatge porta un error com a màxim?*

Exercici 7.11.24 *Un dispositiu electrònic consta de 40 circuits integrats. La probabilitat que qualsevol circuit integrat siga defectuós és 0.01, i els circuits integrats són independents. El dispositiu funciona només si no té circuits integrats defectuosos. Quina és la probabilitat que el dispositiu funcione?*

Exercici 7.11.25 *Una companyia aèria, conscient de les cancel·lacions d'última hora, estudia la possibilitat d'infringir la llei oferint més de 300 reserves, per a un vol d'un avió amb només 300 places.*

Si la companyia pensa que els passatgers es comporten independentment i segons un estudi, cancel·len un 8% dels que reserven, quin nombre de reserves màxim podria oferir la companyia, per que la probabilitat d'incòrrer en overbooking siga inferior al 5%?

Nota 1: la situació d'overbooking es dóna quan es presenten més persones amb reserva que seients hi ha disponibles per al viatje.

Nota 2: la formació universitària ha de ser una formació intel·lectual, professional i ètica. Els autors desaproven l'ús de l'Estadística per ajudar a la realització de pràctiques il·legals com la d'aquest problema.

Exercici 7.11.26 *Un lot de 75 lectors de DVD en conté 5 que tenen un funcionament inacceptable. Una mostra de 10 lectors se selecciona a l'atzar, sense reemplaçament:*

- *Quina és la probabilitat que cap dels lectors inacceptables forme part de la mostra?*
- *Quina és la probabilitat que hi haja algun lector inacceptable en la mostra?*
- *Quina és la probabilitat que hi haja exactament un lector inacceptable en la mostra?*
- *Quin és el nombre esperat de lectors inacceptables en fer la selecció?*

7.12 Pràctica R: 7. Càlcul de probabilitats en models coneguts

Objectius

El programa R té implementades les funcions de probabilitat, densitat de probabilitat (denotades comunament per f) i distribució acumulada (denotada comunament per F) de les distribucions discretes i contínues més conegudes. Per tant, els problemes de càlcul de probabilitats es poden resoldre de manera natural amb l'ordinador, sense les imcompletes calculadores i les farragoses taules.

Tanmateix, R és una potent calculadora especialitzada en funcions estadístiques, però és l'usuari el que ha de tenir els coneixements teòrics necessaris per plantejar i resoldre els problemes.

D'altra banda, R també disposa d'un algorisme generador de nombres aleatoris, base de la simulació numèrica d'experiments.

Alguns models programats en R

El següent llistat mostra els models de variable aleatòria implementats en R i els noms atribuïts als seus paràmetres:

Distribució	nom R	arguments addicionals
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-quadrat	chisq	df, ncp
exponencial	exp	rate
F	f	df1, df1, ncp
gamma	gamma	shape, scale
geomètrica	geom	prob
hipergeomètrica	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logística	logis	location, scale
binomial negativa	nbinom	size, prob
normal (Gaussiana)	norm	mean, sd
Poisson	pois	lambda
t de Student	t	df, ncp
uniforme	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

Principals funcions de les variables aleatòries

Per a calcular probabilitats relatives a les variables aleatòries amb distribucions implementades a R, una sèrie de funcions són de gran utilitat. En aquesta secció denotarem per:

- `%%%%%` el nom R de la distribució que siga (i que és qualsevol dels que figura al llistat anterior a la columna `nom R`).
- `...` els arguments addicionals corresponents a la distribució (que també figuren al llistat anterior).

Cada funció que anem a presentar es compon d'un prefix i del nom de la distribució.

La funció f: de quantia (en variable discreta) o de densitat (en variable contínua)

La definició de funció f és:

$$f(x) = \begin{cases} P(X = x), & \text{si } X \text{ és v. a. discreta} \\ \text{Versemblança de } x, & \text{si } X \text{ és v. a. contínua} \end{cases}$$

En ambdós casos, la funció f es calcula amb R com:

$$f(x) = \texttt{d\%\%\%\%\%(x=}x\texttt{, ...)}$$

La funció F: de distribució acumulada

Tant si la variable aleatòria X és discreta com si és contínua, la definició de funció de distribució és $F(x) = P(X \leq x)$. En R:

$$F(x) := \texttt{p\%\%\%\%\%(q=}x\texttt{, ...)}$$

La funció de quantils F^{-1}

En una variable aleatòria X, el quantil d'ordre p es denota per x_p i és el menor valor possible que deixa per davall una probabilitat d'almenys p, és a dir, $P(X \leq x_p) \geq p$. En variable continua és el que compleix la igualtat $F(x_p) = p$, i per tant, el que compleix la igualtat $x_p = F^{-1}(p)$. Així doncs, la funció inversa de la funció de distribució acumulada s'anomena funció de quantils, i es calcula amb R com:

$$x_p = \texttt{q\%\%\%\%\%(p=}p\texttt{, ...)}$$

Simulació de nombres aleatoris

Per simular una quantitat n de valors que pertanyen a una variable aleatòria X cal usar la funció:

$$\texttt{r\%\%\%\%\%(n=}\ n\ \texttt{, ...)}$$

El resultat és un vector que emmagatzema els n valors aleatoris demanats. Els valors són pseudoaleatoris (no aleatoris del tot) perquè es generen amb una fórmula recurrent.

Models estudiats a la teoria vs models programats en R

Els models de variable aleatòria es poden definir de maneres distintes (encara que similars). Per això és molt important conéixer com estan programats en R, perquè les parametritzacions poden ser diferents.

Escrivint `help(d%%%%%)`, o qualsevol altre prefixe, obtenim l'ajuda que R dóna sobre el model `%%%%%`. Entre altres, la informació d'ajuda inclou el significat de la variable aleatòria i dels seus paràmetres.

Així, l'usuari que modelitza un problema amb una variable aleatòria i fa servir directament el plantejament de R, no ha de fer res més que aplicar les fórmules necessàries. Mentrestant, un usuari que usa una parametrització d'una variable aleatòria que no coincideix amb la que està implementada a R ha de fer algunes modificacions perquè el programa calcule exactament el que vol calcular l'usuari.

En aquesta secció anem a comparar les parametritzacions dels models presentades al curs de teoria IG12 impartit en 2006-2007, amb les de R, explicant les modificacions que caldria fer.

Binomial

- IG12-0607: una variable aleatòria X segueix la distribució binomial de paràmetres n i p, quan X representa el "nombre d'èxits" observats en realitzar n proves de Bernoulli (èxit-fracàs) independents, on p és la probabilitat de l'èxit de cada prova individual. Notació: $X \sim \text{Bin}(n, p)$.

- R: una variable aleatòria X segueix la distribució binomial de paràmetres `size` i `prob`, quan X representa el "nombre d'èxits" observats en realitzar `size` proves de Bernoulli (èxit-fracàs) independents, on `prob` és la probabilitat de l'èxit de cada prova individual.

- Per tant:

  $f(x) =$ `dbinom(x=`x`, size=`n`, prob=`p`)`

  $F(x) =$ `pbinom(q=`x`, size=`n`, prob=`p`)`

  $x_q =$ `qbinom(p=`q`, size=`n`, prob=`p`)`

  `simul.k.valors <- rbinom(n=`k`, size=`n`, prob=`p`)`

Binomial negativa

- IG12-0607: una variable aleatòria X segueix el model de distribució binomial negativa de paràmetres r i p quan X indica el "nombre de proves de Bernoulli necessàries" per obtenir `r` èxits (incloent-hi els èxits) on `p` és la probabilitat de cada èxit. Notació: $X \sim \text{BinNeg}(r, p)$.

- R: una variable aleatòria X segueix el model de distribució binomial negativa de paràmetres `size` i `prob` quan X indica el "nombre de fracassos (no èxits) ocorreguts abans d'obtenir `size` èxits" (per tant no compta els èxits) en realitzar proves de Bernoulli successives on `prob` és la probabilitat de cada èxit.

- Per tant:

  $f(x) =$ `dnbinom(x=`$x - r$`, size=`r`, prob=`p`)`

  $F(x) =$ `pnbinom(q=`$x - r$`, size=`r`, prob=`p`)`

  $x_q = r +$ `qnbinom(p=`q`, size=`r`, prob=`p`)`

  `simul.k.valors <-`$r +$ `rnbinom(n=`k`, size=`r`, prob=`p`)`

Hipergeomètrica

- IG12-0607: una variable aleatòria X segueix la distribució hipergeomètrica de paràmetres n, N i K quan X representa el "nombre d'èxits" observats en fer una extracció, sense reemplaçament, de n boles, d'una urna on hi ha un total de N boles equiprobables, de les quals K estan marcades com a "èxit". Notació: $X \sim \text{Hyper}(n, N, K)$.

- R: una variable aleatòria X segueix la distribució binomial de paràmetres `m`, `n` i `k`, quan X representa el "nombre d'èxits" observats en fer una extracció, sense reemplaçament, de `k` boles, d'una urna on hi ha `m` boles marcades com a "èxit" i `n` boles marcades com "no èxit", i totes són equiprobables.

- Per tant:

  $f(x) =$ `dhyper(x=`x`, m=`K`, n=`$N-K$`, k=`n`)`

  $F(x) =$ `phyper(q=`x`, m=`K`, n=`$N-K$`, k=`n`)`

  $x_p =$ `qhyper(p=`p`, m=`K`, n=`$N-K$`, k=`n`)`

  `simul.j.valors <- rhyper(nn=`j`, m=`K`, n=`$N-K$`, k=`n`)`

Poisson

- IG12-0607: una variable aleatòria X segueix la distribució de Poisson amb paràmetre λ quan X representa el "nombre d'ocasions" en les quals s'aprecia un esdeveniment en un interval d'espai o temps, quan es coneix que l'esdeveniment ocorre una mitjana de λ ocasions en cada interval d'espai o temps de la mateixa llargària. Notació: $X \sim \text{Po}(\lambda)$.

- R: una variable aleatòria X segueix la distribució de Poisson amb paràmetre `lambda` quan X representa el "nombre d'ocasions" en les quals s'aprecia un esdeveniment en un interval d'espai o temps, quan es coneix que l'esdeveniment ocorre una mitjana de `lambda` ocasions en cada interval d'espai o temps de la mateixa llargària.

- Per tant:

  $f(x) =$ `dpois(x=`x`, lambda=`λ`)`

  $F(x) =$ `ppois(q=`x`, lambda=`λ`)`

  $x_p =$ `qpois(p=`p`, lambda=`λ`)`

  `simul.k.valors <- rpois(n=`k`, lambda=`λ`)`

Uniforme contínua

- IG12-0607: una variable aleatòria X segueix la distribució uniforme de paràmetres a i b, quan X representa el "valor observat" en triar-se un valor de l'interval (a, b) completament a l'atzar, és a dir, on tots els valors tenen la mateixa versemblança de ser triats. Notació: $X \sim \text{U}(a, b)$.

- R: una variable aleatòria X segueix la distribució uniforme de paràmetres `min` i `max` quan X representa el "valor observat" en triar-se un valor de l'interval [`min`, `max`] completament a l'atzar, és a dir, on tots els valors tenen la mateixa versemblança de ser triats.

- Per tant:

  $f(x) =$ `dunif(x=`x`, min=`a`, max=`b`)`

$F(x) =$ `punif(q=`x`, min=`a`, max=`b`)`

$x_p =$ `qunif(p=`p`, min=`a`, max=`b`)`

`simul.k.valors <- runif(n=`k`, min=`a`, max=`b`)`

Exponencial

- IG12-0607: una variable aleatòria X segueix la distribució exponencial de paràmetre λ quan X representa la "llargària de l'interval" (d'espai o temps) que transcorre fins a l'aparició d'un esdeveniment assenyalat (o entre dues aparicions consecutives d'un esdeveniment assenyalat), quan es coneix que l'aparició de l'esdeveniment segueix la llei de Poisson de paràmetre λ. Notació: $X \sim \text{Exp}(\lambda)$.

- R: una variable aleatòria X segueix la distribució exponencial de paràmetre `rate` quan X representa la "llargària de l'interval" (d'espai o temps) que transcorre fins a l'aparició d'un esdeveniment assenyalat (o entre dues aparicions consecutives d'un esdeveniment assenyalat), quan es coneix que l'aparició de l'esdeveniment segueix la llei de Poisson de paràmetre `rate`.

- Per tant:

  $f(x) =$ `dexp(x=`x`, rate=`λ`)`

  $F(x) =$ `pexp(q=`x`, rate=`λ`)`

  $x_p =$ `qexp(p=`p`, rate=`λ`)`

  `simul.k.valors <- rexp(n=`k`, rate=`λ`)`

Normal

- IG12-0607: una variable aleatòria X que segueix la distribució normal pot vindre parametritzada per la mitjana i la desviació típica, μ i σ, o per la mitjana i la variància, μ i σ^2. Notació: $X \sim \text{N}(\mu, \sigma)$ o $X \sim \text{N}(\mu, \sigma^2)$.

- R: una variable aleatòria X que segueix la distribució normal ve parametritzada per la mitjana i la desviació típica, `mean` i `sd`.

- Per tant:

  $f(x) =$ `dnorm(x=`x`, mean=`μ`, sd=`σ`)`

  $F(x) =$ `pnorm(q=`x`, mean=`μ`, sd=`σ`)`

  $x_p =$ `qnorm(p=`p`, mean=`μ`, sd=`σ`)`

  `simul.k.valors <- rnorm(n=`k`, mean=`μ`, sd=`σ`)`

  Nota: En tots els casos anteriors, si la normal ve parametritzada per la variància σ^2, aleshores `sd = sqrt(`σ^2`)`.

Altres distribucions d'ús freqüent

Les quatre funcions (de densitat, de distribució, de quantils i de simulacions) de les distribucions que mostrem a continuació s'invoquen amb el prefix i arguments adequats:

- χ^2 amb n graus de llibertat:

  `%chisq(..., df =` n`...)`.

- t de Student amb n graus de llibertat:

  `%t(..., df=` n`...)`.

- F de Snedecor amb n_1 i n_2 graus de llibertat:

  `%f(..., df1=` n_1 `, df2=` n_2`...)`.

`%` pot ser qualsevol dels prefixos `d`, `p`, `q` o `r`, segons el càlcul que interesse realitzar.

Exercicis d'ensinistrament

Responeu a les següents qüestions usant les funcions convenients.

> **No oblideu que els models que es mencionen als exercicis corresponen a la parametrització exposada en el curs de teoria, i que per tant en usar R és necessari replantejar els paràmetres segons està explicat a les subseccions de la Secció 7.12**.

1. Un problema involucra una variable aleatòria X que representa el "nombre de bits invertits en transmetre una cadena de 125 bits" i que es cataloga com a $X \sim \text{Bin}(n = 125, p = 0.00021)$. Calculeu:

   (a) Probabilitat que no hi haja cap bit invertit. Sol.: 0.9740889

   (b) Probabilitat que hi haja entre 2 i 12 bits invertits.
   Sol.: 0.0003359458

   (c) Usant la llavor `set.seed(123)`, simula que s'han rebut 50 missatges i contesta en quants d'aquests no s'ha rebut cap bit invertit. Sol.: 49

2. La NASA estima que la probabilitat que falle un *component crític* dins del motor principal d'un transbordador espacial és d'aproximadament 1 sobre 5000. L'errada d'un component crític durant el vol conduiria directament a una catàstrofe del transbordador. El transbordador va fent missions a l'espai de manera que considerem X la variable aleatòria que registra el "nombre de missions realitzades fins que falle el component" (incloent-hi la missió on falle). Es considera que segueix la distribució binomial negativa amb paràmetres $r = 1$ (geomètrica) i $p = 1/5000$, és a dir $X \sim \text{BinNeg}(r = 1, p = 1/5000)$.

(a) Calculeu la probabilitat que volen almenys 15 missions sense que es produïsca cap errada. Sol: 0.002995804

(b) La direcció vol substituir el component crític després d'un nombre de missions, de manera que la probabilitat que es porte a terme aquest nombre de missions sense problemes siga d'almenys un 99%. Quin seria aquest nombre de missions preventiu? Sol: 51

3. El cost de provar **T**ubs de **R**aigs **C**atòdics (TRC) per terminals d'ordinadors és molt elevat. Imagina que vols comprar un lot de 20 TRC, i que et voldries assegurar que no hi ha aparells defectuosos. Tanmateix, no pots comprovar més que 3 per decidir si et quedes amb el lot sencer o no. Si trobes que algú dels 3 és defectuós rebutges la compra, i si no, l'acceptes.

(a) Si el lot conté un TRC defectuós, la variable aleatòria X que registra el "nombre de TRC defectuosos detectats" segueix la distribució Hyper$(n = 3, N = 20, K = 1)$. Aleshores, quina probabilitat hi ha que acceptes el lot? Sol: 0.85

(b) Si el lot conté 3 TRC defectuosos, la variable aleatòria X que registra el "nombre de TRC defectuosos detectats" segueix la distribució Hyper$(n = 3, N = 20, K = 3)$. Aleshores, quina probabilitat hi ha que acceptes el lot? Sol: 0.5964912

4. El nombre de transaccions de comerç electrònic per hora gestionades per un portal de viatjes segueix una llei de Poisson amb una mitjana de $\lambda = 5.85$ transaccions per hora. Calculeu:

(a) Probabilitat que durant la pròxima hora es gestionen més de 10 transaccions. Sol.: 0.03673122

(b) Probabilitat que no es reba cap petició de transacció durant els pròxims 20 minuts. Sol.: 0.857726

(c) El 99% de les hores, la web gestiona menys de _____ transaccions. Sol.: 12

(d) Usant la llavor `set.seed(123)`, simula que han transcorregut les **hores** d'una setmana sencera i contesta en quantes s'han superat les 10 transaccions. Sol.: 4

5. El gruix d'un tipus de peça usada per construir avions és un valor completament aleatori entre 0.95 i 1.05 mm, raó per la qual es pot considerar que la variable X que registra el "gruix de cada peça" segueix la distribució uniforme a l'interval $[0.95, 1.05]$.

(a) Quin percentatge de les peces excedeix els 1.02 mm de gruix? Sol: 30%

(b) El 90% de les peces té un gruix inferior a _____. Sol: 1.04 mm

6. En un sistema de comunicació de dades els missatges arriben a un node a una mitjana de 30 missatges per minut, raó per la qual es pot considerar que que la variable X, que registra el temps (en minuts!) transcurregut entre dos missatges consecutius, segueix la distribució exponencial de paràmetre $\lambda = 30$.

   (a) Quina és la probabilitat que el temps entre dos missatges siga inferior a 5 s? Sol: 0.002773923

   (b) Quina és la probabilitat que passen més de 10 s sense rebre's un missatge? Sol: 0.9944598

7. La vida d'un làser semiconductor funcionant a potència constant es registra a una variable X que distribueix segons la llei normal amb mitjana de 7000 hores i desviació típica de 600 hores.

   (a) Quina és la probabilitat que el làser falle abans de 5000 hores d'ús? Sol: 0.0004290603

   (b) El 95% dels làsers supera les _____ hores de duració. Sol: 6013.088

8. Calculeu els següents quantils:

   (a) Quantil 0.975 de la distribució t de Student amb 5 graus de llibertat. (Sol.: $(t_5)_{0.975} = 2.570582$)

   (b) Quantil 0.025 de la distribució chi-quadrat amb 9 graus de llibertat. (Sol.: $(\chi^2_9)_{0.025} = 2.700389$)

   (c) Quantil 0.99 de la distribució F de Snedecor amb 14 i 19 graus de llibertat.
   (Sol.: $(F_{14,19})_{0.99} = 3.194915$)

PART IV
INFERÈNCIA SOBRE POBLACIONS (INFERÈNCIA ESTADÍSTICA)

Capítol 8

Mostratge i estadístics de mostratge

8.1 Introducció

Quan una quantitat aleatòria, del nostre interès, pot modelitzar-se usant una distribució concreta de les presentades al capítol anterior, tenim informació sobre les probabilitats del que pot passar, i per tant podem considerar-les i prendre decisions més racionals.

Exemple 8.1.1 *Si el "nombre de telefonades rebudes per hora en la secció d'atenció al client d'una empresa" pot modelitzar-se com una variable aleatòria* $X \sim \text{Po}(\lambda = 3.2)$*, podem estimar a llarg termini el nombre d'hores on es rep cada quantitat de telefonades com es mostra a la Taula 8.1.*

Taula 8.1: Estudi de la intensitat de telefonades rebudes en un servei d'atenció al client. Nombre de telefonades rebudes per hora, i percentatge d'hores en què es rep aqueix nombre de telefonades

Nomb. telef.	0	1	2	3	4	5	6	$\cdots$
% d'hores	4.08	13.04	20.87	22.26	17.81	11.40	6.08	$\cdots$

L'empresa pot fer previsions del personal que necessita usant la informació i estimant el risc que els clients tarden molt de temps a ser atesos o queden sense atendre, etc.

L'elecció del tipus de model que ajusta una situació particular pot resultar més o menys controvertida, ja que cada model té un àmbit d'aplicació, i la interpretació de si la situació real segueix les condicions del model pot donar lloc a la subjectivitat. No obstant, una volta triat el model, queda per triar el valor concret del paràmetre. A l'Exemple 8.1.1, com es podria saber que $\lambda = 3.2$ i no un valor similar però diferent?

És impossible obtenir la resposta correcta. Per açò és important obtenir respostes raonables. La primera solució raonable involucra interpretar la variable aleatòria com una població infinita de dades. Seguint l'Exemple 8.1.1,

si observem els valors de la variable aleatòria en hores anteriors (és a dir, observem una mostra) i calculem la mitjana mostral, podem pensar que el més raonable és que la mitjana de la població (esperança de X) vinga ben aproximada per la de la mostra. Per tant $\overline{x} = \mathbb{E}(X) = \lambda$, i ja tenim una elecció raonable pel paràmetre λ. Quin risc té aquest procediment?

Encara que, malhauradament, l'Estadística no pot donar la solució exacta al problema, almenys pot informar sobre els riscos de prendre solucions com l'explicada.

8.2 Mostratge aleatori simple i estadístics

La forma en què s'obtenen les mostres influeix sobre la informació que es pot obtenir d'aquestes. Hi ha molta teoria sobre aquest punt, però ens restringirem a una forma, simple i de qualitat.

Definició 8.2.1 (Mostratge aleatori simple) *Siga X una variable aleatòria del nostre interès, i suposem que X segueix una distribució qualsevol amb mitjana $\mathbb{E}(X)$ i variància $\mathbb{V}(X)$.*

Es diu que la variable aleatòria conjunta $(X_1, X_2, \ldots, X_n)$ forma un mostratge aleatori simple (m.a.s.) de mida n de la distribució de X si:

- *Cada component X_i segueix la distribució de X.*
- *Cada component X_i és independent de totes les altres.*

Definició 8.2.2 (Estadístics de mostratge) *Si $X_1, X_2, \ldots, X_n$ és un m.a.s. de mida n d'una distribució, un estadístic de mostratge és qualsevol operació definida sobre $X_1, X_2, \ldots, X_n$.*

Exemple 8.2.1 (Mitjana i variància mostral) *Si $X_1, X_2, \ldots, X_n$ és un m.a.s. de mida n d'una distribució, la mitjana i variància mostrals (de mida n) són dos estadístics especials que es defineixen com:*

$$\overline{X} = \frac{1}{n}\sum_{i=1} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}(X_i - \overline{X})^2$$

8.3 Tres noves distribucions necessàries

Presentem tres nous models de variable aleatòria necessaris per a la inferència.

Definició 8.3.1 (Distribució chi-quadrat) *Si $Z_1, Z_2, \ldots, Z_k$ és un m.a.s. d'una distribució $\mathrm{N}(\mu = 0, \sigma^2 = 1)$ aleshores*

$$H = Z_1^2 + Z_2^2 + \ldots + Z_k^2$$

*segueix una distribució anomenada **chi-quadrat de k graus de llibertat** (notació $H \sim \chi^2_k$).*

Propietat 8.3.1 *Si $H \sim \chi^2_k$ aleshores $\mathbb{E}(H) = k$ i $\mathbb{V}(H) = 2k$.*

Definició 8.3.2 (Distribució t de Student) *Si $Z \sim \mathrm{N}(\mu = 0, \sigma^2 = 1)$ i $H \sim \chi^2_k$ són independents, aleshores*

$$T = \frac{Z}{\sqrt{H/k}}$$

segueix una distribució anomenada t ***de Student de k graus de llibertat*** *(notació $T \sim t_k$).*

Propietat 8.3.2 *Si $T \sim t_k$ i $k > 2$ aleshores $\mathbb{E}(T) = 0$ i $\mathbb{V}(T) = \frac{k}{k-2}$.*

Definició 8.3.3 (Distribució F de Snedecor) *Si $H_1 \sim \chi^2_{k_1}$ i $H_2 \sim \chi^2_{k_2}$ són independents, aleshores*

$$F = \frac{H_1/k_1}{H_2/k_2}$$

segueix una distribució anomenada F ***de Snedecor de k_1 i k_2 graus de llibertat*** *(notació $F \sim F_{k_1,k_2}$).*

Propietat 8.3.3 *Si $F \sim F_{k_1,k_2}$ aleshores $\mathbb{E}(F) = \frac{k_2}{k_2-2}$.*

8.4 Distribucions d'estadístics en el mostratge

Definició 8.4.1 (Distribució en el mostratge d'un estadístic) *Cada estadístic de mostratge és una variable aleatòria (ja que cada X_i del mostratge ho és). Per tant cada estadístic segueix una distribució, que s'anomena distribució en el mostratge.*

8.5 Usos de les noves distribucions

8.5.1 Per a la mitjana mostral

Propietat 8.5.1 (Esperança i variància de $\overline{X}$) *Si $\overline{X}$ és la mitjana mostral (de mida n) d'una variable aleatòria X de distribució amb esperança $\mathbb{E}(X)$ i variància $\mathbb{V}(X)$, aleshores:*

$$\mathbb{E}(\overline{X}) = \mathbb{E}(X) \qquad \mathbb{V}(\overline{X}) = \frac{\mathbb{V}(X)}{n}$$

Propietat 8.5.2 (Distribució de $\overline{X}$) *En general no se sap res sobre la distribució de $\overline{X}$ (de mida n), però sí en dos casos particulars:*

- *Si $X \sim \mathrm{N}(\mu = \mu_X, \sigma^2 = \sigma^2_X)$, aleshores $\overline{X} \sim \mathrm{N}(\mu = \mu_X, \sigma^2 = \frac{\sigma^2_X}{n})$*
- *Si n és "gran", aleshores $\overline{X} \sim_{\text{aprox.}} \mathrm{N}(\mu = \mathbb{E}(X), \sigma^2 = \frac{\mathbb{V}(X)}{n})$*

8.5.2 Per a la variància mostral

Propietat 8.5.3 (Esperança i variància de S^2) *Si S^2 és la variància mostral (de mida n) d'una variable aleatòria X de distribució amb esperança $\mathbb{E}(X)$ i variància $\mathbb{V}(X)$, aleshores:*

$$\mathbb{E}(S^2) = \mathbb{V}(X) \qquad \mathbb{V}(S^2) = \frac{2\mathbb{V}(X)^2}{n-1}$$

Propietat 8.5.4 (Distribució de S^2) *En general no se sap res sobre la distribució de S^2 (de mida n), però sí en un cas particular:*

- *Si $X \sim \mathrm{N}(\mu, \sigma^2)$, aleshores $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$*

8.5.3 Per a altres estadístics vinculats a la mitjana i variància mostrals

Propietat 8.5.5 *Si $X \sim \mathrm{N}(\mu, \sigma^2)$ i tenim la mitjana i variància mostrals (de mida n), $\overline{X}$ i S^2, aleshores l'estadístic següent es distribueix:*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Propietat 8.5.6 *Si $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ (per $i = 1, 2$) són independents i tenim les mitjanes i variàncies mostrals respectives (de mides n_1 i n_2 resp.), $\overline{X}_1, S_1^2$ i $\overline{X}_2, S_2^2$, aleshores l'estadístic següent es distribueix:*

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(S_1^2/n_1 + S_2^2/n_2)}} \sim t_{n_1+n_2-2}$$

Propietat 8.5.7 *Si $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ (per $i = 1, 2$) són independents i tenim les mitjanes i variàncies mostrals respectives (de mides n_1 i n_2 resp.), $\overline{X}_1, S_1^2$ i $\overline{X}_2, S_2^2$, aleshores l'estadístic següent es distribueix:*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{(n_1-1),(n_2-1)}$$

Exercici 8.5.1 *Si $X_1 \sim \mathrm{N}(\mu = 3.7, \sigma^2 = 1.6)$ i $X_2 \sim \mathrm{N}(\mu = 99.9, \sigma^2 = 1.2)$ són independents, calculeu la probabilitat que, en calcular una variància mostral de mida 10 en cada distribució, la primera siga major que la segona.*

Capítol 9

Estimació dels paràmetres dels models coneguts

9.1 Introducció

A la introducció del capítol anterior ja tractàvem de la importància de conéixer el model que ajustava una situació real d'interès. Quan el model està clar, però no el valor concret del paràmetre, l'estimació d'aquest és clau per tenir més informació sobre el procés aleatori en qüestió. Usant notació matemàtica, el problema que es planteja és el següent.

Siga $X \sim \mathcal{D}(\theta)$ on:

- $\mathcal{D}$ és una distribució coneguda (binomial, Poisson, exponencial, normal...)
- θ és el paràmetre de la distribució, i es desconeix.

Quant val θ?

- En principi, s'ha d'acceptar que és impossible conéixer-lo amb total seguretat. Per exemple, a la prova de Bernoulli de paràmetre p desconegut, encara que férem 100 proves, i totes resultaren èxit, no es podria asegurar que $p = 1$. Podria ser que $p = 0.999$, o $p = 0.999$, o fins i tot $p = 0.5$, ja que en eixos casos, les probabilitats de fer 100 proves que resulten totes èxit són, respectivament, 0.9047, 0.3660 i 7.88×10^{-31} (és a dir, possible en els tres casos, encara que més fàcil en uns que en altres).

Podem calcular el risc de les possibles estimacions de θ?

- L'estimació de θ es calcularà de manera raonada, però no serà el valor correcte. El fet important serà controlar que no estiga molt lluny del valor correcte. Donar informació sobre l'error (la distància al valor real) és útil per saber el risc que es pren quan s'accepta l'estimació per continuar treballant.

La proposta de solució consisteix en agafar una mostra (a partir d'un mostratge aleatori simple) i fer un càlcul amb aquesta (és a dir, usar un estadístic) amb la intenció de donar un valor prop del valor real però desconegut de θ.

9.2 Estimadors

Definició 9.2.1 (Estimador) *Un estimador del paràmetre θ de la distribució de la variable aleatòria X és un estadístic de mostratge denotat per $\widehat{\Theta}$ i amb l'objectiu d'estimar el paràmetre θ.*

Si $(X_1, X_2, \ldots, X_n)$ és un mostratge aleatori simple de X,

$$\widehat{\Theta} = \widehat{\Theta}(X_1, X_2, \ldots, X_n)$$

Una estimació és el valor de l'estimador calculat per a una mostra concreta. És a dir, si $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, aleshores, l'estimació és:

$$\widehat{\theta} = \widehat{\Theta}(x_1, x_2, \ldots, x_n)$$

L'estimador, com estadístic de mostratge que és, segueix un model de distribució, i té una esperança i una variància. A més, atés que el propòsit de l'estimador és obtenir valors que aproximen el valor real però desconegut de θ, i que l'estimació és un procés que es pot repetir indefinidament (si no materialment, almenys conceptualment), es poden analitzar unes propietats que són raonables per a fer *bones* estimacions.

Definició 9.2.2 (Propietats desitjables d'un estimador) *Si $\widehat{\Theta}$ és un estimador del paràmetre θ, es desitja que siga:*

- ***No esbiaixat:*** $\mathbb{E}(\widehat{\Theta}) = \theta$.
- ***Consistent:*** $\lim_{n\to\infty} \mathbb{E}(\widehat{\Theta}) = \theta$ *i* $\lim_{n\to\infty} \mathbb{V}(\widehat{\Theta}) = 0$ *(n és la mida del mostratge).*
- ***Eficient:*** $\mathbb{V}(\widehat{\Theta})$ *mínima.*

La definició d'estimadors és una tasca dels matemàtics. Els usuaris poden utilitzar els estimadors més populars, com són la mitjana i la variància mostral, i que, afortunadament, tenen les propietats desitjables.

Exemple 9.2.1 *Siga $X \sim \mathrm{N}(\mu, \sigma^2)$ amb μ i σ^2 desconegudes. Els estimadors $\widehat{\Theta}_1 = \overline{X}$ i $\widehat{\Theta}_2 = S^2$ són estimadors, respectivament, dels paràmetres μ i σ^2, i són:*

- *No esbiaixats: ja que $\mathbb{E}(\overline{X}) = \mu$ i $\mathbb{E}(S^2) = \sigma^2$.*
- *Consistents: ja que $\mathbb{V}(\overline{X}) = \frac{\sigma^2}{n}$ i $\mathbb{V}(S^2) = \frac{2\sigma^4}{n-1}$.*

9.3 Estimació puntual

L'estimació puntual és el procés que dóna lloc a valors concrets d'estimació, i s'obtenen calculant els valors dels estimadors per a les mostres concretes.

Encara que hi ha molts mètodes per fabricar estimadors puntuals, comentem només l'estimador pel mètode de la màxima versemblança.

9.3.1 Estimació puntual pel mètode de la màxima versemblança

Si X és una variable aleatòria distribuïda segons un model conegut amb paràmetre θ conegut, que podem denotar amb $X \sim \mathcal{D}(\theta)$, la funció f, anomenada funció de probabilitat (variable discreta) o funció de densitat de probabilitat (variable contínua), calcula, per a cada possible valor x de la variable X el valor de la seua probabilitat o densitat de probabilitat.

Si tant el valor de θ com el possible resultat de l'experiment són desconeguts, aleshores podem considerar que la funció f és una funció de dues variables:

$$f(x) = f(x, \theta).$$

Si s'agafa una mostra aleatòria de mida n, la probabilitat o densitat de probabilitat d'un possible resultat $x_1, x_2, \ldots, x_n$ és:

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

Per tant, si θ és desconegut, i considerem la mostra $x_1, x_2, \ldots, x_n$, tenim una funció amb $n+1$ variables:

$$f(x_1, x_2, \ldots, x_n, \theta) = f(x_1, \theta)f(x_2, \theta)\cdots f(x_n, \theta)$$

Atés que el significat de f és quantificar la credibilitat dels seus arguments, quan la mostra és coneguda, la funció:

$$f(x_1, x_2, \ldots, x_n, \theta) = f(x_1, \theta)f(x_2, \theta)\cdots f(x_n, \theta)$$

com a funció de θ s'anomena **versemblança**.

Definició 9.3.1 (Versemblança del paràmetre θ) *Si $X \sim \mathcal{D}(\theta)$, amb funció de probabilitat o de densitat de probabilitat f, i es té la mostra x_1, x_2, ..., x_n obtinguda per mostratge aleatori simple de X, s'anomena funció de versemblança de θ a la funció:*

$$V(\theta) = f(x_1, x_2, \ldots, x_n, \theta) = f(x_1, \theta)f(x_2, \theta)\cdots f(x_n, \theta)$$

El logarisme de la funció $V(\theta)$ s'anomena log-versemblança, i s'ha popularitzat per aprofitar la propietat que tenen els logarismes, respecte al producte, de transformar-lo en sumes:

$$\begin{aligned}\log V(\theta) &= \log\left[f(x_1, \theta)f(x_2, \theta)\cdots f(x_n, \theta)\right] \\ &= \log f(x_1, \theta) + \log f(x_2, \theta) + \cdots + \log f(x_n, \theta)\end{aligned}$$

Definició 9.3.2 (Estimador de màxima versemblança) *S'anomena **Estimador de Màxima Versemblança** (EMV) l'estimador que maximitza la funció $V(\theta)$. Es denota per $\widehat{\Theta}_{\text{EMV}}$, i és:*

$$\widehat{\Theta}_{\text{EMV}} = \text{argmax}_\theta V(\theta)$$

Nota 9.3.1 *El valor que maximitza la funció $V(\theta)$ és el mateix que maximitza la funció $\log V(\theta)$, ja que la funció $\log$ és estrictament creixent.*

Exercici 9.3.1 *Demostrem la fórmula de l'estimador EMV del paràmetre λ per a la distribució de Poisson. Denotem la mostra per x_1, x_2, ..., x_n. Aleshores:*

1. *La funció de probabilitat per cada x_i és:*

$$f(x_i, \lambda) =$$

2. *La funció de versemblança és, per tant:*

$$\begin{aligned} V(\lambda) &= f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) \\ &= \\ &= \\ &= C(x_1, x_2, \ldots, x_n) \cdot \end{aligned}$$

   on $C(x_1, x_2, \ldots, x_n)$ és un valor que involucra la mostra però no el valor de λ, i que per tant és constant per a la funció V.

3. *La funció log-versemblança, que és més fàcil de treballar, és:*

$$\log V(\lambda) = \log C(x_1, x_2, \ldots, x_n) +$$

4. *El màxim de la funció $\log V(\lambda)$ es troba derivant (respecte de la variable λ) i igualant a zero:*

$$= 0$$

5. *El resultat és*

$$\lambda = \overline{x}$$

Es proposa com exercici trobar l'estimador de màxima versemblança per a una altra distribució de les estudiades al curs.

Exercici 9.3.2 *Els temps de vida útil d'una sèrie de monitors LCD ha sigut (en mesos):*

$$32.5, 56.1, 9.3, 19.6, 24.5, 13.1, 16.6$$

Trieu la distribució teòrica raonable que podria seguir la variable "temps de vida útil dels aparells" i estimeu el(s) valor(s) del(s) seu(s) paràmetre(s) pel mètode de la màxima versemblança.

9.4 Estimació per interval

9.4.1 Introducció

Un estimador puntual $\widehat{\Theta}$ dóna, per cada mostra obtinguda, un possible valor $\widehat{\theta}$ del paràmetre θ però:

Com d'allunyat pot estar $\widehat{\theta}$ de θ en realitat?

Per respondre esta qüestió és molt necessari conéixer la distribució de l'estimador $\widehat{\Theta}$ (posició, dispersió...). Si la dispersió és alta, les estimacions que fa l'estimador són valors molt dispars, per tant, una estimació feta serà poc de fiar. Si la dispersió és baixa, les estimacions seran sempre similars, el qual és molt desitjable. Si la posició central és correcta, les estimacions variaran al voltant del valor desconegut de manera equitativa, però si la posició central no és correcta, les estimacions infravaloraran (o sobrevaloraran) generalment el valor que tracten d'estimar.

Farem servir un nou concepte: la **confiança** (o el concepte contrari, el **risc**).

Exemple 9.4.1 (Interval de confiança per a una variable aleatòria) *Imaginem que tenim una variable aleatòria X amb una distribució coneguda i un experiment real que podria seguir la distribució de X.*

L'experiment resulta en un valor x i ens demanem: és creïble que l'experiment segueix la distribució de X?

Com podem contestar? Un procediment (lògic, però no l'únic) per a decidir, a partir del resultat x del nou experiment, si X és la variable aleatòria del nou experiment seria:

- *Si x forma part dels valors **menys versemblants** de la distribució de X, **decidirem que** l'experiment no segueix la distribució de la variable aleatòria X, i tindrem un risc d'equivocar-nos.*
- *Si x forma part dels valors **més versemblants** de la distribució de X, **decidirem que** l'experiment **sí que segueix** la distribució de la variable aleatòria X, i tindrem una confiança d'encertar.*

Si la distribució X és la de la Figura 9.1, i ens diuen que la dada $x = 100$ ha eixit de la distribució de X, acceptem l'afirmació?

Exemple 9.4.2 (Interval de confiança, continuació) *Com separem els valors més versemblants dels menys versemblants? Prendrem un **nivell de risc** α (petit) de manera que:*

- $\Pr(X \in \{\textit{Valors menys versemblants}\}) = \alpha$
- $\Pr(X \in \{\textit{Valors més versemblants}\}) = 1 - \alpha$.

(veure Figura 9.1). Amb aquest criteri (si l'experiment segueix en realitat la distribució X, però no ho sabem) decidirem que:

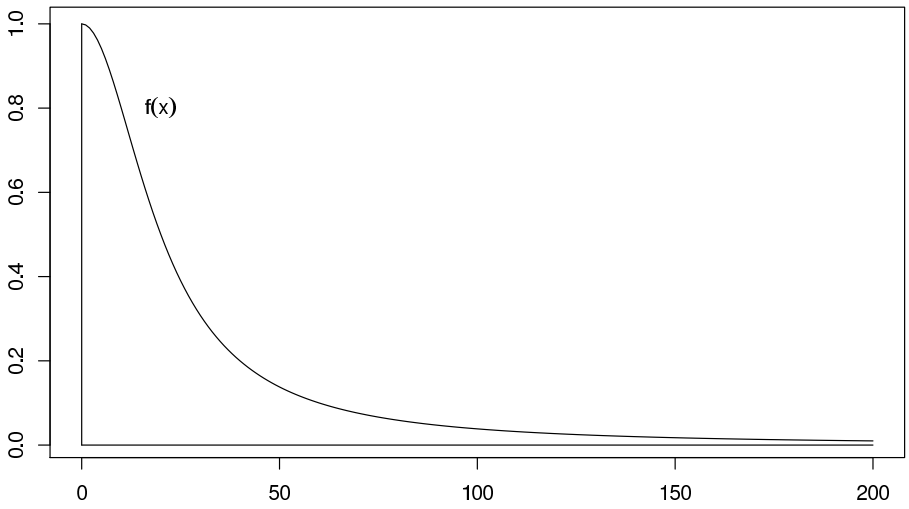


Figura 9.1: Funció de densitat suposada per a X. Acceptem l'afirmació si $x = 100$?

- *No segueix la distribució de X en un $\alpha \times 100\%$ de les ocasions.*
- *Sí segueix la distribució de X en un $(1-\alpha) \times 100\%$ de les ocasions.*

El valor $1-\alpha$ es coneix com ***nivell de confiança****.*

Els valors més habituals per a la confiança són 90%, 95% i 99%, encara que es podrien triar altres.

Si la distribució de X té els valors menys versemblants als costats (vegeu la Figura 9.2), els valors més versemblants vénen donats per un interval, anomenat **interval de confiança**, delimitat per quantils (Atenció! En la literatura també els delimiten per "punts crítics", que són com els quantils, però amb l'àrea cap a la dreta). Si el nivell de confiança és $1-\alpha$ (risc α), la forma de l'interval és $[x_{\frac{\alpha}{2}}, x_{1-\frac{\alpha}{2}}]$, és a dir $X \in [x_{\frac{\alpha}{2}}, x_{1-\frac{\alpha}{2}}]$ amb confiança $1-\alpha$, encara que en aquesta frase, la paraula *confiança* és sinònim de *probabilitat*.

9.4.2 Aplicació a les principals distribucions de mostratge

L'interval de confiança de nivell (de confiança) $1-\alpha$ és:

- Si $Z \sim \mathrm{N}(0,1)$, l'interval de probabilitat $1-\alpha$ format pels valors més versemblants és, per simetria, $[-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]$. Per tant es pot dir que:

$$Z \in [-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}] \text{ amb confiança } 1-\alpha.$$

  Valors usuals són: $z_{0.95} = 1.64$, $z_{0.975} = 1.96$, $z_{0.995} = 2.58$

- Si $H \sim \chi^2_n$, l'interval de probabilitat $1-\alpha$ format pels valors més versemblants és, $[(\chi^2_n)_{\frac{\alpha}{2}}, (\chi^2_n)_{1-\frac{\alpha}{2}}]$. Per tant es pot dir que:

$$H \in [(\chi^2_n)_{\frac{\alpha}{2}}, (\chi^2_n)_{1-\frac{\alpha}{2}}] \text{ amb confiança } 1-\alpha.$$

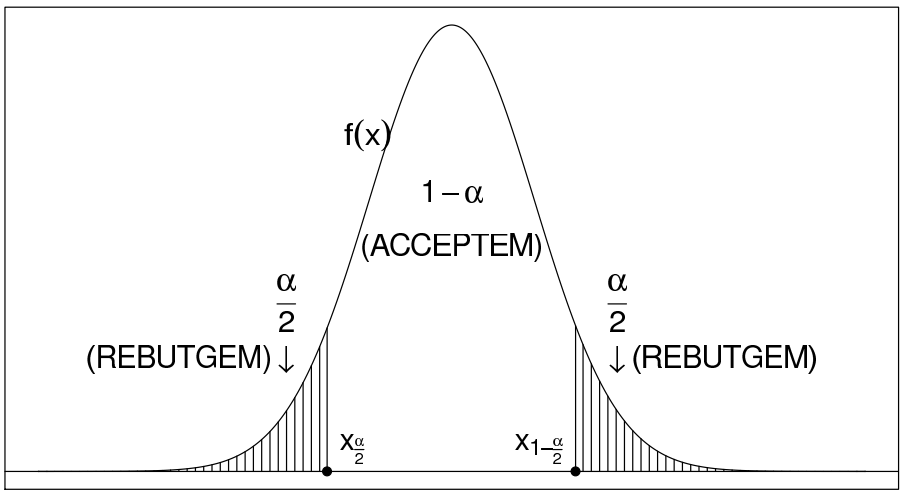


Figura 9.2: Quantils que determinen la regió de valors més versemblants en una distribució simètrica i campaniforme

- Si $T \sim t_n$, l'interval de probabilitat $1-\alpha$ format pels valors més versemblants és, per simetria, $[-(t_n)_{1-\frac{\alpha}{2}}, (t_n)_{1-\frac{\alpha}{2}}]$. Per tant es pot dir que:

$$T \in [-(t_n)_{1-\frac{\alpha}{2}}, (t_n)_{1-\frac{\alpha}{2}}] \text{ amb confiança } 1-\alpha.$$

- Si $F \sim F_{n_1,n_2}$, l'interval de probabilitat $1-\alpha$ format pels valors més versemblants és, $[(F_{n_1,n_2})_{\frac{\alpha}{2}}, (F_{n_1,n_2})_{1-\frac{\alpha}{2}}]$. Per tant es pot dir que:

$$F \in [(F_{n_1,n_2})_{\frac{\alpha}{2}}, (F_{n_1,n_2})_{1-\frac{\alpha}{2}}] \text{ amb confiança } 1-\alpha.$$

En el fons, els intervals relacionats més amunt són intervals de probabilitat, ja que les variables Z, H, T i F disposen d'una probabilitat, i la probabilitat de que el valor d'eixes variables estiga dins de cada interval respectiu és exactament $1-\alpha$.

9.4.3 Aplicació a l'estimació de paràmetres

En aquest apartat passem dels intervals de probabilitat (sobre variables aleatòries que tenen una distribució de probabilitat) als intervals de confiança (sobre valors desconeguts de paràmetres que no tenen una distribució de probabilitat).

Com a notació, es denotarà simplificadament com $[c \pm r]$ a l'interval de la forma $[c-r, c+r]$.

Interval de confiança per a la mitjana μ en poblacions normals (o en poblacions qualsevol si n és "gran") assumint variància σ^2 coneguda

1. $Z := \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$
2. Per tant, $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \in [-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]$ amb probabilitat $1-\alpha$.

3. Aïllant μ, es troba que $\mu \in [\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$ amb confiança $1-\alpha$.

Interval de confiança per a la mitjana μ en poblacions normals amb variància σ^2 desconeguda

1. $T := \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$
2. Per tant, $\dfrac{\overline{X}-\mu}{S/\sqrt{n}} \in [-(t_{n-1})_{(1-\frac{\alpha}{2})}, (t_{n-1})_{(1-\frac{\alpha}{2})}]$ amb probabilitat $1-\alpha$.
3. Aïllant μ, es troba que $\mu \in [\overline{X} \pm (t_{n-1})_{(1-\frac{\alpha}{2})} \dfrac{S}{\sqrt{n}}]$ amb confiança $1-\alpha$.

Nota: Si n és "gran", $t_{n-1} \approx N(\mu=0, \sigma^2=1)$ i es poden usar els quantils de la normal tipificada (z).

Interval de confiança per a la variància σ^2 en poblacions normals

1. $H := \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
2. Per tant, $\dfrac{(n-1)S^2}{\sigma^2} \in [(\chi^2_{n-1})_{\frac{\alpha}{2}}, (\chi^2_{n-1})_{1-\frac{\alpha}{2}}]$ amb probabilitat $1-\alpha$.
3. Aïllant σ^2, es troba que $\sigma^2 \in [\dfrac{(n-1)S^2}{(\chi^2_{n-1})_{1-\frac{\alpha}{2}}}, \dfrac{(n-1)S^2}{(\chi^2_{n-1})_{\frac{\alpha}{2}}}]$ amb confiança $1-\alpha$.

Interval de confiança per a la proporció p en prova de Bernoulli (binomial) amb n "gran"

1. Prenem la "proporció mostral", $\widehat{P} := \frac{\text{nº èxits}}{n} = \overline{X} \sim_{\text{aprox.}} \text{N}(\mu=p, \sigma^2 = \frac{p(1-p)}{n})$
2. Definim $Z := \dfrac{\widehat{P}-p}{\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}} \sim_{\text{aprox.}} \text{N}(0,1)$
3. Per tant, $\dfrac{\widehat{P}-p}{\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}} \in [-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]$ amb probabilitat $1-\alpha$.
4. Aïllant p, es troba que $p \in [\widehat{P} \pm z_{1-\frac{\alpha}{2}} \sqrt{\dfrac{\widehat{P}(1-\widehat{P})}{n}}]$ amb confiança $1-\alpha$.

S'observa que l'interval té la major amplària possible quan la proporció estimada és de 0.5.

Definició 9.4.1 (Error d'estimació) *S'anomena error d'estimació associat a un interval de confiança de nivell de confiança $1-\alpha$ a la meitat de la seua amplària.*

És a dir, si l'interval és $[a, b]$ l'error d'estimació és

$$\text{Error} = \frac{b-a}{2}$$

Observació 9.4.1 (Reduir l'error de l'interval) *Per a l'estimació sempre interessarà un interval amb una confiança suficient (90, 95, 99%) però precís (el més reduït possible).*

La mida de la mostra (n) serveix per a disminuir l'error. De vegades la mida de la mostra ve determinada per l'error màxim que es puga assumir.

En ocasions (per exemple per raons de cost econòmic del mostratge) interessa conéixer a priori el mínim tamany de mostra necessari per tal de poder fer una estimació del paràmetre amb un error "controlat" per una fita superior, i amb un nivell de confiuança prefixat.

Observació 9.4.2 *Encara que no es pot calcular l'interval de confiança abans de recollir les dades, en l'estimació de la proporció p de proves de Bernoulli (binomial), l'error d'estimació és el valor*

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}}.$$

Si es vol afitar l'error per un valor fixat E_0, es pot anar al pitjor escenari, on $\widehat{P} = 0.5$, quedant:

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{P}(1-\widehat{P})}{n}} \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{0.25}{n}} \leq E_0.$$

Ara, aïllant la n obtenim que si

$$n \geq \left(\frac{z_{1-\frac{\alpha}{2}}}{2E_0}\right)^2,$$

l'error d'estimació en els pitjors dels casos serà inferior a E_0, i per tant també ho serà siga quina siga la proporció mostral observada a la mostra.

9.5 Exercicis proposats

Exercici 9.5.1 *Una màquina fabrica peces de llargària distribuïda normalment amb mitjana i variància desconegudes. Si agafem una mostra de 50 peces i calculem $\overline{x} = 14.3$ i $s^2 = 1.82$, calculeu un interval de confiança del 95% per a la mitjana i un altre per a la variància de les llargàries de les peces.*

Exercici 9.5.2 *Una màquina fabrica peces de llargària distribuïda normalment amb mitjana desconeguda i variància $\sigma^2 = 0.1$ cm^2.*

Quina mida de mostra caldria triar per a estimar μ amb un error màxim d'una centèsima de centímetre (0.01 cm) amb una confiança del 99%?

Exercici 9.5.3 *En un sondeig electoral es consulta l'opinió de* 567 *persones, de les quals,* 259 *votarien al partit* XXX*. Calculeu l'interval de confiança al* 95% *per a la proporció* p *de votants del partit* XXX *entre l'electorat.*

Exercici 9.5.4 *Calculeu la mida mínima de la consulta de l'exercici anterior per tal de poder estimar el valor de la proporció real de votants* p *amb un error màxim del* 0.01 *(usant una confiança del* 95%*).*

Exercici 9.5.5 *Una mostra aleatòria de 36 cigarretes d'una determinada marca donà un contingut mitjà de nicotina de 3 mil·ligrams. El contingut en nicotina d'aquestes cigarretes segueix la llei normal amb una desviació estàndard d'1 mil·ligram. Obteniu i interpreteu un interval de confiança del 95% per al vertader contingut mitjà de nicotina en aquestes cigarretes.*

Exercici 9.5.6 *Els següents nombres representen el temps (en minuts) que van tardar 15 operaris a familiaritzar-se amb el funcionament d'una nova màquina adquirida per l'empresa: 3.4, 2.8, 4.4, 2.5, 3.3, 4, 4.8, 2.9, 5.6, 5.2, 3.7, 3, 3.6, 2.8, 4.8. Suposem que els temps es distribueixen normalment. Determineu i interpreteu un interval del 95% de confiança per al vertader temps mitjà.*

Exercici 9.5.7 *Una marca de rentadores vol saber la proporció de llars en què prefereixen usar la seua marca. Prenen a l'atzar una mostra de 100 llars i en 20 diuen que la usarien. Calculeu un interval de confiança del 95% per a la verdadera proporció de llars que preferirien l'esmentada marca de rentadores.*

Exercici 9.5.8 *Volem ajustar una màquina de refrescos de manera que la mitjana del líquid dispensat quede dins de cert rang. La quantitat de líquid abocat per la màquina segueix una distribució normal amb desviació estàndard de 0.15 decilitres. Desitgem que el valor estimat que es vaja a obtenir comparat amb el vertader no siga superior a 0.2 decilitres amb una confiança del 95%. De quina grandària hem de triar la mostra?*

Exercici 9.5.9 *Una màquina ompli caixes amb un cert cereal. El supervisor desitja conéixer amb un error d'estimació de màxim 0.1 i un nivell de confiança del 90%, una mitjana estimada del pes. Com que la variància era desconeguda es va procedir a triar una mostra pilot. Els resultats van ser els següents: 11.02, 11.14, 10.78, 11.59, 11.58, 11.19, 11.71, 11.27, 10.93, 10.94. Quantes caixes ha de triar perquè es complisquen els requisits proposats?*

Exercici 9.5.10 *Es desitja fer una enquesta per a determinar la proporció de famílies que no tenen mitjans econòmics per a atendre els problemes de salut. Hi ha la impressió que aquesta proporció està pròxima a 0.35. Es desitja determinar un interval de confiança del 95% amb un error d'estimació de 0.05. De quina grandària ha de prendre's la mostra?*

Exercici 9.5.11 *Un productor de llavors desitja saber amb un error d'estimació de l'1% el percentatge de llavors que germinen en la granja del seu competidor. Quina grandària de mostra ha de prendre's per a obtenir un nivell de confiança del 95%?*

Exercici 9.5.12 *Es desitja realitzar una enquesta entre la població juvenil d'una determinada localitat per a determinar la proporció de joves que estaria a favor d'una nova zona d'oci. El nombre de joves de la població és de 2000. Determineu la grandària de mostra necessària per a estimar la proporció d'estudiants que estan a favor amb un error d'estimació de 0.05 i un nivell de confiança del 95%.*

Capítol 10

Proves d'hipòtesi sobre paràmetres de models coneguts

10.1 Definicions

Situació 10.1.1 *Suposem que un experiment aleatori pot modelitzar-se segons una variable aleatòria X amb una certa distribució. Volem contrastar:*

- *una caracterìstica de la distribució que es suposa certa, contra*
- *una desviació de la característica anterior, que es sospita que puga estar passant.*

Per tant, interessa posar a prova la característica inicial (i acceptar que és certa o decidir que és falsa, i que és certa la desviació sospitada).

Exemple 10.1.1 *Una maquina fabricada per a omplir ampolles d'aigua mineral (1500 ml) té l'especificació: $X \sim \mathrm{N}(\mu = 1500, \sigma^2 = 15)$ on X ="volum omplit per ampolla". Sospitem que la màquina podria funcionar malament, omplint per davall d'allò especificat. Aleshores, el que es pot contrastar és el parell:*

H_0 $\mu = 1500$ *(es suposa cert), contra*

H_1 $\mu < 1500$ *(es sospita que ompli per defecte)*

Definició 10.1.1 Prova o contrast d'hipòtesi() *En una situació com l'anterior, s'anomena:*

- *Hipòtesi nul·la (H_0): propietat que se suposa a la variable aleatòria X fins que no hi haja evidència del contrari, i que es vol posar a prova.*
- *Hipòtesi alternativa (H_1): propietat distinta a H_0 i se sospita que està passant o es vol contraposar a la hipòtesi nul·la.*

Les hipòtesis nul·les poden ser de molts tipus. En aquest curs ens centrem en proves on la hipòtesi nul·la expressa el valor suposat d'un paràmetre d'un

model conegut. S'anomena hipòtesi simple, front a hipòtesis on es suposa que el paràmetre pertany a un conjunt de valors (hipòtesi composta).

La hipòtesi alternativa doncs, pot expressar el complementari d'H_0 ($\neq$) o una desigualtat només ($>$, $<$).

Nota 10.1.1 (Procediment estadístic estàndard) *Una prova d'hipòtesi és un problema per al qual no hi ha cap tècnica que porte a la decisió correcta amb un 100% de seguretat. L'estadística, per tant, no pot donar la resposta perfecta, però sí que és capaç de donar procediments de decisió que inclouen el nivell de risc d'equivocar-se. L'usuari final és qui aprofita aquesta informació per a triar la seua decisió personal.*

D'altra banda, la confecció d'una prova d'hipòtesi no és única. Es poden dissenyar dos procediments distints per a decidir una mateixa prova d'hipòtesi, ambdós amb la mateixa significació, i en aplicar-los sobre la mateixa mostra, poden donar lloc a dues decisions diferents.

Per tant, l'Estadística ataca les proves d'hipòtesi de la manera següent:

1. *Crea un estadístic (anomenat **de contrast**) amb la condició que, si es suposa H_0 certa, aleshores es coneix la distribució d'aqueix estadístic i els valors més versemblants són els que més s'aproximen a la H_0 certa.*

2. *Dins de la suposició que H_0 siga certa, es calcula la zona de valors de manera que la probabilitat que l'estadístic se n'isca de la zona, en la direcció indicada per la hipòtesi alternativa, siga exactament igual a α.*

3. *Si una mostra dóna lloc a un valor d'estadístic de contrast fora de la zona indicada a l'apartat anterior, aleshores la decisió és rebutjar H_0.*

Nota 10.1.2 (Casuística en les proves d'hipòtesi) *Una vegada està plantejada la prova, es poden donar les situacions indicades a la Taula 10.1. Valorar*

Taula 10.1: Quatre situacions que es poden donar en resoldre una prova d'hipòtesi amb una decisió (acceptar o rebutjar H_0)

		REALITAT	
		H_0 és certa	H_0 és falsa
DECISIÓ	Acceptes H_0	ENCERT	ERROR TIPUS II
	Rebutges H_0	ERROR TIPUS I	ENCERT

la gravetat de l'error depén de les situacions particulars. Generalment és molt més greu l'error tipus I, raó per la qual és important controlar-la amb un valor petit.

Definició 10.1.2 (Significació i potència d'un contrast) *S'anomena* ***significació del contrast*** *la probabilitat de l'error tipus I. Es denota per* α *i ha de ser un valor petit (es sol prendre* 0.05*,* 0.01 *o* 0.1*).*

S'anomena ***potència del contrast*** *la probabilitat de rebutjar* H_0 *quan és falsa (el complementari de l'error tipus II). Si es denota per* β *el valor de la probabilitat de cometre l'error tipus II, aleshores la potència és* $1-\beta$*, i interessa que siga com més gran millor.*

Malhauradament, no sempre es pot fer pujar la potència sense fer pujar col·lateralment la significació. Normalment serà més important tenir α controlat i petit, per la gravetat que implica l'error tipus I.

Encara que el procediment mencionat a la Nota 10.1.1 és la forma estàndard d'abordar les proves d'hipòtesi, una xicoteta variant d'aquest procediment és lleugerament més informativa, i és utilitzada per la majoria de programes informàtics que tracten el tema.

Nota 10.1.3 (Procediment estadístic alternatiu) *El procediment estàndard de la Nota 10.1.1 consisteix en crear una regió de valors d'acceptació del contrast, usant el nivell de significació* α*. Una volta creada la regió, per cada mostra que trobem, calcularem l'estadístic i, segons estiga o no en aqueixa regió, acceptarem o no* H_0*. Alternativament:*

1. *Es calcula el valor concret de l'estadístic de contrast (del qual es coneix la distribució si es suposa* H_0 *certa) per a la mostra concreta que es tinga.*
2. *Dins de la suposició que* H_0 *siga certa, es calcula la probabilitat que l'estadístic de contrast done un valor tan versemblant o menys (és a dir, tan estrany o més) que el valor que ha donat amb la mostra concreta. Aquesta probabilitat d'estranyesa del resultat s'anomena p-valor (o p-value, en anglés).*
3. *Ara, comparant el p-valor amb* α*, i decidint:*

   - *Acceptar* H_0 *si p-valor* $\geq \alpha$
   - *Rebutjar* H_0 *si p-valor* $< \alpha$

   tindrem un contrast d'hipòtesi de nivell de significació exactament igual a α*.*

Així, els programes informàtics d'estadística calculen el p-valor de cada mostra i el responsable del contrast decideix si accepta o no H_0 *segons el nivell de risc que pot assumir.*

10.2 Alguns contrastos paramètrics habituals

Contrast sobre la mitjana μ d'una població normal (o una població qualsevol si la mida mostral és prou gran) amb variància σ^2 coneguda

- Contrast $\left\{ \begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{array} \right\}$ o $\left\{ \begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{array} \right\}$ o $\left\{ \begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{array} \right\}$, respectivament, amb significació α.

- Si H_0 és certa: $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ i
  - $\Pr(Z \notin [-z_{1-\alpha/2}, z_{1-\alpha/2}]) = \alpha$
  - $\Pr(Z \notin [-z_{1-\alpha}, +\infty)) = \alpha$
  - $\Pr(Z \notin (-\infty, z_{1-\alpha}]) = \alpha$

  respectivament.

- Per tant, si acceptem H_0 en el cas:
  - $\overline{X} \in \left[\mu_0 - z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right]$
  - $\overline{X} \in \left[\mu_0 - z_{1-\alpha}\dfrac{\sigma}{\sqrt{n}}, +\infty\right)$
  - $\overline{X} \in \left(-\infty, \mu_0 + z_{1-\alpha}\dfrac{\sigma}{\sqrt{n}}\right]$

  i rebutgem H_0 en el cas contrari tindrem, respectivament, procediments de contrast de significació exactament igual a α.

Contrast sobre la mitjana μ d'una població normal (o una població qualsevol si la mida mostral és prou gran) amb variància σ^2 desconeguda

- Contrast $\left\{\begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{array}\right\}$ o $\left\{\begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{array}\right\}$ o $\left\{\begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{array}\right\}$, respectivament, amb significació α.

- Si la variància és desconeguda, l'estadístic amb el qual es treballa és $T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$.

- Per tant, els tres contrastos anteriors queden d'igual manera reemplaçant cada quantil de la z pel mateix quantil de la t_{n-1}.

Contrast sobre la igualtat de mitjanes μ_1 i μ_2 de dues poblacions normals dependents o emparellades amb variàncies desconegudes

En ocasions, quan es comparen dues poblacions normals, el disseny del mostratge fa que les dades obtingudes estiguen lligades. Per exemple, per a comparar la velocitat de funcionament de dos algorismes d'ordenació de bases de dades, A i B, podem usar una única bateria de bases de dades, i ordenar cada base de dades amb els dos algorismes. Per a cada base de dades tindrem dos valors, un temps d'ordenació segons A i un altre temps segons B. Una base de dades complexa farà que els dos algorismes tarden prou de temps en ordenar-la, i una base de dades senzilla s'ordenarà en un temps reduït, per als dos algorismes. En aquest cas les poblacions es consideraran dependents o emparellades (pel mètode de mostratge utilitzat).

- Contrast $\left\{ \begin{array}{ll} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{array} \right\}$ amb significació α.
- Si H_0 és certa, X_1 i X_2 representen les variables aleatòries de les poblacions respectives, i es forma la variable $D = X_1 - X_2$, aleshores, la mostra feta a partir de les diferències de les dades emparellades de les dues mostres originals, és una mostra de D. La mitjana de les diferències (denotada per $\overline{D}$) estima la veritable mitjana de D (que coincideix amb $\mu_1 - \mu_2$, i la desviació típica de les diferències (denotada per S_D) estima la veritabledesviació típica de D. Per tant:

$$\frac{\overline{D}}{S_D/\sqrt{n}} \sim t_{n-1},$$

  on n és la mida comuna de les mostres.
- Així doncs, si acceptem H_0 en el cas:

$$\overline{D} \in [-(t_{n-1})_{1-\alpha/2} S_D/\sqrt{n}, (t_{n-1})_{1-\alpha/2} S_D/\sqrt{n}]$$

  i rebutgem H_0 en el cas contrari, tidrem un procediment de contrast de significació exactament igual a α.

Contrast sobre la igualtat de mitjanes μ_1 i μ_2 de dues poblacions normals independents amb variàncies desconegudes però que es poden considerar iguals

Dins el context de l'exemple comentat al contrast anterior, agafar aleatòriament dues bateries de bases de dades, i passar cada bateria per un únic algorisme, fa que els valors de temps obtinguts amb cada algorisme no tinguen cap vincle, i es consideren independents.

- Contrast $\left\{ \begin{array}{ll} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{array} \right\}$ amb significació α.
- Si H_0 és certa, podem estimar millor la variància comuna a les poblacions amb la variància "apilada" $S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$. Aleshores:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}.$$

- Per tant, si acceptem H_0 en el cas:

$$\overline{X}_1 - \overline{X}_2 \in \left[-(t_{n_1+n_2-2})_{1-\alpha/2} S\sqrt{1/n_1 + 1/n_2}, (t_{n_1+n_2-2})_{1-\alpha/2} S\sqrt{1/n_1 + 1/n_2}\right]$$

  i rebutgem H_0 en el cas contrari, tidrem un procediment de contrast de significació exactament igual a α.

Contrast sobre la igualtat de mitjanes μ_1 i μ_2 de dues poblacions normals independents amb variàncies σ_1^2 i σ_2^2 desconegudes i que no es poden considerar iguals

- Contrast $\left\{ \begin{array}{ll} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{array} \right\}$ amb significació α.
- Si H_0 és certa:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t_k, \text{ on } k = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

- Per tant, si acceptem H_0 en el cas:

$$\overline{X}_1 - \overline{X}_2 \in \left[-(t_k)_{1-\alpha/2}\sqrt{S_1^2/n_1 + S_2^2/n_2}, (t_k)_{1-\alpha/2}\sqrt{S_1^2/n_1 + S_2^2/n_2} \right]$$

  i rebutgem H_0 en el cas contrari, tidrem un procediment de contrast de significació exactament igual a α.

Contrast sobre la variància σ^2 d'una població normal

- Contrast $\left\{ \begin{array}{ll} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 \neq \sigma_0^2 \end{array} \right\}$ amb significació α.
- Si H_0 és certa:

$$H = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} \qquad \text{i} \qquad \Pr(H \notin [(\chi^2_{n-1})_{\alpha/2}, (\chi^2_{n-1})_{1-\alpha/2}]) = \alpha.$$

- Per tant, si acceptem H_0 en el cas:

$$S^2 \in \left[\frac{\sigma_0^2(\chi^2_{n-1})_{\alpha/2}}{n-1}, \frac{\sigma_0^2(\chi^2_{n-1})_{1-\alpha/2}}{n-1} \right]$$

  i rebutgem H_0 en el cas contrari, tidrem un procediment de contrast de significació exactament igual a α.

- Els contrastos unilaterals es resolen de manera similar als anteriors, usant els quantils convenients.

Contrast sobre la igualtat de variàncies σ_1^2 i σ_2^2 de dues poblacions normals independents

- Contrast $\left\{ \begin{array}{ll} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{array} \right\}$ amb significació α.
- Si H_0 és certa:

$$F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}.$$

- Per tant, si acceptem H_0 en el cas:

$$\frac{S_1^2}{S_2^2} \in [(F_{n_1-1,n_2-1})_{\alpha/2}, (F_{n_1-1,n_2-1})_{1-\alpha/2}]$$

i rebutgem H_0 en el cas contrari, tidrem un procediment de contrast de significació α.

Contrast sobre la proporció p d'una població binomial o de Bernoulli amb una mostra de gran mida i p no prop dels extrems 0 ni 1

- Contrast $\left\{ \begin{array}{ll} H_0: & p = p_0 \\ H_1: & p \neq p_0 \end{array} \right\}$ amb significació α.

- Si H_0 és certa:

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim_{\text{aprox.}} \mathrm{N}(0,1).$$

- Per tant, si acceptem H_0 en el cas:

$$\hat{P} \in \left[p_0 - z_{1-\alpha/2}\sqrt{p_0(1-p_0)/n}, p_0 + z_{1-\alpha/2}\sqrt{p_0(1-p_0)/n}\right]$$

i rebutgem H_0 en el cas contrari, tidrem un procediment de contrast de significació exactament igual a α.

Contrast sobre la igualtat de proporcions p_1 i p_2 de dues poblacions binomials o de Bernoulli amb mostres de grans mides

- Contrast $\left\{ \begin{array}{ll} H_0: & p_1 = p_2 \\ H_1: & p_1 \neq p_2 \end{array} \right\}$ amb significació α.

- Si H_0 és certa:

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}_3(1-\hat{P}_3)(1/n_1 + 1/n_2)}} \sim_{\text{aprox.}} \mathrm{N}(0,1), \text{ on } \hat{P}_3 = \frac{\hat{P}_1 n_1 + \hat{P}_2 n_2}{n_1 + n_2}.$$

- Per tant, si acceptem H_0 en el cas:

$$\hat{P}_1 - \hat{P}_2 \in [\, -z_{1-\alpha/2}\sqrt{\hat{P}_3(1-\hat{P}_3)(1/n_1 + 1/n_2)}, \\ z_{1-\alpha/2}\sqrt{\hat{P}_3(1-\hat{P}_3)(1/n_1 + 1/n_2)}\,]$$

i rebutgem H_0 en el cas contrari, tidrem un procediment de contrast de significació exactament igual a α.

Hi ha molts més contrastos: per a la bondat d'ajustament d'una mostra a una població, per a la independència entre dues poblacions, per a l'adequació d'un model de regressió lineal, etc. Alguns d'aquestos es mostren amb detall a la continuació [11].

Un tipus de contrast molt important és el **contrast de normalitat**. No entra en l'exposició teòrica del curs, però sí que es treballa un d'aquests (el de Shapiro-Wilks) a les pràctiques d'ordinador. Són importants sobretot quan es disposa de mostres petites, ja que per a utilitzar els contrastos paramètrics amb aquestes mostres, és necessari que les dades pertanguen a una població normal. Si no, les conclusions sobre el nivell de significació no són correctes.

10.3 Exercicis proposats

Exercici 10.3.1 *Els temps de lectura de dues memòries A i B segueixen lleis normals, i es van a comparar. Se suposen iguals de partida, però es sospita que són diferents. Les mostres de temps de lectura són*

A	*2.1*	*2.2*	*2.5*	*2.3*	*2.5*	*2.4*	*2.2*	*2.6*	*2.5*	*2.1*
B	*1.8*	*2.2*	*2.1*	*2.3*	*2.1*	*2.2*	*2.5*	*2.1*	*2.2*	*2.1*

Prenent una significació del 5%, hi ha evidències estadístiques d'una diferència en els temps?

Exercici 10.3.2 *La insatisfacció dels usuaris amb el sistema operatiu amb què treballen, es sondeja per enquesta: dels 35 usuaris de Windows, 12 es queixen dels seus defectes, mentre que això passa amb només 6 dels 29 usuaris de Linux.*

Demostren aquestes dades una evidència (amb significació del 5%) que els usuaris de Linux estan més satisfets?

Exercici 10.3.3 *Una mostra aleatòria de 36 cigarretes d'una determinada marca donà un contingut mitjà de nicotina de 3 mil·ligrams. El contingut en nicotina d'aquestes cigarretes segueix la llei normal amb una desviació estàndard d'1 mil·ligram. El fabricant garanteix que el contingut mitjà de nicotina és de* 2.9 *mil·ligrams, què pot dir-se d'acord amb les dades obtingudes?*

Exercici 10.3.4 *Els següents nombres representen el temps (en minuts) que van tardar 15 operaris en familiaritzar-se amb el funcionament d'una nova màquina adquirida per l'empresa: 3.4, 2.8, 4.4, 2.5, 3.3, 4, 4.8, 2.9, 5.6, 5.2, 3.7, 3, 3.6, 2.8, 4.8. Suposem que els temps es distribueixen normalment i que el representant que ven la màquina afirma que és suficient amb una mitjana de 3 minuts per a familiaritzar-se amb ella. No obstant el comprador sospita que el temps mitjà requerit pels treballadors és major que 3 minuts. Què es pot dir d'acord amb les dades?*

Exercici 10.3.5 *Un fabricant de bateries per automòbils assegura que les bateries que produeix duren una mitjana de 2 anys, amb una desviació típica de 0.5 anys. Si 5 d'aquestes bateries tenen una duració de 1.5, 2.5, 2.9, 3.2, 4*

anys, determineu un interval de confiança del 95% per a la variància i indica si és vàlida l'afirmació del fabricant.

10.4 Pràctica R: 8. Estimació i proves d'hipòtesi sobre paràmetres de models coneguts

Objectius

L'objectiu d'aquesta pràctica consisteix en fer servir el programa per les tasques d'inferència estadística paramètrica sobre poblacions aleatòries, com són:

- L'estimació de paràmetres per intervals de confiança.
- La decisió sobre acceptar o rebutjar valors de paràmetres en poblacions normals o binomials.

Intervals de confiança de nivell de confiança $1-\alpha$ sobre el valor del paràmetre θ d'una distribució

Quan assumim que un procés aleatori està governat per una distribució concreta (binomial, Poisson, uniforme, exponencial, normal...), queda per determinar quin és el valor concret del paràmetre, que denotarem amb la lletra θ.

Decidir sobre aquest valor desconegut θ és la principal tasca de l'inferència estadística, i l'element essencial per donar una estimació d'eixe valor és obtenir una mostra, que denotarem per $x_1, x_2, \ldots, x_n$.

Un mètode per a ubicar aquest valor desconegut del paràmetre és el de l'interval de confiança. En grans trets, aquest mètode fa el següent:

- Triar un nivell de confiança $1-\alpha$ alt, pròxim a 1 (i per tant un nivell de risc α pròxim a 0). Valors estàndard són 0.95 i 0.99 (és a dir, 0.05 i 0.01 per a α).
- Definir un estadístic T que involucre la mostra i el valor desconegut θ, de manera que tinga una distribució coneguda. Aleshores podem escriure $T = f(\theta)$ amb f creixent.
- Calcular l'interval $[a, b]$ de manera que $P(T \in [a, b]) = 1-\alpha$ amb els valors més versemblants de l'estadístic T.
- Ara aïllem el paràmetre, és a dir
  - $T \in [a, b]$ amb probabilitat $1-\alpha$,
  - $a \leq T \leq b$ amb probabilitat $1-\alpha$,
  - $a \leq f(\theta) \leq b$ amb probabilitat $1-\alpha$,
  - $f^{-1}(a) \leq \theta \leq f^{-1}(b)$ amb "probabilitat" $1-\alpha$,
  - $\theta \in [f^{-1}(a), f^{-1}(b)]$ amb "probabilitat" $1-\alpha$.

Per tant $[f^{-1}(a), f^{-1}(b)]$ és un interval que té dins el valor desconegut del paràmetre θ amb una confiança de $1-\alpha$ (no podem dir 'probabilitat' perquè θ no és una variable aleatòria, sinó un valor concret, encara que desconegut).

Proves d'hipòtesi amb significació α sobre el valor del paràmetre θ d'una distribució

Un segon mètode d'inferència estadística és el de comprovar que una situació inicial (suposadament certa) ha canviat a una situació alternativa. L'estructura és:

$$\left\{ \begin{array}{l} H_0 : \text{situació inicial} \\ H_1 : \text{situació alternativa} \end{array} \right\}$$

On les situacions representen condicions sobre el valor θ desconegut. Per exemple, una màquina està dissenyada per a tallar peces de 3.04 mm, i després d'un mes volem verificar que encara fa les peces d'aqueixa llargària mitjana. Aleshores la prova d'hipòtesi seria:

$$\left\{ \begin{array}{l} H_0 : \mu = 3.04 \\ H_1 : \mu \neq 3.04 \end{array} \right\}$$

S'ha de decidir entre acceptar H_0 i rebutjar H_0 (en favor d'H_1), usant un procediment estadístic, que parteix d'una mostra obtinguda per a tal fi.

No es pot saber si H_0 és certa o no, però la probabilitat de rebutjar H_0 quan és vertadera ha de ser xicoteta i controlada. Aqueixa probabilitat s'anomena "error tipus I" o "significació" de la prova, i es representa amb α.

A grans trets, el procediment per a decidir una prova d'hipòtesi és el següent:

- Triar una significació α xicoteta, pròxima a 0. Valors estàndard són 0.05 i 0.01.
- Definir un estadístic T que involucre la mostra i el valor desconegut θ, de manera que si H_0 es suposa veritable, aleshores tinga una distribució coneguda.
- Calcular T per a la mostra concreta (denotat com T_M)
- Calcular la probabiliat $P(T$ "més rara que" $T_M)$, on l'esdeveniment, ser "més rara que" significa que siguen valors menys versemblants per a la distribució. Aquest valor s'anomena p-valor.
- Ara es decideix: si rebutgem H_0 només quan p-valor$< \alpha$, aleshores tindrem exactament una probabilitat d'equivocar-nos de α, que és el que voliem controlar.

La hipòtesi H_0 es sol referir com a **hipòtesi nul·la**, mentre que la hipòtesi H_1 es diu **hipòtesi alternativa**.

Relació entre intervals de confiança i proves d'hipòtesi paramètriques

El mecanisme de calcular els intervals de confiança i els p-valors de les proves d'hipòtesis nul·les és, en el fons, el mateix. Per això R integra en una sola funció (prova d'hipòtesi = test) els dos procediments.

Per tant, si només vols calcular un interval de confiança, has d'invocar una prova d'hipòtesi, inventant-te (no importa quin valor poses) el valor de la hipòtesi nul·la

Recorda que en una prova d'hipòtesi, l'investigador fixa la significació α a priori (que és el complementari del nivell de confiança als intervals de confiança). Després, la decisió d'acceptar H_0 o rebutjar-la es pot fer de dues maneres:

- Accepta H_0 només si el valor de comparació està dins l'interval de confiança.
- Accepta H_0 només si el p-valor de la mostra és major o igual a la significació α.

Atés que R torna sempre el p-valor de la mostra, és més senzill optar per la segona metodologia.

La prova t per decidir sobre hipòtesis relatives a la mitjana d'una població o mitjanes de dues poblacions

Per a contrastar la hipòtesi del valor de la mitjana μ d'una població normal (o d'una població qualsevol si la mida de la mostra és gran) o per calcular l'interval de confiança, una estadístic relacionat amb la mitjana mostral segueix la distribució t-Student. La funció que realitza la prova d'hipòtesi i l'interval de confiança és:

```
t.test(x, y = NULL,
        alternative = c("two.sided", "less", "greater"),
        mu = 0, paired = FALSE, var.equal = FALSE,
        conf.level = 0.95, ...)
```

on:

- `x`: vector que conté la mostra observada de la població
- `y`: `NULL` per defecte (si només es contrasta el valor de la mitjana d'una població). Vector que conté la mostra observada de la segona població quan es comparen les mitjanes de dues poblacions.
- `alternative`: indica la forma de la hipòtesi alternativa; $\neq$, $<$,$>$, respectivament `"two.sided"` (per defecte), `"less"`, `"greater"`.
- `mu`: valor de μ_0 (o de $\mu_1 - \mu_2$) en la hipòtesi nul·la. Val 0 per defecte.

- `paired`: només en el cas de contrastar dues poblacions, valor lògic que indica si les mostres són emparellades (no independents). És un detall important i es pot deduir de la forma en què s'han obtingut les mostres.
- `var.equal`: lògic que indica si les variàncies de les poblacions contrastades es poden considerar iguals.
- `conf.level`: nivell de confiança (valor entre 0 i 1, val 0.95 per defecte). És el complementari del nivell de significació α.

La funció torna un objecte de classe llista amb les següents components:

- `statistic`: el valor de l'estadístic del contrast.
- `parameter`: els graus de llibertat de l'estadístic.
- `p.value`: el p-valor calculat en la mostra.
- `conf.int`: l'interval de confiança per a la mitjana (o la diferència entre les dues mitjanes) corresponent al nivell de confiança i a la hipòtesi alternativa especificades.
- `estimate`: la mitjana (o diferència de mitjanes) mostral.
- `null.value`: el valor determinat per la hipòtesi nul·la.
- `alternative`: descripció de la hipòtesi alternativa.
- `method`: descripció del tipus de contrast realitzat (una o dues mostres).
- `data.name`: descripció de les dades usades com mostra.

La informació més important s'obté per pantalla en fer la invocació a la funció.

Per a les seccions que continuen assumim que si es té la mostra d'una sola població, aquesta estarà emmagatzemada a la variable `mostra`, mentre que si tenim dues mostres de dues poblacions respectives, aquestes estaran emmagatzemades e les variables `mostra1` i `mostra2`, respectivament.

També assumim que el nivell de significació de les proves d'hipòtesi és α, i el corresponent nivell de confiança és $1-\alpha$.

Les proves d'hipòtesi sobre la mitjana es presenten a continuació en totes les seues variants.

Sobre el valor de la mitjana μ d'una població normal (o qualsevol si la mida mostral és gran) amb variància desconeguda

- $\left\{ \begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{array} \right\}$

```
t.test(x=mostra, mu = μ0, conf.level = 1 − α)
```

- $\left\{ \begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{array} \right\}$

```
t.test(x=mostra, mu = μ0, conf.level = 1 − α,
          alternative="greater")
```

- $\left\{ \begin{array}{ll} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{array} \right\}$

```
t.test(x=mostra, mu = μ0, conf.level = 1 − α,
          alternative="less")
```

Sobre la comparació de les mitjanes μ_1 i μ_2 de dues poblacions normals (o qualsevol si la mida mostral és gran) dependents o emparellades

El contrast d'igualtat de mitjanes es correspon amb l'elecció de $\mu_0 = 0$.

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 \neq \mu_0 \end{array} \right\}$

```
t.test(x=mostra1, y=mostra2, mu = μ0, conf.level = 1 − α,
          paired=TRUE)
```

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 > \mu_0 \end{array} \right\}$

```
t.test(x=mostra1, y=mostra2, mu = μ0, conf.level = 1 − α,
          alternative="greater", paired=TRUE)
```

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 < \mu_0 \end{array} \right\}$

```
t.test(x=mostra1, y=mostra2, mu = μ0, conf.level = 1 − α,
          alternative="less", paired=TRUE)
```

Sobre la comparació de les mitjanes μ_1 i μ_2 de dues poblacions normals (o qualsevol si la mida mostral és gran) independents amb variàncies completament desconegudes

El contrast d'igualtat de mitjanes es correspon amb l'elecció de $\mu_0 = 0$.

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 \neq \mu_0 \end{array} \right\}$

```
t.test(x=mostra1, y=mostra2, mu = μ0, conf.level = 1 − α)
```

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 > \mu_0 \end{array} \right\}$

```
t.test(x=mostra1, y=mostra2, mu = μ0, conf.level = 1 − α,
          alternative="greater")
```

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 < \mu_0 \end{array} \right\}$

```
t.test(x=mostra1, y=mostra2, mu = μ0, conf.level = 1 − α,
          alternative="less")
```

Sobre la comparació de les mitjanes μ_1 i μ_2 de dues poblacions normals (o qualsevol si la mida mostral és gran) independents amb variàncies desconegudes però suposadament iguals

El contrast d'igualtat de mitjanes es correspon amb l'elecció de $\mu_0 = 0$.

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 \neq \mu_0 \end{array} \right\}$

  `t.test(x=mostra1, y=mostra2, mu =` μ_0`, conf.level =` $1-\alpha$`, var.equal=TRUE)`

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 > \mu_0 \end{array} \right\}$

  `t.test(x=mostra1, y=mostra2, mu =` μ_0`, conf.level =` $1-\alpha$`, alternative="greater", var.equal=TRUE)`

- $\left\{ \begin{array}{ll} H_0: & \mu_1 - \mu_2 = \mu_0 \\ H_1: & \mu_1 - \mu_2 < \mu_0 \end{array} \right\}$

  `t.test(x=mostra1, y=mostra2, mu =` μ_0`, conf.level =` $1-\alpha$`, alternative="less", var.equal=TRUE)`

La prova F per a decidir sobre hipòtesis relatives a la comparació de variàncies de dues poblacions

R no té implementat un contrast sobre el valor de la variància d'una població normal, però sí el contrast per a la comparació de les variàncies σ_1^2 i σ_2^2 de dues poblacions normals mitjançant un estadístic F-Snedecor. Per tant, si les mostres de dues poblacions normals estan emmagatzemades a les variables `mostra1` i `mostra2`, per trobar l'interval de confiança de nivell $1-\alpha$ de la ratio σ_1^2/σ_2^2 o la decisió sobre el contrast

$$\left\{ \begin{array}{lll} H_0: & \sigma_1^2/\sigma_2^2 = \sigma_0 & \\ H_1: & \sigma_1^2/\sigma_2^2 \neq \sigma_0 & (\text{o } \sigma_1^2/\sigma_2^2 < \sigma_0, \text{ o } \sigma_1^2/\sigma_2^2 > \sigma_0) \end{array} \right\}$$

amb significació α, s'ha d'invocar la funció `var.test()` amb l'estructura:

`var.test(x=mostra1, y=mostra2, ratio =`σ_0`, ++conf.level =` $1-\alpha$`)`

on s'han d'usar les opcions `alternative="less"` o `alternative="greater"` segons el cas, si la hipòtesi H_1 no és la usada per defecte ($\neq$). L'argument `ratio=1` pot servir per hipòtesis del tipus $H_0 : \sigma_1^2 = \sigma_2^2$ d'igualtat entre variàncies.

La funció torna un objecte de classe llista amb les següents components:

- `statistic`: el valor de l'estadístic del contrast.
- `parameter`: els graus de llibertat de l'estadístic.
- `p.value`: el p-valor calculat en la mostra.

- `conf.int`: l'interval de confiança per a la ràtio entre les dues variàncies corresponent al nivell de confiança i a la hipòtesi alternativa especificades.
- `estimate`: la ràtio de variàncies mostrals.
- `null.value`: el valor determinat per la hipòtesi nul·la.
- `alternative`: descripció de la hipòtesi alternativa.
- `method`: descripció del tipus de contrast realitzat (una o dues mostres).
- `data.name`: descripció de les dades usades com a mostra.

La informació més important s'obté per pantalla en fer la invocació a la funció.

El contrast sobre les proporcions de la binomial o Bernoulli

- **Una mostra**: si sabem que una mostra de mida n d'una prova de Bernoulli amb paràmetre p desconegut, consta de x èxits, aleshores podem realitzar la prova d'hipòtesi

$$\left\{ \begin{array}{ll} H_0: & p = p_0 \\ H_1: & p \neq p_0 \quad (\text{o } p < p_0, \text{ o } p > p_0) \end{array} \right\}$$

  de significació α o un interval de confiança de nivell $1-\alpha$ sobre el valor de la proporció p_0, mitjançant la funció:

  `prop.test(x=`x`, n=`n`, p=`p_0`, conf.level=`$1-\alpha$`)`

  Recorda usar `alternative="less"` o `alternative="greater"` segons el cas, si la hipòtesi H_1 no és la usada per defecte ($\neq$).

- **Dues mostres**: Si sabem que una mostra de mida n_1 d'una prova de Bernoulli amb paràmetre p_1 desconegut, consta de x_1 èxits, i que una altra mostra de mida n_2 d'una prova de Bernoulli amb paràmetre p_2 desconegut, consta de x_2 èxits, aleshores podem realitzar la prova d'hipòtesi sobre la igualtat de les proporcions

$$\left\{ \begin{array}{ll} H_0: & p_1 = p_2 \\ H_1: & p_1 \neq p_2 \quad (\text{o } p_1 < p_2, \text{ o } p_1 > p_2) \end{array} \right\}$$

  de significació α o un interval de confiança de nivell $1-\alpha$ sobre el valor de la diferència $p_1 - p_2$, mitjançant la funció:

  `prop.test(x=c(`x_1`,`x_2`), n=c(`n_1`,`n_2`), conf.level=`$1-\alpha$`)`

  Recorda usar `alternative="less"` o `alternative="greater"` segons el cas, si la hipòtesi H_1 no és la usada per defecte ($\neq$).

L'objecte tornat per la funció és de tipus llista amb les components:

- `statistic`: el valor de l'estadístic del contrast.
- `parameter`: els graus de llibertat de l'estadístic.

- `p.value`: el p-valor calculat en la mostra.
- `estimate`: les proporcions mostrals.
- `conf.int`: l'interval de confiança per a la proporció (o diferència entre proporcions si són dues poblacions) al nivell de confiança i a la hipòtesi alternativa especificades.
- `null.value`: el valor determinat per la hipòtesi nul·la.
- `alternative`: descripció de la hipòtesi alternativa.
- `method`: descripció del mètode i de si la correcció per continuïtat s'ha usat.
- `data.name`: descripció de les dades usades com a mostra.

La informació més important s'obté per pantalla en fer la invocació a la funció. Cal destacar que aquest contrast és el mateix que el presentat a la teoria (que utilitza l'aproximació normal), raó per la qual s'ha de comprovar que la mida de la mostra és prou gran.

Una prova de "normalitat"

El contrast de normalitat sobre una població és molt important perquè, per poder aplicar moltes tècniques estadístiques, es necessita acceptar que les dades que s'estudien siguen generades per una distribució normal (en cas contrari, les conclusions d'aquestes tècniques no tindrien validesa científica).

Aleshores, diversos investigadors han desenvolupat proves de normalitat que, com sempre, no poden demostrar res, però donen una confiança (un percentatge concret i que es pot prendre com a base per prendre decisions) de si la població de la que s'ha tret la mostra segueix una llei normal o no.

La prova de normalitat

$$\left\{ \begin{array}{l} H_0 : \text{la mostra pertany a una variable amb distribució normal} \\ H_1 : \text{no } H_0 \end{array} \right\}$$

que usarem i que està implementada en R és la prova de Shapiro-Wilks. Si la mostra està emmagatzemada en la variable `mostra`, la prova s'invoca mitjançant:

```
shapiro.test(x=mostra)
```

Aquesta funció calcula un estadístic (denotat per W i no donat en la teoria del curs IG12-0607) i el p-valor corresponent a la mostra disponible. Comparant-lo amb la significació es decideix si s'accepta o no que la mostra pertany a una població de dades que segueix la llei normal.

Exercicis d'ensinistrament

1. Una màquina fabrica peces de 50 mm, i ha estat funcionant correctament durant un temps. Ha canviat l'operador de manteniment, i aquest home ha agafat una mostra de 40 peces, mesurant-les, perquè sospita d'un mal funcionament de la màquina. L'arxiu `dades-pr4-s5-peces.txt` conté les llargàries mesurades.

   Com expert a l'empresa, l'encarregat et demana que investigues si la màquina ha canviat realment de llargària mitjana o ha sigut degut a l'atzar (usant un nivell de confiança del 95%). Per tant el contrast seria:

$$\left\{ \begin{array}{ll} H_0: & \mu = 50 \\ H_1: & \mu \neq 50 \end{array} \right\}$$

2. Un client que rep barrils de petroli es queixa al seu proveïdor perquè no estan plens (42 galons americans, o 158.99 litres). Amb la notificació de la queixa acompanya les dades mesurades del darrers 176 barrils, que figuren a l'arxiu `dades-pr4-s5-barrils.txt`.

   Com encarregat del sistema d'ompliment de barrils, contrasta la hipòtesi de l'empresa amb la del client, usant una significació de l'1%. En aquest cas el contrast seria:

$$\left\{ \begin{array}{ll} H_0: & \mu = 158.99 \\ H_1: & \mu < 158.99 \end{array} \right\}$$

3. Una xicoteta empresa informàtica, de la qual ets soci, tracta de donar un nou servei a la comunitat, inexistent fins ara. Per a no arriscar la inversió econòmica inicial que es necessita, es té l'idea que aproximadament un 10% de la clientela està interessada a contractar aquest nou servei i que, per tant, sí que paga la pena intentar-ho.

   Tú, com a soci i dubtós d'aqueix optimista 10%, penses que no estaria malament sondejar el mercat per comprovar el percentatge, usant una significació del 15%, ja que equivocar-se en estimar el percentatge tindria greus conseqüències. Es tractaria d'enviar una sèrie de correus on es demanaria als clients si contractarien aquest nou servei (a un preu ja pactat). Mirant-te de reüll, però com que no costa molt, els teus socis decideixen fer-te cas, i trobeu que, dels 92 correus enviats i contestats (bé per correu o bé insistint al client per telèfon), només 5 han contestat afirmativament.

   Es pot acceptar encara que hi ha un 10% d'usuaris potencials del nou servei amb les dues significacions indicades? En aquest cas, el contrast seria:

$$\left\{ \begin{array}{ll} H_0: & p = 0.1 \\ H_1: & p < 0.1 \end{array} \right\}$$

4. Una fàbrica de cristalls per satèl·lits-telescopi (com el famós *Hubble*) té dues sucursals a Nova York i Bonn. Els cristalls es dissenyen de la manera

més plana possible, mesurant-se aquesta amb un coeficient $\lambda > 0.0$ (a menor valor de λ, més pla i perfecte és el cristall).

S'acaben d'introduir canvis en les dues sucursals. Per a saber si encara treballen de manera similar, es mesura la λ de 7 cristalls enviats per cada sucursal, amb el resultat a l'arxiu `dades-pr4-s5-cristalls.txt`. El procés de mesurar amb molta precisió és molt costós i lent, raó per la qual no et pot agafar una mostra major.

Es pot acceptar que les dues sucursals donen productes de la mateixa qualitat? És a dir, quin resultat dóna el contrast

$$\left\{ \begin{array}{ll} H_0: & \mu_{\text{NY}} = \mu_{\text{B}} \\ H_1: & \mu_{\text{NY}} \neq \mu_{\text{B}} \end{array} \right\}$$

(Ajuda: la mostra és molt xicoteta, per tant has de comprovar alguns detalls tècnics sobre les dades obtingudes).

10.5 Pràctica R: 9. Recopilatòria

Objectius

Resoldre problemes de plantejaments globals, on s'ha d'usar més d'una tècnica indicada al llibre.

Exercicis d'ensinistrament

1. La màquina que talla l'acer que després es doblega convertint-se en clips, està ajustada per tallar trossos de 60.0 mm. Per a analitzar el correcte funcionament de la màquina, es registra la llargària d'una mostra de peces, que figura a la variable `clips`.

   (a) (Descripció de la mostra) Dibuixa un histograma de les llargàries de les peces.

   (b) (Descripció de la mostra) Una peça que té una llargària fora de l'interval $[58.0, 62.0]$ mm es considera "perillosa" perquè pot donar problemes a la màquina que les doblega per formar els clips. Quin percentatge de les peces de la mostra és perillós per a la màquina dobladora? Sol.: 8.1%

   (c) (Models de probabilitat) A la vista de l'histograma, i per la natura de la variable analitzada, quina distribució de probabilitat sembla que segueix la variable "llargària de la peça"?

   (d) (Inferència estadística) Contrasta la possibilitat que, efectivament, la variable "llargària de la peça" seguisca la distribució que has contestat a l'apartat anterior amb un nivell de significació de l'1%? Sol.: Acceptem que sí que la segueix

(e) (Inferència estadística) Contrasta ara el fet que la llargària mitjana amb què està tallant les peces és efectivament 60.0 mm, amb un nivell de risc del 5%. Sol.: Acceptem que és 60.0 mm

Necessitaves la resposta afirmativa de l'apartat anterior per a poder fer aquest? (Sí/No) Per què?

(f) (Models de probabilitat) Si el contrast de l'apartat anterior també és positiu, i usant la variància mostral com si fóra la variància de la distribució de la variable, calculeu un percentatge aproximat de peces que poden posar en perill la màquina doblegadora en un llarg període de temps. Sol.: 8.79%

2. En les darreres enquestes d'intenció de vot d'un país, els resultats obtinguts estan emmagatzemats en les variables `taula.enquesta.passada` i `taula.enquesta.actual`.

   El partit A vol analitzar estadísticament les enquestes per tal d'indagar si realment la intenció de vot real (no només la resultant de les enquestes) sobre el seu partit és la mateixa, o si ha canviat. Per això:

   (a) Calculeu la proporció de votants (o percentatge) del partit A en cada enquesta. Sol.: 39.84% i 35.35%

   (b) (Descripció de la mostra) Crea un diagrama (de sectors) amb el resultat de votants per partit per cada enquesta.

   (c) (Inferència estadística) Contrasta la hipòtesi que la proporció de votants del partit A dins de l'electorat (no només dins la mostra recollida a l'enquesta) ha canviat o no respecte del sondeig anterior amb una significació estàndard. Sol.: S'accepta que no ha canviat

3. Es vol fer un estudi sobre l'efecte de la implantació del carnet per punts a la sinistralitat vial. A tal fi s'arrepleguen dades en la variable `accidents`, que emmagatzema, per cada dia laborable que no és vespra de festiu, el nombre d'accidents ocorreguts i el període (si era abans amb el carnet tradicional o ara que es té el carnet per punts). Per tant:

   (a) (Descripció de la mostra) Calculeu la mitjana d'accidents diaris ocorreguts amb cada tipus de carnet. Amb quin tipus és major la mitjana mostral? Sol.: 8.35 i 6.83 (major amb el carnet tradicional)

   (b) (Descripció de la mostra) Representeu un gràfic on es puga comparar bé les distribucions del nombre d'accidents diaris abans i després del carnet per punts.

   (c) (Models de probabilitat) A priori, per la natura de la variable, quina distribució teòrica, de les estudiades en el curs, hauria de seguir la variable "nombre d'accidents ocorreguts en un dia" (tant abans com després d'introduir-se el carnet per punts)?

   (d) (Models de probabilitat) Si la distribució que has contestat en l'apartat anterior és correcta, i agafes com valor del seu paràmetre el valor de la mitjana mostral corresponent, quina probabilitat existeix

que un dia laborable no vespra de festiu de la pròxima setmana, hi haja menys de 10 accidents? Sol.: 0.8464

(e) (Inferència estadística) Es pot acceptar científicament, en base a les dades de la mostra, amb un nivell estàndard que el carnet per punts ha fet baixar la sinistralitat o podria ser la mateixa? Fes el contrast que calga. Sol.: Acceptem que la mitjana ha baixat.

4. Es tracta de determinar l'efecte que una droga té sobre el nivell de colesterol 'dolent' en sang en pacients amb el colesterol elevat. Per tant es tria un grup de 50 pacients amb nivell de colesterol quasi idèntic (al voltant de 200 mg/dl) als quals se'ls administra una quantitat variable de la droga, mesurant-se aquesta quantitat (en mg) i el nivell de colesterol del dia següent, amb els resultats que figuren en la variable `colest`.

   Tenint en compte que un nivell de colesterol 'dolent' acceptable és de 100 mg/dl, i que no convé prendre més quantitat de droga que la necessària, quina dosi es recomanaria a aquest perfil de pacients per tal d'aconseguir aqueix nivell acceptable de colesterol? D'acord amb quina tècnica? (Sol.: 12.41 mg)

PART V
TAULES ESTADÍSTIQUES

Presentem les taules estadístiques més habituals per al càlcul manual de probabilitats en els models binomial, de Poisson, normal tipificat, i el càlcul de quantils (útils pel càlcul d'intervals de confiança i per a la decisió de les proves d'hipòtesi) de les distribucions χ^2, t-Student i F-Snedecor.

La utilització del programa R, estalvia completament l'ús d'aquest tipus de taules. Concretament, a la secció 7.12, la Pràctica R: 7 descriu, entre d'altres, la forma de calcular els valors de la funció de probabilitat o distribució dels models habituals, així com el càlcul de quantils.

Taula 10.2: Funció de distribució acumulada de la (distribució) binomial de paràmetres n i p.

		p									
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	0	0.9025	0.81	0.7225	0.64	0.5625	0.49	0.4225	0.36	0.3025	0.25
2	1	0.9975	0.99	0.9775	0.96	0.9375	0.91	0.8775	0.84	0.7975	0.75
2	2	1	1	1	1	1	1	1	1	1	1
3	0	0.8574	0.729	0.6141	0.512	0.4219	0.343	0.2746	0.216	0.1664	0.125
3	1	0.9928	0.972	0.9393	0.896	0.8438	0.784	0.7183	0.648	0.5748	0.5
3	2	0.9999	0.999	0.9966	0.992	0.9844	0.973	0.9571	0.936	0.9089	0.875
3	3	1	1	1	1	1	1	1	1	1	1
4	0	0.8145	0.6561	0.522	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
4	1	0.986	0.9477	0.8905	0.8192	0.7383	0.6517	0.563	0.4752	0.391	0.3125
4	2	0.9995	0.9963	0.988	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
4	3	1	0.9999	0.9995	0.9984	0.9961	0.9919	0.985	0.9744	0.959	0.9375
4	4	1	1	1	1	1	1	1	1	1	1
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.116	0.0778	0.0503	0.0312
5	1	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.337	0.2562	0.1875
5	2	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5
5	3	1	0.9995	0.9978	0.9933	0.9844	0.9692	0.946	0.913	0.8688	0.8125
5	4	1	1	0.9999	0.9997	0.999	0.9976	0.9947	0.9898	0.9815	0.9688
5	5	1	1	1	1	1	1	1	1	1	1
6	0	0.7351	0.5314	0.3771	0.2621	0.178	0.1176	0.0754	0.0467	0.0277	0.0156
6	1	0.9672	0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
6	2	0.9978	0.9841	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438
6	3	0.9999	0.9987	0.9941	0.983	0.9624	0.9295	0.8826	0.8208	0.7447	0.6562
6	4	1	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.959	0.9308	0.8906
6	5	1	1	1	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
6	6	1	1	1	1	1	1	1	1	1	1
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.049	0.028	0.0152	0.0078
7	1	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
7	2	0.9962	0.9743	0.9262	0.852	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
7	3	0.9998	0.9973	0.9879	0.9667	0.9294	0.874	0.8002	0.7102	0.6083	0.5
7	4	1	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
7	5	1	1	0.9999	0.9996	0.9987	0.9962	0.991	0.9812	0.9643	0.9375
7	6	1	1	1	1	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
7	7	1	1	1	1	1	1	1	1	1	1
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
8	1	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
8	2	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
8	3	0.9996	0.995	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.477	0.3633
8	4	1	0.9996	0.9971	0.9896	0.9727	0.942	0.8939	0.8263	0.7396	0.6367
8	5	1	1	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
8	6	1	1	1	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
8	7	1	1	1	1	1	0.9999	0.9998	0.9993	0.9983	0.9961
8	8	1	1	1	1	1	1	1	1	1	1
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.002
9	1	0.9288	0.7748	0.5995	0.4362	0.3003	0.196	0.1211	0.0705	0.0385	0.0195
9	2	0.9916	0.947	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
9	3	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
9	4	1	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5
9	5	1	0.9999	0.9994	0.9969	0.99	0.9747	0.9464	0.9006	0.8342	0.7461
9	6	1	1	1	0.9997	0.9987	0.9957	0.9888	0.975	0.9502	0.9102
9	7	1	1	1	1	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
9	8	1	1	1	1	1	1	0.9999	0.9997	0.9992	0.998
9	9	1	1	1	1	1	1	1	1	1	1
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.006	0.0025	0.001
10	1	0.9139	0.7361	0.5443	0.3758	0.244	0.1493	0.086	0.0464	0.0233	0.0107
10	2	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
10	3	0.999	0.9872	0.95	0.8791	0.7759	0.6496	0.5138	0.3823	0.266	0.1719
10	4	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.377
10	5	1	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.623
10	6	1	1	0.9999	0.9991	0.9965	0.9894	0.974	0.9452	0.898	0.8281
10	7	1	1	1	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
10	8	1	1	1	1	1	0.9999	0.9995	0.9983	0.9955	0.9893
10	9	1	1	1	1	1	1	1	0.9999	0.9997	0.999
10	10	1	1	1	1	1	1	1	1	1	1

Continua darrere

		p									
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
11	0	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005
11	1	0.8981	0.6974	0.4922	0.3221	0.1971	0.113	0.0606	0.0302	0.0139	0.0059
11	2	0.9848	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327
11	3	0.9984	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133
11	4	0.9999	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5328	0.3971	0.2744
11	5	1	0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5
11	6	1	1	0.9997	0.998	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256
11	7	1	1	1	0.9998	0.9988	0.9957	0.9878	0.9707	0.939	0.8867
11	8	1	1	1	1	0.9999	0.9994	0.998	0.9941	0.9852	0.9673
11	9	1	1	1	1	1	1	0.9998	0.9993	0.9978	0.9941
11	10	1	1	1	1	1	1	1	1	0.9998	0.9995
11	11	1	1	1	1	1	1	1	1	1	1
12	0	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002
12	1	0.8816	0.659	0.4435	0.2749	0.1584	0.085	0.0424	0.0196	0.0083	0.0032
12	2	0.9804	0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193
12	3	0.9978	0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.073
12	4	0.9998	0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938
12	5	1	0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872
12	6	1	0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128
12	7	1	1	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062
12	8	1	1	1	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.927
12	9	1	1	1	1	1	0.9998	0.9992	0.9972	0.9921	0.9807
12	10	1	1	1	1	1	1	0.9999	0.9997	0.9989	0.9968
12	11	1	1	1	1	1	1	1	1	0.9999	0.9998
12	12	1	1	1	1	1	1	1	1	1	1
13	0	0.5133	0.2542	0.1209	0.055	0.0238	0.0097	0.0037	0.0013	0.0004	0.0001
13	1	0.8646	0.6213	0.3983	0.2336	0.1267	0.0637	0.0296	0.0126	0.0049	0.0017
13	2	0.9755	0.8661	0.692	0.5017	0.3326	0.2025	0.1132	0.0579	0.0269	0.0112
13	3	0.9969	0.9658	0.882	0.7473	0.5843	0.4206	0.2783	0.1686	0.0929	0.0461
13	4	0.9997	0.9935	0.9658	0.9009	0.794	0.6543	0.5005	0.353	0.2279	0.1334
13	5	1	0.9991	0.9925	0.97	0.9198	0.8346	0.7159	0.5744	0.4268	0.2905
13	6	1	0.9999	0.9987	0.993	0.9757	0.9376	0.8705	0.7712	0.6437	0.5
13	7	1	1	0.9998	0.9988	0.9944	0.9818	0.9538	0.9023	0.8212	0.7095
13	8	1	1	1	0.9998	0.999	0.996	0.9874	0.9679	0.9302	0.8666
13	9	1	1	1	1	0.9999	0.9993	0.9975	0.9922	0.9797	0.9539
13	10	1	1	1	1	1	0.9999	0.9997	0.9987	0.9959	0.9888
13	11	1	1	1	1	1	1	1	0.9999	0.9995	0.9983
13	12	1	1	1	1	1	1	1	1	1	0.9999
13	13	1	1	1	1	1	1	1	1	1	1
14	0	0.4877	0.2288	0.1028	0.044	0.0178	0.0068	0.0024	0.0008	0.0002	0.0001
14	1	0.847	0.5846	0.3567	0.1979	0.101	0.0475	0.0205	0.0081	0.0029	0.0009
14	2	0.9699	0.8416	0.6479	0.4481	0.2811	0.1608	0.0839	0.0398	0.017	0.0065
14	3	0.9958	0.9559	0.8535	0.6982	0.5213	0.3552	0.2205	0.1243	0.0632	0.0287
14	4	0.9996	0.9908	0.9533	0.8702	0.7415	0.5842	0.4227	0.2793	0.1672	0.0898
14	5	1	0.9985	0.9885	0.9561	0.8883	0.7805	0.6405	0.4859	0.3373	0.212
14	6	1	0.9998	0.9978	0.9884	0.9617	0.9067	0.8164	0.6925	0.5461	0.3953
14	7	1	1	0.9997	0.9976	0.9897	0.9685	0.9247	0.8499	0.7414	0.6047
14	8	1	1	1	0.9996	0.9978	0.9917	0.9757	0.9417	0.8811	0.788
14	9	1	1	1	1	0.9997	0.9983	0.994	0.9825	0.9574	0.9102
14	10	1	1	1	1	1	0.9998	0.9989	0.9961	0.9886	0.9713
14	11	1	1	1	1	1	1	0.9999	0.9994	0.9978	0.9935
14	12	1	1	1	1	1	1	1	0.9999	0.9997	0.9991
14	13	1	1	1	1	1	1	1	1	1	0.9999
14	14	1	1	1	1	1	1	1	1	1	1

Continua darrere

		p									
n	x	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
15	0	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0
15	1	0.829	0.549	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005
15	2	0.9638	0.8159	0.6042	0.398	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037
15	3	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176
15	4	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592
15	5	0.9999	0.9978	0.9832	0.9389	0.8516	0.7216	0.5643	0.4032	0.2608	0.1509
15	6	1	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036
15	7	1	1	0.9994	0.9958	0.9827	0.95	0.8868	0.7869	0.6535	0.5
15	8	1	1	0.9999	0.9992	0.9958	0.9848	0.9578	0.905	0.8182	0.6964
15	9	1	1	1	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491
15	10	1	1	1	1	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408
15	11	1	1	1	1	1	0.9999	0.9995	0.9981	0.9937	0.9824
15	12	1	1	1	1	1	1	0.9999	0.9997	0.9989	0.9963
15	13	1	1	1	1	1	1	1	1	0.9999	0.9995
15	14	1	1	1	1	1	1	1	1	1	1
15	15	1	1	1	1	1	1	1	1	1	1

Taula 10.3: Funció de distribució acumulada de la (distribució) de Poisson de paràmetre λ.

	λ									
x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
0	0.6065	0.3679	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067
1	0.9098	0.7358	0.5578	0.406	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404
2	0.9856	0.9197	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247
3	0.9982	0.981	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.265
4	0.9998	0.9963	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405
5	1	0.9994	0.9955	0.9834	0.958	0.9161	0.8576	0.7851	0.7029	0.616
6	1	0.9999	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622
7	1	1	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666
8	1	1	1	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319
9	1	1	1	1	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682
10	1	1	1	1	0.9999	0.9997	0.999	0.9972	0.9933	0.9863
11	1	1	1	1	1	0.9999	0.9997	0.9991	0.9976	0.9945
12	1	1	1	1	1	1	0.9999	0.9997	0.9992	0.998
13	1	1	1	1	1	1	1	0.9999	0.9997	0.9993
14	1	1	1	1	1	1	1	1	0.9999	0.9998
15	1	1	1	1	1	1	1	1	1	0.9999
16	1	1	1	1	1	1	1	1	1	1
	λ									
x	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
0	0.0041	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001	0
1	0.0266	0.0174	0.0113	0.0073	0.0047	0.003	0.0019	0.0012	0.0008	0.0005
2	0.0884	0.062	0.043	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042	0.0028
3	0.2017	0.1512	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149	0.0103
4	0.3575	0.2851	0.2237	0.173	0.1321	0.0996	0.0744	0.055	0.0403	0.0293
5	0.5289	0.4457	0.369	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885	0.0671
6	0.686	0.6063	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649	0.1301
7	0.8095	0.744	0.6728	0.5987	0.5246	0.453	0.3856	0.3239	0.2687	0.2202
8	0.8944	0.8472	0.7916	0.7291	0.662	0.5925	0.5231	0.4557	0.3918	0.3328
9	0.9462	0.9161	0.8774	0.8305	0.7764	0.7166	0.653	0.5874	0.5218	0.4579
10	0.9747	0.9574	0.9332	0.9015	0.8622	0.8159	0.7634	0.706	0.6453	0.583
11	0.989	0.9799	0.9661	0.9467	0.9208	0.8881	0.8487	0.803	0.752	0.6968
12	0.9955	0.9912	0.984	0.973	0.9573	0.9362	0.9091	0.8758	0.8364	0.7916
13	0.9983	0.9964	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981	0.8645
14	0.9994	0.9986	0.997	0.9943	0.9897	0.9827	0.9726	0.9585	0.94	0.9165
15	0.9998	0.9995	0.9988	0.9976	0.9954	0.9918	0.9862	0.978	0.9665	0.9513
16	0.9999	0.9998	0.9996	0.999	0.998	0.9963	0.9934	0.9889	0.9823	0.973
17	1	0.9999	0.9998	0.9996	0.9992	0.9984	0.997	0.9947	0.9911	0.9857
18	1	1	0.9999	0.9999	0.9997	0.9993	0.9987	0.9976	0.9957	0.9928
19	1	1	1	1	0.9999	0.9997	0.9995	0.9989	0.998	0.9965
20	1	1	1	1	1	0.9999	0.9998	0.9996	0.9991	0.9984
21	1	1	1	1	1	1	0.9999	0.9998	0.9996	0.9993
22	1	1	1	1	1	1	1	0.9999	0.9999	0.9997
23	1	1	1	1	1	1	1	1	0.9999	0.9999
24	1	1	1	1	1	1	1	1	1	1

Continua darrere

	λ									
x	11	12	13	14	15	16	17	18	19	20
0	0	0	0	0	0	0	0	0	0	0
1	0.0002	0.0001	0	0	0	0	0	0	0	0
2	0.0012	0.0005	0.0002	0.0001	0	0	0	0	0	0
3	0.0049	0.0023	0.0011	0.0005	0.0002	0.0001	0	0	0	0
4	0.0151	0.0076	0.0037	0.0018	0.0009	0.0004	0.0002	0.0001	0	0
5	0.0375	0.0203	0.0107	0.0055	0.0028	0.0014	0.0007	0.0003	0.0002	0.0001
6	0.0786	0.0458	0.0259	0.0142	0.0076	0.004	0.0021	0.001	0.0005	0.0003
7	0.1432	0.0895	0.054	0.0316	0.018	0.01	0.0054	0.0029	0.0015	0.0008
8	0.232	0.155	0.0998	0.0621	0.0374	0.022	0.0126	0.0071	0.0039	0.0021
9	0.3405	0.2424	0.1658	0.1094	0.0699	0.0433	0.0261	0.0154	0.0089	0.005
10	0.4599	0.3472	0.2517	0.1757	0.1185	0.0774	0.0491	0.0304	0.0183	0.0108
11	0.5793	0.4616	0.3532	0.26	0.1848	0.127	0.0847	0.0549	0.0347	0.0214
12	0.6887	0.576	0.4631	0.3585	0.2676	0.1931	0.135	0.0917	0.0606	0.039
13	0.7813	0.6815	0.573	0.4644	0.3632	0.2745	0.2009	0.1426	0.0984	0.0661
14	0.854	0.772	0.6751	0.5704	0.4657	0.3675	0.2808	0.2081	0.1497	0.1049
15	0.9074	0.8444	0.7636	0.6694	0.5681	0.4667	0.3715	0.2867	0.2148	0.1565
16	0.9441	0.8987	0.8355	0.7559	0.6641	0.566	0.4677	0.3751	0.292	0.2211
17	0.9678	0.937	0.8905	0.8272	0.7489	0.6593	0.564	0.4686	0.3784	0.297
18	0.9823	0.9626	0.9302	0.8826	0.8195	0.7423	0.655	0.5622	0.4695	0.3814
19	0.9907	0.9787	0.9573	0.9235	0.8752	0.8122	0.7363	0.6509	0.5606	0.4703
20	0.9953	0.9884	0.975	0.9521	0.917	0.8682	0.8055	0.7307	0.6472	0.5591
21	0.9977	0.9939	0.9859	0.9712	0.9469	0.9108	0.8615	0.7991	0.7255	0.6437
22	0.999	0.997	0.9924	0.9833	0.9673	0.9418	0.9047	0.8551	0.7931	0.7206
23	0.9995	0.9985	0.996	0.9907	0.9805	0.9633	0.9367	0.8989	0.849	0.7875
24	0.9998	0.9993	0.998	0.995	0.9888	0.9777	0.9594	0.9317	0.8933	0.8432
25	0.9999	0.9997	0.999	0.9974	0.9938	0.9869	0.9748	0.9554	0.9269	0.8878
26	1	0.9999	0.9995	0.9987	0.9967	0.9925	0.9848	0.9718	0.9514	0.9221
27	1	0.9999	0.9998	0.9994	0.9983	0.9959	0.9912	0.9827	0.9687	0.9475
28	1	1	0.9999	0.9997	0.9991	0.9978	0.995	0.9897	0.9805	0.9657
29	1	1	1	0.9999	0.9996	0.9989	0.9973	0.9941	0.9882	0.9782
30	1	1	1	0.9999	0.9998	0.9994	0.9986	0.9967	0.993	0.9865
31	1	1	1	1	0.9999	0.9997	0.9993	0.9982	0.996	0.9919
32	1	1	1	1	1	0.9999	0.9996	0.999	0.9978	0.9953
33	1	1	1	1	1	0.9999	0.9998	0.9995	0.9988	0.9973
34	1	1	1	1	1	1	0.9999	0.9998	0.9994	0.9985
35	1	1	1	1	1	1	1	0.9999	0.9997	0.9992
36	1	1	1	1	1	1	1	0.9999	0.9998	0.9996
37	1	1	1	1	1	1	1	1	0.9999	0.9998
38	1	1	1	1	1	1	1	1	1	0.9999
39	1	1	1	1	1	1	1	1	1	0.9999
40	1	1	1	1	1	1	1	1	1	1

Taula 10.4: Funció de distribució acumulada d'una variable Z que segueix la distribució normal tipificada. Utilitzar la relació $F(z) = 1 - F(-z)$ per z negatiu. Per a trobar z_q, el quantil d'ordre q, cal buscar q a l'interior de la taula i z_q serà el z associat (per exemple $z_{0.975} = 1.96$).

	Segon decimal de z									
z	*.*0	*.*1	*.*2	*.*3	*.*4	*.*5	*.*6	*.*7	*.*8	*.*9
0.0	0.5	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999
3.1	0.999	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1	1	1	1	1	1	1	1	1	1

Taula 10.5: Quantils de la distribució χ^2_n (chi-quadrat amb n graus de llibertat). Cada valor de la taula és $(\chi^2_n)_q$ per al corresponent ordre q del quantil i els graus de llibertat n de la distribució.

	Ordre q del quantil									
n	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.00004	0.00016	0.00098	0.00393	0.01579	2.705	3.841	5.023	6.634	7.879
2	0.01002	0.0201	0.05063	0.1025	0.2107	4.605	5.991	7.377	9.21	10.59
3	0.07172	0.1148	0.2157	0.3518	0.5843	6.251	7.814	9.348	11.34	12.83
4	0.2069	0.2971	0.4844	0.7107	1.063	7.779	9.487	11.14	13.27	14.86
5	0.4117	0.5542	0.8312	1.145	1.61	9.236	11.07	12.83	15.08	16.74
6	0.6757	0.872	1.237	1.635	2.204	10.64	12.59	14.44	16.81	18.54
7	0.9892	1.239	1.689	2.167	2.833	12.01	14.06	16.01	18.47	20.27
8	1.344	1.646	2.179	2.732	3.489	13.36	15.5	17.53	20.09	21.95
9	1.734	2.087	2.7	3.325	4.168	14.68	16.91	19.02	21.66	23.58
10	2.155	2.558	3.246	3.94	4.865	15.98	18.3	20.48	23.2	25.18
11	2.603	3.053	3.815	4.574	5.577	17.27	19.67	21.92	24.72	26.75
12	3.073	3.57	4.403	5.226	6.303	18.54	21.02	23.33	26.21	28.29
13	3.565	4.106	5.008	5.891	7.041	19.81	22.36	24.73	27.68	29.81
14	4.074	4.66	5.628	6.57	7.789	21.06	23.68	26.11	29.14	31.31
15	4.6	5.229	6.262	7.26	8.546	22.3	24.99	27.48	30.57	32.8
16	5.142	5.812	6.907	7.961	9.312	23.54	26.29	28.84	31.99	34.26
17	5.697	6.407	7.564	8.671	10.08	24.76	27.58	30.19	33.4	35.71
18	6.264	7.014	8.23	9.39	10.86	25.98	28.86	31.52	34.8	37.15
19	6.843	7.632	8.906	10.11	11.65	27.2	30.14	32.85	36.19	38.58
20	7.433	8.26	9.59	10.85	12.44	28.41	31.41	34.16	37.56	39.99
21	8.033	8.897	10.28	11.59	13.23	29.61	32.67	35.47	38.93	41.4
22	8.642	9.542	10.98	12.33	14.04	30.81	33.92	36.78	40.28	42.79
23	9.26	10.19	11.68	13.09	14.84	32	35.17	38.07	41.63	44.18
24	9.886	10.85	12.4	13.84	15.65	33.19	36.41	39.36	42.97	45.55
25	10.51	11.52	13.11	14.61	16.47	34.38	37.65	40.64	44.31	46.92
26	11.16	12.19	13.84	15.37	17.29	35.56	38.88	41.92	45.64	48.28
27	11.8	12.87	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.3	16.92	18.93	37.91	41.33	44.46	48.27	50.99
29	13.12	14.25	16.04	17.7	19.76	39.08	42.55	45.72	49.58	52.33
30	13.78	14.95	16.79	18.49	20.59	40.25	43.77	46.97	50.89	53.67
39	19.99	21.42	23.65	25.69	28.19	50.65	54.57	58.12	62.42	65.47
49	27.24	28.94	31.55	33.93	36.81	62.03	66.33	70.22	74.91	78.23
59	34.77	36.69	39.66	42.33	45.57	73.27	77.93	82.11	87.16	90.71
69	42.49	44.63	47.92	50.87	54.43	84.41	89.39	93.85	99.22	102.9
79	50.37	52.72	56.3	59.52	63.37	95.47	100.7	105.4	111.1	115.1
89	58.38	60.92	64.79	68.24	72.38	106.4	112	116.9	122.9	127.1
99	66.51	69.22	73.36	77.04	81.44	117.4	123.2	128.4	134.6	138.9
109	74.72	77.61	81.99	85.9	90.55	128.2	134.3	139.7	146.2	150.7
119	83.01	86.07	90.69	94.81	99.7	139.1	145.4	151	157.7	162.4
129	91.38	94.59	99.45	103.7	108.8	149.9	156.5	162.3	169.2	174.1

Taula 10.6: Quantils de la distribució t_n (t-Student amb n graus de llibertat). Cada valor de la taula és $(t_n)_q$ per al corresponent ordre q del quantil i els graus de llibertat n de la distribució.

	Ordre q del quantil									
n	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	-63.66	-31.83	-12.71	-6.314	-3.078	3.077	6.313	12.7	31.82	63.65
2	-9.925	-6.965	-4.303	-2.92	-1.886	1.885	2.919	4.302	6.964	9.924
3	-5.841	-4.541	-3.183	-2.354	-1.638	1.637	2.353	3.182	4.54	5.84
4	-4.605	-3.747	-2.777	-2.132	-1.534	1.533	2.131	2.776	3.746	4.604
5	-4.033	-3.365	-2.571	-2.016	-1.476	1.475	2.015	2.57	3.364	4.032
6	-3.708	-3.143	-2.447	-1.944	-1.44	1.439	1.943	2.446	3.142	3.707
7	-3.5	-2.998	-2.365	-1.895	-1.415	1.414	1.894	2.364	2.997	3.499
8	-3.356	-2.897	-2.307	-1.86	-1.397	1.396	1.859	2.306	2.896	3.355
9	-3.25	-2.822	-2.263	-1.834	-1.384	1.383	1.833	2.262	2.821	3.249
10	-3.17	-2.764	-2.229	-1.813	-1.373	1.372	1.812	2.228	2.763	3.169
11	-3.106	-2.719	-2.201	-1.796	-1.364	1.363	1.795	2.2	2.718	3.105
12	-3.055	-2.681	-2.179	-1.783	-1.357	1.356	1.782	2.178	2.68	3.054
13	-3.013	-2.651	-2.161	-1.771	-1.351	1.35	1.77	2.16	2.65	3.012
14	-2.977	-2.625	-2.145	-1.762	-1.346	1.345	1.761	2.144	2.624	2.976
15	-2.947	-2.603	-2.132	-1.754	-1.341	1.34	1.753	2.131	2.602	2.946
16	-2.921	-2.584	-2.12	-1.746	-1.337	1.336	1.745	2.119	2.583	2.92
17	-2.899	-2.567	-2.11	-1.74	-1.334	1.333	1.739	2.109	2.566	2.898
18	-2.879	-2.553	-2.101	-1.735	-1.331	1.33	1.734	2.1	2.552	2.878
19	-2.861	-2.54	-2.094	-1.73	-1.328	1.327	1.729	2.093	2.539	2.86
20	-2.846	-2.528	-2.086	-1.725	-1.326	1.325	1.724	2.085	2.527	2.845
21	-2.832	-2.518	-2.08	-1.721	-1.324	1.323	1.72	2.079	2.517	2.831
22	-2.819	-2.509	-2.074	-1.718	-1.322	1.321	1.717	2.073	2.508	2.818
23	-2.808	-2.5	-2.069	-1.714	-1.32	1.319	1.713	2.068	2.499	2.807
24	-2.797	-2.493	-2.064	-1.711	-1.318	1.317	1.71	2.063	2.492	2.796
25	-2.788	-2.486	-2.06	-1.709	-1.317	1.316	1.708	2.059	2.485	2.787
26	-2.779	-2.479	-2.056	-1.706	-1.315	1.314	1.705	2.055	2.478	2.778
27	-2.771	-2.473	-2.052	-1.704	-1.314	1.313	1.703	2.051	2.472	2.77
28	-2.764	-2.468	-2.049	-1.702	-1.313	1.312	1.701	2.048	2.467	2.763
29	-2.757	-2.463	-2.046	-1.7	-1.312	1.311	1.699	2.045	2.462	2.756
30	-2.75	-2.458	-2.043	-1.698	-1.311	1.31	1.697	2.042	2.457	2.749
39	-2.708	-2.426	-2.023	-1.685	-1.304	1.303	1.684	2.022	2.425	2.707
49	-2.68	-2.405	-2.01	-1.677	-1.3	1.299	1.676	2.009	2.404	2.679
59	-2.662	-2.392	-2.001	-1.672	-1.297	1.296	1.671	2	2.391	2.661
69	-2.649	-2.382	-1.995	-1.668	-1.294	1.293	1.667	1.994	2.381	2.648
79	-2.64	-2.375	-1.991	-1.665	-1.293	1.292	1.664	1.99	2.374	2.639
89	-2.633	-2.369	-1.987	-1.663	-1.292	1.291	1.662	1.986	2.368	2.632
99	-2.627	-2.365	-1.985	-1.661	-1.291	1.29	1.66	1.984	2.364	2.626
109	-2.622	-2.362	-1.982	-1.659	-1.29	1.289	1.658	1.981	2.361	2.621
119	-2.618	-2.359	-1.981	-1.658	-1.289	1.288	1.657	1.98	2.358	2.617
129	-2.615	-2.356	-1.979	-1.657	-1.289	1.288	1.656	1.978	2.355	2.614

Taula 10.7: Quantils 0.90 de la distribució F_{n_1,n_2} (F-Snedecor amb n_1 i n_2 graus de llibertat). Cada valor de la taula és el quantil $(F_{n_1,n_2})_{0.90}$ (no confondre amb el punt crític d'altres llibres) corresponent als graus de llibertat n_1 i n_2 alineats.

Quantil $q = 0.90$, és a dir $F_{n_1,n_2,(0.90)}$ (punt crític 0.10)											
n_1 / n_2	4	9	14	19	24	29	34	39	44	49	99
4	4.107	3.935	3.877	3.848	3.83	3.819	3.81	3.804	3.799	3.795	3.778
9	2.692	2.44	2.351	2.304	2.276	2.257	2.244	2.233	2.225	2.219	2.189
14	2.394	2.121	2.022	1.97	1.937	1.915	1.899	1.887	1.877	1.869	1.834
19	2.266	1.983	1.878	1.822	1.787	1.763	1.745	1.732	1.721	1.712	1.673
24	2.194	1.906	1.797	1.738	1.701	1.676	1.657	1.643	1.631	1.622	1.579
29	2.149	1.856	1.745	1.684	1.646	1.619	1.6	1.585	1.573	1.563	1.517
34	2.117	1.822	1.709	1.647	1.607	1.58	1.559	1.544	1.531	1.521	1.473
39	2.094	1.797	1.682	1.619	1.578	1.55	1.529	1.513	1.5	1.49	1.439
44	2.077	1.777	1.661	1.597	1.556	1.528	1.506	1.489	1.476	1.465	1.413
49	2.063	1.762	1.645	1.58	1.539	1.509	1.488	1.471	1.457	1.446	1.392
54	2.051	1.75	1.632	1.566	1.524	1.495	1.472	1.455	1.441	1.43	1.375
59	2.042	1.739	1.621	1.555	1.512	1.482	1.46	1.442	1.428	1.417	1.36
64	2.034	1.731	1.612	1.545	1.502	1.472	1.449	1.431	1.417	1.405	1.348
69	2.028	1.723	1.604	1.537	1.494	1.463	1.44	1.422	1.408	1.396	1.337
74	2.022	1.717	1.597	1.53	1.486	1.455	1.432	1.414	1.399	1.387	1.328
79	2.017	1.712	1.591	1.524	1.48	1.449	1.425	1.407	1.392	1.38	1.32
84	2.013	1.707	1.586	1.518	1.474	1.443	1.419	1.4	1.386	1.373	1.312
89	2.009	1.702	1.581	1.513	1.469	1.437	1.414	1.395	1.38	1.367	1.306
94	2.005	1.699	1.577	1.509	1.464	1.433	1.409	1.39	1.375	1.362	1.3
99	2.002	1.695	1.573	1.505	1.46	1.428	1.404	1.385	1.37	1.357	1.295
104	1.999	1.692	1.57	1.501	1.457	1.424	1.4	1.381	1.366	1.353	1.29
109	1.997	1.689	1.567	1.498	1.453	1.421	1.397	1.378	1.362	1.349	1.285
114	1.994	1.687	1.564	1.495	1.45	1.418	1.393	1.374	1.359	1.346	1.281
119	1.992	1.684	1.562	1.493	1.447	1.415	1.39	1.371	1.355	1.343	1.278
124	1.99	1.682	1.559	1.49	1.445	1.412	1.388	1.368	1.353	1.339	1.274
129	1.988	1.68	1.557	1.488	1.442	1.41	1.385	1.366	1.35	1.337	1.271
134	1.987	1.678	1.555	1.486	1.44	1.407	1.383	1.363	1.347	1.334	1.268
139	1.985	1.676	1.553	1.484	1.438	1.405	1.38	1.361	1.345	1.332	1.265
144	1.984	1.675	1.552	1.482	1.436	1.403	1.378	1.359	1.343	1.33	1.263
149	1.982	1.673	1.55	1.48	1.434	1.401	1.376	1.357	1.341	1.327	1.26

Taula 10.8: Quantils 0.95 de la distribució F_{n_1,n_2} (F-Snedecor amb n_1 i n_2 graus de llibertat). Cada valor de la taula és el quantil $(F_{n_1,n_2})_{0.95}$ (no confondre amb el punt crític d'altres llibres) corresponent als graus de llibertat n_1 i n_2 alineats.

Quantil $q = 0.95$, és a dir $F_{n_1,n_2,(0.95)}$ (punt crític 0.05)											
n_1 / n_2	4	9	14	19	24	29	34	39	44	49	99
4	6.388	5.998	5.873	5.811	5.774	5.749	5.732	5.719	5.709	5.7	5.664
9	3.633	3.178	3.025	2.947	2.9	2.868	2.846	2.828	2.815	2.804	2.756
14	3.112	2.645	2.483	2.4	2.348	2.313	2.288	2.269	2.254	2.242	2.187
19	2.895	2.422	2.255	2.168	2.114	2.077	2.05	2.029	2.013	2	1.94
24	2.776	2.3	2.129	2.039	1.983	1.945	1.917	1.895	1.878	1.864	1.801
29	2.701	2.222	2.05	1.958	1.9	1.86	1.831	1.809	1.791	1.777	1.71
34	2.649	2.169	1.994	1.901	1.842	1.802	1.772	1.749	1.73	1.715	1.646
39	2.612	2.13	1.954	1.859	1.8	1.758	1.728	1.704	1.685	1.67	1.598
44	2.583	2.1	1.923	1.828	1.767	1.725	1.694	1.67	1.65	1.635	1.56
49	2.561	2.077	1.899	1.802	1.741	1.698	1.667	1.642	1.623	1.607	1.53
54	2.542	2.058	1.879	1.782	1.72	1.677	1.645	1.62	1.6	1.584	1.506
59	2.527	2.042	1.863	1.765	1.703	1.659	1.627	1.601	1.581	1.565	1.485
64	2.515	2.029	1.849	1.751	1.688	1.644	1.611	1.586	1.566	1.549	1.468
69	2.504	2.018	1.837	1.739	1.676	1.631	1.598	1.573	1.552	1.535	1.453
74	2.495	2.009	1.827	1.728	1.665	1.62	1.587	1.561	1.54	1.523	1.44
79	2.487	2	1.818	1.719	1.655	1.61	1.577	1.551	1.53	1.513	1.428
84	2.48	1.993	1.811	1.711	1.647	1.602	1.568	1.542	1.521	1.503	1.418
89	2.474	1.986	1.804	1.704	1.64	1.594	1.56	1.534	1.513	1.495	1.409
94	2.468	1.981	1.798	1.698	1.633	1.588	1.554	1.527	1.505	1.488	1.401
99	2.463	1.975	1.792	1.692	1.627	1.582	1.547	1.521	1.499	1.481	1.394
104	2.459	1.971	1.788	1.687	1.622	1.576	1.542	1.515	1.493	1.475	1.387
109	2.454	1.966	1.783	1.682	1.617	1.571	1.537	1.509	1.488	1.47	1.381
114	2.451	1.962	1.779	1.678	1.613	1.567	1.532	1.505	1.483	1.465	1.375
119	2.447	1.959	1.775	1.674	1.609	1.562	1.528	1.5	1.478	1.46	1.37
124	2.444	1.956	1.772	1.671	1.605	1.559	1.524	1.496	1.474	1.456	1.365
129	2.441	1.953	1.769	1.667	1.602	1.555	1.52	1.493	1.47	1.452	1.361
134	2.439	1.95	1.766	1.664	1.598	1.552	1.517	1.489	1.467	1.448	1.357
139	2.436	1.947	1.763	1.661	1.595	1.549	1.513	1.486	1.464	1.445	1.353
144	2.434	1.945	1.761	1.659	1.593	1.546	1.511	1.483	1.461	1.442	1.349
149	2.432	1.943	1.758	1.656	1.59	1.543	1.508	1.48	1.458	1.439	1.346

Taula 10.9: Quantils 0.975 de la distribució F_{n_1,n_2} (F-Snedecor amb n_1 i n_2 graus de llibertat). Cada valor de la taula és el quantil $(F_{n_1,n_2})_{0.975}$ (no confondre amb el punt crític d'altres llibres) corresponent als graus de llibertat n_1 i n_2 alineats.

Quantil $q = 0.975$, és a dir $F_{n_1,n_2,(0.975)}$ (punt crític 0.025)											
n_1 / n_2	4	9	14	19	24	29	34	39	44	49	99
4	9.604	8.904	8.683	8.575	8.51	8.468	8.437	8.415	8.397	8.383	8.32
9	4.718	4.025	3.797	3.683	3.614	3.567	3.534	3.509	3.49	3.474	3.404
14	3.891	3.209	2.978	2.86	2.788	2.74	2.705	2.678	2.658	2.641	2.565
19	3.558	2.88	2.646	2.526	2.452	2.401	2.365	2.337	2.315	2.298	2.217
24	3.379	2.702	2.467	2.345	2.269	2.217	2.179	2.15	2.128	2.109	2.025
29	3.267	2.591	2.355	2.231	2.154	2.1	2.062	2.032	2.009	1.99	1.902
34	3.19	2.516	2.278	2.153	2.074	2.02	1.981	1.95	1.926	1.907	1.816
39	3.135	2.461	2.222	2.095	2.016	1.961	1.921	1.89	1.866	1.846	1.752
44	3.093	2.419	2.179	2.052	1.972	1.916	1.876	1.844	1.819	1.799	1.703
49	3.06	2.386	2.146	2.018	1.937	1.881	1.84	1.808	1.782	1.762	1.663
54	3.033	2.36	2.119	1.99	1.909	1.852	1.81	1.778	1.752	1.731	1.631
59	3.011	2.338	2.096	1.967	1.885	1.828	1.786	1.754	1.728	1.706	1.604
64	2.993	2.32	2.078	1.948	1.866	1.808	1.766	1.733	1.707	1.685	1.582
69	2.977	2.304	2.062	1.932	1.849	1.791	1.749	1.715	1.689	1.667	1.562
74	2.964	2.291	2.048	1.918	1.835	1.777	1.734	1.7	1.673	1.651	1.545
79	2.952	2.279	2.036	1.905	1.822	1.764	1.721	1.687	1.66	1.638	1.531
84	2.942	2.269	2.026	1.895	1.811	1.753	1.709	1.675	1.648	1.626	1.517
89	2.933	2.26	2.017	1.885	1.801	1.743	1.699	1.665	1.637	1.615	1.506
94	2.925	2.252	2.008	1.877	1.793	1.734	1.69	1.656	1.628	1.605	1.495
99	2.917	2.245	2.001	1.869	1.785	1.726	1.682	1.647	1.62	1.597	1.486
104	2.911	2.238	1.994	1.862	1.778	1.718	1.674	1.64	1.612	1.589	1.477
109	2.905	2.232	1.988	1.856	1.771	1.712	1.667	1.633	1.605	1.582	1.469
114	2.9	2.227	1.983	1.85	1.766	1.706	1.661	1.627	1.599	1.576	1.462
119	2.895	2.222	1.978	1.845	1.76	1.7	1.656	1.621	1.593	1.57	1.456
124	2.89	2.218	1.973	1.84	1.755	1.695	1.651	1.616	1.587	1.564	1.449
129	2.886	2.214	1.969	1.836	1.751	1.691	1.646	1.611	1.582	1.559	1.444
134	2.882	2.21	1.965	1.832	1.747	1.686	1.641	1.606	1.578	1.554	1.439
139	2.879	2.206	1.961	1.828	1.743	1.682	1.637	1.602	1.574	1.55	1.434
144	2.875	2.203	1.958	1.825	1.739	1.679	1.634	1.598	1.57	1.546	1.429
149	2.872	2.2	1.955	1.821	1.736	1.675	1.63	1.595	1.566	1.542	1.425

Taula 10.10: Quantils 0.99 de la distribució F_{n_1,n_2} (F-Snedecor amb n_1 i n_2 graus de llibertat). Cada valor de la taula és el quantil $(F_{n_1,n_2})_{0.99}$ (no confondre amb el punt crític d'altres llibres) corresponent als graus de llibertat n_1 i n_2 alineats.

Quantil $q = 0.99$, és a dir $F_{n_1,n_2,(0.99)}$ (punt crític 0.01)											
n_1 / n_2	4	9	14	19	24	29	34	39	44	49	99
4	15.97	14.65	14.24	14.04	13.92	13.85	13.79	13.75	13.72	13.69	13.57
9	6.422	5.351	5.005	4.832	4.728	4.659	4.61	4.573	4.544	4.52	4.416
14	5.035	4.029	3.697	3.529	3.427	3.358	3.309	3.272	3.242	3.219	3.112
19	4.5	3.522	3.194	3.027	2.924	2.855	2.805	2.767	2.737	2.713	2.603
24	4.218	3.255	2.93	2.762	2.659	2.588	2.537	2.498	2.468	2.443	2.33
29	4.044	3.092	2.767	2.598	2.494	2.423	2.371	2.332	2.301	2.275	2.158
34	3.927	2.981	2.656	2.487	2.382	2.31	2.258	2.218	2.186	2.16	2.04
39	3.842	2.9	2.576	2.407	2.301	2.229	2.175	2.135	2.103	2.076	1.954
44	3.778	2.84	2.516	2.346	2.24	2.166	2.113	2.072	2.039	2.012	1.887
49	3.728	2.793	2.469	2.298	2.191	2.118	2.064	2.022	1.989	1.962	1.835
54	3.688	2.755	2.431	2.26	2.153	2.078	2.024	1.982	1.949	1.921	1.792
59	3.654	2.724	2.399	2.228	2.121	2.046	1.991	1.949	1.915	1.888	1.757
64	3.627	2.697	2.373	2.202	2.094	2.019	1.964	1.921	1.887	1.859	1.727
69	3.603	2.675	2.351	2.179	2.071	1.996	1.94	1.897	1.863	1.835	1.701
74	3.583	2.656	2.332	2.16	2.052	1.976	1.92	1.877	1.843	1.814	1.679
79	3.566	2.64	2.316	2.143	2.034	1.959	1.903	1.859	1.825	1.796	1.659
84	3.551	2.626	2.301	2.129	2.02	1.944	1.887	1.844	1.809	1.78	1.642
89	3.537	2.613	2.288	2.116	2.006	1.93	1.874	1.83	1.795	1.766	1.627
94	3.525	2.601	2.277	2.104	1.995	1.918	1.861	1.818	1.782	1.753	1.613
99	3.514	2.591	2.267	2.094	1.984	1.907	1.851	1.806	1.771	1.742	1.601
104	3.505	2.582	2.258	2.084	1.975	1.898	1.841	1.796	1.761	1.732	1.59
109	3.496	2.574	2.249	2.076	1.966	1.889	1.832	1.787	1.752	1.723	1.58
114	3.488	2.566	2.242	2.068	1.958	1.881	1.824	1.779	1.743	1.714	1.57
119	3.48	2.559	2.235	2.061	1.951	1.874	1.816	1.772	1.736	1.706	1.562
124	3.474	2.553	2.228	2.055	1.944	1.867	1.809	1.765	1.729	1.699	1.554
129	3.468	2.547	2.223	2.049	1.938	1.861	1.803	1.758	1.722	1.693	1.547
134	3.462	2.542	2.217	2.043	1.933	1.855	1.797	1.752	1.716	1.686	1.54
139	3.457	2.537	2.212	2.038	1.927	1.85	1.792	1.747	1.711	1.681	1.534
144	3.452	2.532	2.208	2.033	1.923	1.845	1.787	1.742	1.705	1.676	1.528
149	3.447	2.528	2.203	2.029	1.918	1.84	1.782	1.737	1.701	1.671	1.523

Taula 10.11: Quantils 0.995 de la distribució F_{n_1,n_2} (F-Snedecor amb n_1 i n_2 graus de llibertat). Cada valor de la taula és el quantil $(F_{n_1,n_2})_{0.995}$ (no confondre amb el punt crític d'altres llibres) corresponent als graus de llibertat n_1 i n_2 alineats.

Quantil $q = 0.995$, és a dir $F_{n_1,n_2,(0.995)}$ (punt crític 0.005)											
n_1 / n_2	4	9	14	19	24	29	34	39	44	49	99
4	23.15	21.13	20.51	20.21	20.03	19.91	19.82	19.76	19.71	19.67	19.49
9	7.955	6.541	6.088	5.863	5.729	5.639	5.575	5.526	5.489	5.459	5.323
14	5.998	4.717	4.299	4.088	3.961	3.875	3.814	3.767	3.731	3.702	3.57
19	5.268	4.042	3.637	3.431	3.306	3.221	3.16	3.113	3.077	3.048	2.914
24	4.889	3.694	3.296	3.091	2.966	2.881	2.82	2.773	2.736	2.707	2.57
29	4.659	3.483	3.088	2.884	2.759	2.673	2.611	2.564	2.527	2.497	2.358
34	4.503	3.341	2.948	2.745	2.619	2.533	2.471	2.423	2.385	2.355	2.213
39	4.392	3.239	2.847	2.644	2.518	2.432	2.369	2.321	2.283	2.252	2.107
44	4.308	3.162	2.772	2.569	2.442	2.356	2.292	2.244	2.205	2.174	2.027
49	4.243	3.102	2.713	2.51	2.383	2.296	2.232	2.183	2.144	2.113	1.963
54	4.19	3.054	2.666	2.462	2.335	2.248	2.184	2.134	2.095	2.063	1.912
59	4.147	3.015	2.627	2.424	2.296	2.208	2.144	2.094	2.055	2.023	1.87
64	4.111	2.982	2.595	2.391	2.264	2.175	2.111	2.061	2.021	1.988	1.834
69	4.081	2.954	2.567	2.364	2.236	2.147	2.082	2.032	1.992	1.959	1.803
74	4.055	2.931	2.544	2.34	2.212	2.123	2.058	2.008	1.967	1.934	1.777
79	4.032	2.91	2.523	2.32	2.191	2.103	2.037	1.986	1.946	1.913	1.754
84	4.012	2.892	2.506	2.302	2.173	2.084	2.018	1.967	1.927	1.894	1.734
89	3.995	2.876	2.49	2.286	2.157	2.068	2.002	1.951	1.91	1.877	1.716
94	3.979	2.862	2.476	2.272	2.143	2.054	1.987	1.936	1.895	1.862	1.699
99	3.965	2.849	2.463	2.259	2.13	2.041	1.974	1.923	1.882	1.848	1.685
104	3.953	2.838	2.452	2.248	2.119	2.029	1.962	1.911	1.87	1.836	1.672
109	3.942	2.827	2.442	2.237	2.108	2.018	1.952	1.9	1.859	1.825	1.66
114	3.931	2.818	2.432	2.228	2.099	2.009	1.942	1.89	1.849	1.815	1.649
119	3.922	2.809	2.424	2.219	2.09	2	1.933	1.881	1.84	1.805	1.639
124	3.913	2.802	2.416	2.212	2.082	1.992	1.925	1.873	1.831	1.797	1.63
129	3.905	2.794	2.409	2.204	2.075	1.985	1.917	1.865	1.824	1.789	1.621
134	3.898	2.788	2.402	2.198	2.068	1.978	1.91	1.858	1.816	1.782	1.613
139	3.891	2.781	2.396	2.192	2.062	1.971	1.904	1.852	1.81	1.775	1.606
144	3.885	2.776	2.391	2.186	2.056	1.965	1.898	1.846	1.804	1.769	1.599
149	3.879	2.77	2.385	2.18	2.051	1.96	1.892	1.84	1.798	1.763	1.593

Bibliografia

[1] Anscombe, F. J. (1973): *Graphs in Statistical Analysis*, The American Statistician 27(1), pp. 17–21

[2] Banfield, J., `http://rweb.stat.umn.edu/Rweb/`

[3] Bardina, X. i M. Farré (2009): *Estadística descriptiva*, Manuals de la UAB, 54, Servei de Publicacions de la Universitat Autònoma de Barcelona.

[4] Bardina, X. i M. Farré (2005): *Estadística: un curs introductori per a estudiants de ciències socials i humanes. Vols. 1 i 2*, Manuals de la UAB, 162 i 166, Servei de Publicacions de la Universitat Autònoma de Barcelona.

[5] Becker, R.A., J.M. Chambers i A.R. Wilks (1988): *The new S language, a programming environment for data analysis and graphics*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, Californa.

[6] Carleton College, Mathematics and Computer Sciences, `http://www.mathcs.carleton.edu/probweb/teaching.html`

[7] DeGroot, M. H. (1988): *Probabilidad y estadística*, Addison Wesley Iberoamericana, México.

[8] Devore, J. i N. Farnum (1999): *Applied Statistics for Engineers and Scientists*, Thomson Learning.

[9] Domingo, J. (1997): *Estadística tècnica, una introducció constructivista*, Universitat Rovira i Virgili, Servei Lingüístic.

[10] Dougherty, E. R. (1990): *Probability and statistics for the engineering, computing and physical sciences*, Prentince Hall International Editions.

[11] Epifanio, I. i P. Gregori (2009): *Ampliación de estadística para la Ingeniería Técnica en Informática de Gestión*, Col·lecció Sapientia, 13, Publicacions de la Universitat Jaume I.

[12] Gelman, A. i D. Nolan (2002): *Teaching Statistics, a bag of tricks*, Oxford University Press, New York.

[13] Gregori, P., `http://www3.uji.es/~gregori/materialsuji.zip`

[14] Holmes, S., http://www-stat.stanford.edu/%7Esusan/surprise/

[15] Johnson, R. A. i G. K. Bhattacharyya (1996): *Statistics: principles and methods*, John Wiley and Sons, New York.

[16] Khazanie, R. (1996): *Statistics in a world of applications*, Harper Collins College Publishers, New York.

[17] Lane, D.M., http://www.mathcs.carleton.edu/probweb/teaching.html

[18] Montes, P., http://www.uv.es/~montes/curs.htm

[19] Montgomery, D. C. i G. C. Runger (1994): *Applied statistics and probability for engineers*, John Wiley and Sons, New York.

[20] Peña, D. (1995): *Estadística Modelos y métodos, vol 1 Fundamentos*, Alianza, Madrid, 1995

[21] R Core Development Team, The, http://www.r-project.org