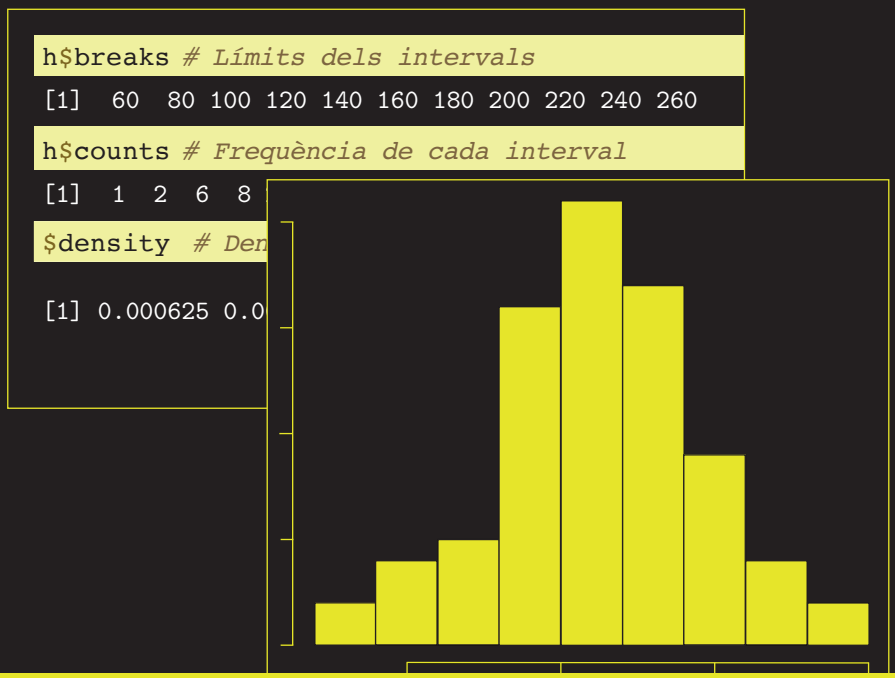
Pràctiques d'estadística amb R

Aplicacions en problemes d'enginyeria



Luis Eduardo Mújica Delgado
Magda L. Ruiz Ordoñez

UPCGRAU 77

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Pràctiques d'estadística amb R

Aplicacions en problemes d'enginyeria

Luis Eduardo Mújica Delgado
Magda L. Ruiz Ordoñez

UPCGRAU 77

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Amb el suport de

Generalitat de Catalunya
Departament de Recerca i Universitats

Traducció del llibre: *Prácticas de estadítica con R. Aplicaciones en problemas de ingeniería*
Traductor: Pere Farrando Canals

Primera edició: maig de 2024

Oficina de Publicacions Acadèmiques Digitals de la UPC
Edificio K2M, Planta S1, Despacho S103-S104
Jordi Girona 1-3, 08034 Barcelona
Tel.: 934 015 885
www.upc.edu/idp
E-mail: info.idp@upc.edu

Producción: Service Point
Pau Casals, 161-163
08820 El Prat de Llobregat (Barcelona)

ISBN:978-84-10008-53-3
ISBN digital: 978-84-10008-54-0
DL: B 10692-2024
DOI: 10.5821/ebook-9788410008540

Índice

Índice

Introducció

No cal insistir en la importància que té l'estadística en les diferents àrees professionals (enginyeria, salut, educació, comerç, etc.) en què es pot recollir i agrupar informació per construir informes que permetin inferir des del punt de vista quantitatiu i qualitatiu les poblacions, els processos, els sistemes, les fallades, els danys i els estats de les condicions normals i/o anormals, entre altres qüestions, i formar-se'n una idea. A més a més, pel fet de ser l'estadística una part de la ciència que comprèn teoria i pràctica, es fa imprescindible aplicar en el seu ensenyament eines tecnològiques com un element indispensable en la formació dels nostres futurs professionals.

De la mateixa manera, som conscients que la tecnologia està en un desenvolupament continu, que no s'aturarà, perquè la nostra curiositat no té límits i, per tant, continuarem investigant a la recerca de coneixement. Per això, és imprescindible que coneguem i emprem les eines capdavanteres que existeixen i faciliten els càlculs. En aquest aspecte, hi ha diferents paquets de programari estadístic, com SPSS, Minitab i S-Plus, entre d'altres. Aquest programari està molt ben desenvolupat i és molt bo a l'hora de satisfer les necessitats de qualsevol usuari o usuària. No obstant això, per utilitzar-lo cal pagar una llicència. En canvi, ha sorgit amb força una eina *open source* potent anomenada *R*. En general és senzill d'utilitzar i permet modes de treball diferents, amb intenció o bé que l'usuari o usuària entengui i sàpiga generar línies d'instruccions o que generi línies de codi per mitjà de menús. D'aquesta manera, no cal ser un expert de l'estadística ni un programador consumat per entendre els procediments necessaris per obtenir una solució, sinó que, alhora, l'usuari o usuària pot concentrar-se en l'anàlisi dels resultats.

Pel fet de ser codi obert, *R* s'actualitza de manera permanent i la comunitat estadística en va millorant l'abast i reconfigurant els errors. Com a avantatges addicionals de *R* es pot enumerar els següents:

1. Disponibilitat per a Linux, Windows i Mac.
2. Solvència, robustesa i estabilitat.
3. Qualitat i facilitat de creació de gràfics.
4. Hi ha l'opció de treballar-hi en xarxa sense necessitat d'instal·lar-lo, cosa que hi afavoreix l'accés i en facilita l'ús.

És així com neix la principal motivació a l'hora de dissenyar aquestes guies: ajudar els nostres estudiants o aprenents a comprendre i entendre conceptes estadístics al mateix temps que s'apliquen i es resolen utilitzant *R*. Com a dada interessant i sense que ens ho proposéssim en dissenyar les guies, topem amb tres paradigmes propis de l'ensenyament de temes científics: El *paradigma de l'ensenyament per transmissió* estableix els treballs pràctics com a activitats de descobriment de fets i conceptes, en aquest cas estadístics, mitjançant la utilització de *R*. Això ens condueix irremeiablement al paradigma següent, *el descobriment guiat i el descobriment autònom*, que es basa en el concepte que els treballs pràctics són activitats encaminades a aprendre per mitjà de l'observació, la classificació, l'elaboració d'hipòtesis, la realització, etc. Finalment, tenim el paradigma de *la ciència dels processos*, en què els treballs pràctics s'utilitzen per fomentar l'adquisició d'habilitats i per posar l'estudiant en situació de resoldre problemes pràctics. Com a resultat, l'objectiu en dissenyar aquestes guies és encaminar el treball de l'aprenent de *R* proporcionant la informació bàsica i necessària per a la comprensió del tema. D'aquesta manera observaran com els seus coneixements i capacitats de resolució van evolucionant mentre resolen els reptes proposats (els exercicis que presentem) en cadascuna de les guies.

Hem dissenyat les guies de manera atractiva, començant amb una pregunta que inicialment no es podria resoldre. Al final, però, obtindrem totes les respostes. Els temes es van presentant de manera seqüencial, començant per l'ús bàsic, com es faria per a qualsevol altre programa, passant per la simulació d'experiments aleatoris, fins a arribar finalment a facilitar l'anàlisi en temes molt més especialitzats, com és el contrast d'hipòtesi.

En cap moment no esperem que aquestes guies siguin una substitució del professor o professora. És primordial que, en l'aprenentatge de temes científics, totes les situacions siguin guiades per un expert o experta, per afavorir el desenvolupament dels aprenents i la seva familiarització amb la tecnologia. El nostre desig és que aquestes guies afavoreixin l'aprenentatge actiu promovent el contacte entre els uns i els altres amb la realimentació del coneixement com a etapa essencial en l'assimilació dels conceptes. Com a resultat, al final, els nostres estudiants o aprenents trobaran sentit als conceptes explicats basant-se en l'experiència, l'aplicació, l'anàlisi i finalment la inferència dels reptes.

1 Introducció a R

1.1. Introducció i objectius

En aquesta sessió es fa una introducció a l'eina informàtica per a l'anàlisi estadística *R*. És un llenguatge de programació en codi obert (*free software*) i gratuït (*freeware*) que últimament està suscitant l'interès de l'acadèmia, del món de la recerca i fins i tot de la indústria.

En primer lloc, s'explica breument què és *R* i els seus diversos entorns d'ús. A més a més, s'ofereixen les indicacions per descarregar i instal·lar el programa, els seus paquets i l'entorn de treball. A continuació, se'n detallen algunes funcions bàsiques, així com el procediment per crear i manipular taules de dades (*data.frame*). També es descriu com es desen i recuperen les ordres o instruccions (*commands*) executades en una sessió i l'espai de treball. Finalment, es mostra com es crea i desa un script o guió per poder-lo obrir i executar des d'un altre ordinador o en una altra sessió. En finalitzar aquesta sessió, l'estudiant ha de ser capaç de:

- Descarregar i instal·lar els fitxers de *R*, el paquet *R-Commander* i l'entorn de treball *Rstudio*.
- Identificar les principals característiques de la *consola de R*, el paquet *R-Commander* i l'entorn de treball *Rstudio*.
- Iniciar una sessió de treball en la *consola de R*, en *R-Commander* i en *Rstudio*.
- Crear i manipular una taula de dades.
- Desar i recuperar l'historial d'una sessió treballada i el seu espai de treball.
- Crear, desar i recuperar un script o guió amb una seqüència d'ordres.

1.2. R, R-Commander i Rstudio

1.2.1. Què són R, R-Commander i Rstudio?

Encara que, per a algunes persones, *R* sigui un programa, el podem considerar un llenguatge de programació enfocat a l'anàlisi estadística de dades i la seva representació

gràfica. Es pot executar en qualsevol ordinador i té un suport en línia molt amigable i actiu (https://www.r-project.org). Proporciona una gran quantitat d'eines amb la capacitat de cridar altres funcions i de desenvolupar noves funcions molt senzilles de manejar. A més a més, la seva gran capacitat de visualització de les dades permet generar gràfics molt variats i d'una extraordinària qualitat i flexibilitat. Permet la integració amb diferents bases de dades i amb altres llenguatges de programació, com Matlab, Maple, Mathematica, Python, Perl, SPSS, etc. Així mateix, com que és un projecte obert i col·laboratiu, hi ha un repositori oficial de paquets o biblioteques (*libraries*) (https://cran.r-project.org/web/packages/).

R permet treballar amb una finestra d'interacció amb l'usuari, *R-Console*. En el seu entorn bàsic, *R* no té una interfície de tipus finestra. Per obtenir els resultats desitjats, les seves funcions s'executen per mitjà d'ordres en el seu propi llenguatge. No obstant això, *R* disposa d'un mòdul addicional (o paquet) anomenat *R-Commander*, que proporciona una sèrie de menús que faciliten l'ús inicial del programa, sense haver d'escriure les ordres, és a dir, amb l'ús del ratolí.

R-Commander és una interfície gràfica d'usuari bàsica (*graphical user interface*, GUI). Els seus menús permeten executar moltes de les funcions bàsiques per a l'anàlisi estadística de dades (tot i que no totes) i crear gràfics sense escriure les ordres; encara més, genera el codi en llenguatge *R* perquè després es pugui executar des de *R-Console*, si així es vol. Tota la informació (els fitxers, l'ajuda i els manuals) es pot consultar al seu lloc web (http://www.rcommander.com/).

D'altra banda, hi ha un entorn per al desenvolupament integrat (*integrated development environment*, IDE) anomenat *Rstudio*, que és bàsicament una interfície agradable que inclou una consola, un editor més complet i funcional, una finestra de gràfics i la visualització de les variables en l'espai de treball, entre altres coses. Està completament integrat en *R* i *R-Commander* i permet executar el codi directament des de l'editor i gestionar múltiples directoris i fitxers.

1.2.2. Instal·lació

Com que *R* és gratuït, es poden trobar a internet molts llocs per descarregar els fitxers necessaris per instal·lar-lo. No obstant això, el lloc oficial de *R* té a disposició l'última versió per a Windows, Linux i Mac (macOS). CRAN (Comprehensive R Archive Network) és una xarxa de servidors web i FTP distribuïda per tot el món que emmagatzema les versions més actualitzades del codi i la documentació de *R* (https://cran.r-project.org/).

La instal·lació sol ser una miqueta complicada, depenent del sistema operatiu i naturalment de la seva versió. Com que *R* està constantment en desenvolupament, és difícil definir en aquest document els passos exactes i definitius per instal·lar-lo. Malgrat això, les instruccions senzilles que s'ofereixen a continuació són la base per a una

instal·lació correcta en Windows. Si es vol instal·lar en una altra plataforma o sorgeix algun inconvenient, el millor és consultar els fòrums i les preguntes freqüents de CRAN.

La versió més actualitzada de *R* es pot descarregar des de CRAN seguint la ruta: Download R for Windows > base > Download R 3.x.x for Windows. S'executa el fitxer deixant totes les opcions d'instal·lació per defecte. A la figura 1.1 es mostra la icona d'accés directe que apareix després de la instal·lació.

Rstudio es descarrega des de http://rstudio.org/download/desktop. S'executa el fitxer d'instal·lació seleccionant de nou totes les opcions de configuració per defecte. La icona d'accés directe es mostra a la figura 1.2.

Fig. 1.1: Icona d'accés a *R*.

Fig. 1.2: Icona d'accés a *Rstudio*.

La instal·lació de *R-Commander* es pot fer des de *R* o des de *Rstudio*. Si es vol fer des de *R*, s'executa el programa i en la consola s'executa l'ordre:

```
install.packages("Rcmdr",dependencies=TRUE)
```

Se selecciona el servidor desitjat i immediatament comença la descàrrega i la instal·lació de totes les biblioteques necessàries per executar-lo.

Ara, si es vol instal·lar *R-Commander* des de *Rstudio*, s'executa el programa i se selecciona la pestanya Packages. I es fa clic a Install Packages (figura 1.3).

A la nova finestra, es comença a escriure Rcmdr en l'espai de Packages assegurant-se que l'opció Install dependencies estigui seleccionada. Finalment, es fa clic a Install, tal com s'observa a la figura 1.4.

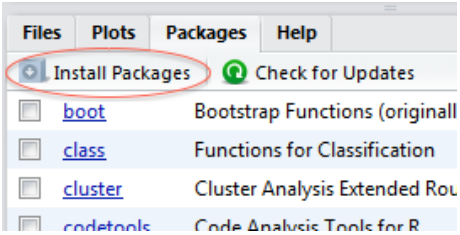


Fig. 1.3: Instal·lació d'un paquet (o biblioteca) des de *Rstudio*.

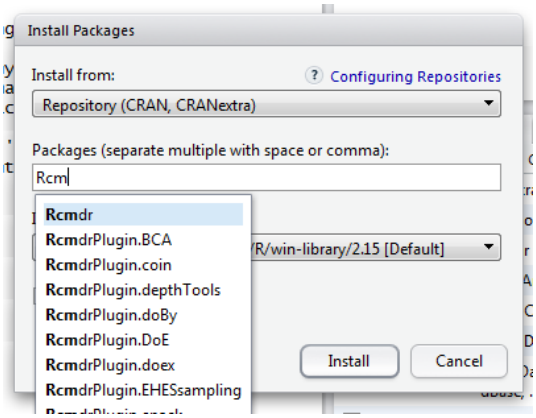


Fig. 1.4: Instal·lació de *R-Commander* des de *Rstudio*.

Com que *R-Commander* és un paquet de *R*, no es pot executar de manera independent i, per tant, no genera cap icona d'accés directe. Per executar *R-Commander* (des de *R* o *Rstudio*), cal fer-ho des de la consola mitjançant la instrucció següent:

```
library(Rcmdr)
```

1.2.3. Primeres impressions

R-Console

Una vegada iniciat el programa, es pot observar que s'obre una finestra de treball denominada *R-Console*, tal com mostra la figura 1.5.

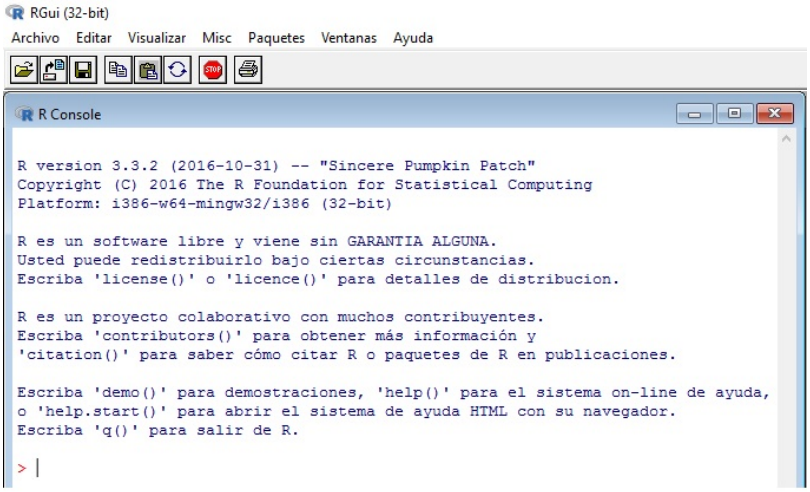


Fig. 1.5: Vista de la consola de *R*.

El cursor > | indica que el programa està preparat per acceptar ordres i efectuar els càlculs corresponents. Aquestes instruccions s'han de donar en forma d'ordres, operadors i funcions. Els més importants s'aniran introduint progressivament al llarg de les diferents sessions de pràctiques. Addicionalment, la consola té una barra de menús principal amb diverses opcions, com ara les típiques de qualsevol programa en l'entorn Windows i la configuració de paquets i finestres.

- **Arxiu:** Per efectuar operacions bàsiques amb els fitxers (scripts, àrea de treball, historial).
- **Editar:** Es tracta del típic menú d'edició (copiar, enganxar, etc.). També s'usa per netejar la consola i editar les dades.
- **Visualitzar:** Per visualitzar o ocultar la barra d'eines i la barra d'estat.
- **Misc:** Per configurar opcions avançades.
- **Paquets:** Gestiona els diversos paquets que es poden carregar en *R*.
- **Finestres:** Per configurar les finestres.
- **Ajuda:** Facilita informació sobre el programa *R*.

Rstudio

Rstudio és un entorn lliure i de codi obert per al desenvolupament integrat (IDE) de *R*. Es pot executar a l'escriptori o fins i tot a través d'internet, mitjançant el servidor *Rstudio*. Aquest programa aplega tots els entorns i assumeix la filosofia de les expressions, però aporta algunes "ajudes" que fan més suportable el dia a dia. Està organitzat en quatre zones de treball diferents, com s'aprecia a la figura 1.6.

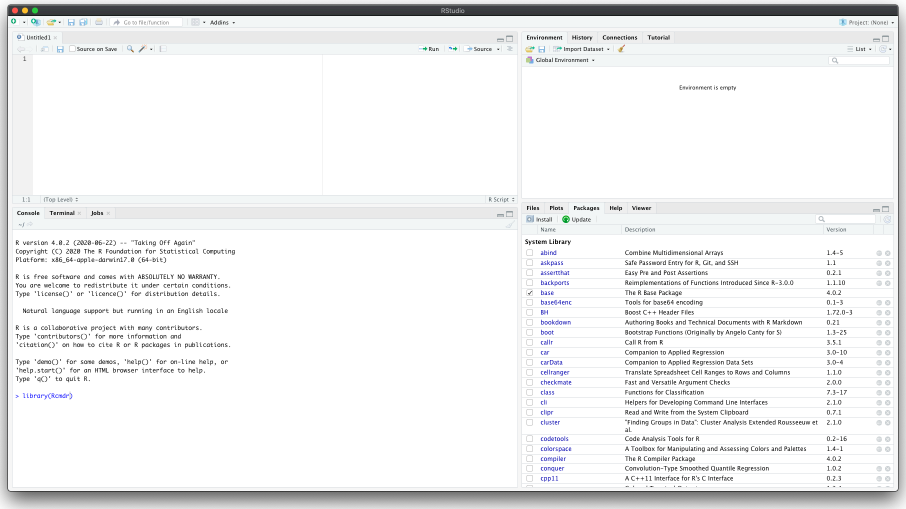


Fig. 1.6: Vista de *Rstudio*.

A la part superior esquerra, es pot obrir i editar un o diversos fitxers amb codi *R* (scripts) alhora. A la part inferior esquerra, hi ha una consola de *R* en què poden executar-se ordres de *R* individualment. La part superior dreta té quatre pestanyes, que són: Workspace, on apareix la llista dels objectes creats en la memòria; History, que conté l'historial de les línies de codi executades; Connections, on es pot fer una connexió a fonts de dades existents, i finalment Tutorial, on es pot obtenir informació addicional sobre els paquets desenvolupats en *R*. La part inferior dreta disposa de quatre pestanyes: Files, que dona accés a l'arbre de directoris i fitxers del disc dur; Plots, on apareixen els gràfics creats a la consola; Packages, que facilita l'administració dels paquets de *R* instal·lats en la màquina, i Help, on s'obren les pàgines d'ajuda.

Des de la barra del menú principal es pot accedir a tots els menús de *Rstudio*. Els menús Arxiu, Edició, Visualitza i Ajuda són habituals en els programes de l'entorn de Windows. La resta de menús són específics de *Rstudio*. Aquests permeten gestionar la interfície, és a dir, editar els fitxers, importar dades, instal·lar paquets, gestionar els gràfics, etc. Però el programa en cap moment permet fer un càlcul estadístic o una representació gràfica. Tot això ho hem de fer per mitjà d'ordres, igual com en *R-console*.

R-Commander

R-Commander és una interfície gràfica d'usuari (GUI, la sigla en anglès) creada per John Fox que permet accedir a moltes funcionalitats de l'entorn estadístic *R* sense que l'usuari hagi de conèixer el llenguatge d'ordres propi d'aquest entorn. Per utilitzar-lo, s'ha d'obrir *R* o *Rstudio* i executar-lo des de la consola (*R-Commander* no és una aplicació que funcioni sola). La finestra que apareix és la següent (figura 1.7):

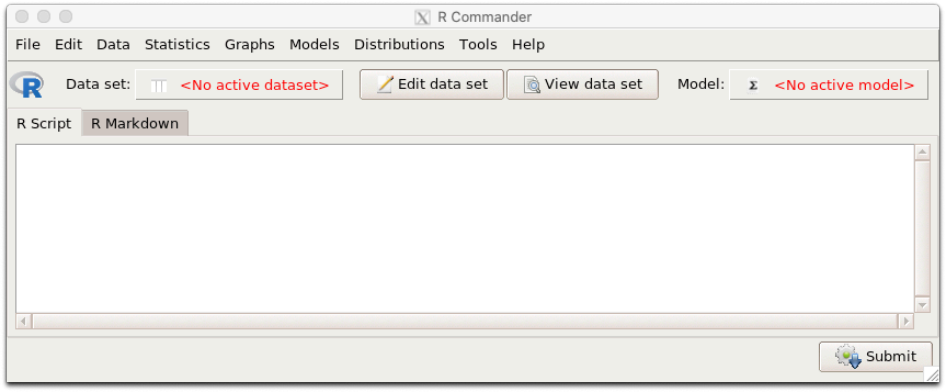


Fig. 1.7: Vista de la finestra de *R-Commander*.

Cada vegada que, a través dels menús, s'accedeix a les funcionalitats de *R* (gràfics, procediments estadístics, models, etc.), es mostra a la finestra d'instruccions (*R Script*) la

instrucció o conjunt d'instruccions que executen la tasca que s'ha sol·licitat, i a la consola (sigui de *R* o de *Rstudio*) es mostra el resultat de la instrucció. D'aquesta manera, encara que l'usuari no conegui el llenguatge d'ordres de *R*, simplement observant el que va apareixent en la finestra d'instruccions s'anirà familiaritzant amb el llenguatge.

Addicionalment, l'usuari pot introduir ordres directament en aquesta finestra i, després de clicar el botó Executar, aquestes ordres s'executen i el seu resultat es visualitza igualment. Les instruccions es poden desar i tornar a executar directament amb conjunts de diferents.

L'accés a les funcions implementades en *R-Commander* és molt simple i es fa utilitzant el ratolí per seleccionar, dins de la barra de menús principal situada a la primera línia de la finestra, l'opció a què vulguem accedir. S'hi pot trobar:

- **Arxiu:** Per obrir fitxers amb instruccions a executar, o per desar dades, resultats, sintaxi, etc.
- **Editar:** Conté les típiques opcions de retallar, enganxar, esborrar, etc.
- **Dades:** Utilitats per gestionar dades (creació de dades, importació des d'altres programes, recodificació de variables, etc.).
- **Estadístiques:** Per executar procediments pròpiament estadístics.
- **Gràfics:** Conté tots els gràfics disponibles.
- **Models:** Permet definir i utilitzar models específics per a l'anàlisi de dades.
- **Distribucions:** Ofereix probabilitats, quantils i gràfics de les distribucions de probabilitat més habituals (Normal, t de Student, F de Fisher, binomial, etc.).
- **Eines:** Permet carregar biblioteques i definir l'entorn.
- **Ajuda:** Ajuda sobre *R-Commander* (en anglès).

Addicionalment, hi ha una barra d'eines sota la barra de menús amb els botons següents:

- **Conjunt de dades:** Mostra el nom de la sèrie de dades activa. Al començament no hi ha cap sèrie de dades activa. En prémer aquest botó, es pot triar entre les sèries de dades que hi ha actualment en la memòria (si n'hi ha més d'una).
- **Editar conjunt de dades:** Permet obrir l'editor de dades de *R* per modificar la sèrie de dades activa.
- **Visualitzar:** Permet obrir l'editor de dades de *R* para examinar la sèrie de dades activa.
- **Model:** Indica el nom del model estadístic actiu, un model lineal (com el model de regressió lineal), un model lineal generalitzat, etc. Inicialment no hi ha cap model actiu.

1.3. Primers passos amb R

Per començar, es descriuen els primers passos per introduir dades, fer operacions, calcular funcions i representar gràficament les dades i l'anàlisi estadística utilitzant el llenguatge *R*, sigui a la consola de *R*, en *Rstudio* o en *R-Commander*. Finalment, es mostrarà com es desa una sessió.

1.3.1. R com a calculadora bàsica

En *R* es poden fer operacions de càlcul numèric bàsiques com ara: suma (+), resta (–), multiplicació (||), divisió (/), divisió entera (%/%), residu (%%), potència (^), etc. A més a més, estan disponibles operacions lògiques com: igual (==), més gran que (>), més petit que (<), més gran o igual que (>=), més petit o igual que (<=), diferent (!=), i (*and*) (&), o (*or*) (||), etc. Per exemple:

```
2+2
```

```
[1] 4
```

```
2+3*5/6+4^2
```

```
[1] 20.5
```

```
31%%7
```

```
[1] 3
```

```
202%%10
```

```
[1] 20
```

Els operadors <– i = s'utilitzen per fer assignacions. És preferible utilitzar el primer, ja que el signe igual té, en algunes ocasions, connotacions lògiques. La variable es crea en el mateix instant de l'assignació. Encara més, no es pot declarar amb anterioritat i deixar-la buida. A continuació es poden veure alguns exemples d'assignació de variables.

```
x <- 4
y = 6
z = x+y # NO mostra el resultat
z = x+y; z # SÍ mostra el resultat
```

```
[1] 10
```

```
(z = x+y) # SÍ mostra el resultat
```

```
[1] 10
```

Els nombres de les variables també poden contenir períodes, tret que es demarca amb el punt (.). Per exemple:

```
x.inicial <- 4
x.final = 10
```

```
x.dif = x.final-x.inicial; x.dif
```

```
[1] 6
```

En comparar dues variables, el resultat és una variable lògica que indica si la declaració és certa o falsa. Per exemple:

```
x==y
```

```
[1] FALSE
```

```
x!=y
```

```
[1] TRUE
```

```
x>y
```

```
[1] FALSE
```

R també té algunes constants integrades: π pi o les lletres de l'alfabet en anglès en majúscules i minúscules (LETTERS, letters), entre d'altres. Per exemple, es pot calcular el perímetre de la circumferència de la Terra a l'equador, sabent que té un radi de 6378 km.

```
pi
```

```
[1] 3.141593
```

```
2*pi*6378
```

```
[1] 40074.16
```

Addicionalment, hi ha moltes funcions matemàtiques integrades en *R*, entre les quals podem destacar: l'arrel quadrada (sqrt()), les funcions exponencials i logarítmiques (exp(), log(), log10()), les funcions trigonomètriques (sin(), cos(), tan()), el valor absolut (abs()), les funcions d'arrodoniment (ceiling(), floor(), trunc(), round()), etc. A continuació se'n poden veure uns quants exemples:

```
sin(45*pi/180)
```

```
[1] 0.7071068
```

```
sqrt(81)
```

```
[1] 9
```

```
exp(2)
```

```
[1] 7.389056
```

```
log(20)
```

```
[1] 2.995732
```

Tips & Tricks!

- Per executar les instruccions que hi ha en una línia, es prem la tecla Enter.
- Tot el que va precedit per coixinets (#), *R* ho considera un comentari i no ho interpreta.
- Diverses instruccions es poden executar en una mateixa línia si se separen per un punt i coma (;).
- Per visualitzar les dades assignades a una variable, s'introdueix el nom de la variable.
- Es poden recuperar línies d'instruccions introduïdes anteriorment prement la tecla de fletxa ascendent del teclat, a fi de tornar a executar-les o modificar-les.
- Per interrompre l'execució d'una instrucció i retornar el control a l'usuari, basta prémer la tecla [Esc] del teclat. Així recuperarem el símbol (>) per tornar a escriure les instruccions.

1.3.2. Vectors i matrius

L'ús de vectors i matrius és fonamental per poder organitzar les dades d'una manera apropiada per a l'anàlisi estadística posterior. Per tant, cal saber com definir-los, utilitzar-los i manipular-los en *R*.

Definició de vectors

Per construir un vector, primer es defineix un nom (per exemple, x). Tot seguit s'insereix l'operador d'assignació i després s'introdueix la lletra c (de *concatenar*). Finalment, s'escriuen els components del vector entre parèntesis i separats per comes.

```
x <- c(1,2,3,4,5); x
```

```
[1] 1 2 3 4 5
```

També es poden introduir les dades mitjançant el teclat amb la instrucció scan(). Els valors es teclegen deixant espais en blanc. Cada vegada que es prem la tecla Enter es canvia de línia i es pot continuar introduint valors. Per acabar, es prem Enter en una línia buida:

```
y = scan()
```

Una vegada definit el vector, mitjançant la funció length() se'n pot conèixer la longitud (nombre d'elements que la componen):

```
length(y)
```

```
[1] 0
```

Si el vector és una seqüència de valors enters (per exemple, d'1 a 10), es pot definir de la manera següent:

```
x1 <- 1:10; x1
```

```
[1]  1  2  3  4  5  6  7  8  9 10
```

Un vector també pot ser definit com una seqüència de valors equidistants mitjançant la funció seq(). Com a informació addicional, s'han de definir el valor inicial del vector (from =), el seu valor final (to =) i la distància entre valors (by =) o la longitud del vector (length =):

```
x2 <- seq(from=2, to=18, by=2); x2; length(x2)
```

```
[1]  2  4  6  8 10 12 14 16 18
[1] 9
```

```
x3 <- seq(from=2, to=18, length=30); x3; length(x3)
```

```
[1]  2.000000  2.551724  3.103448  3.655172  4.206897  4.758621  5.310345
[8]  5.862069  6.413793  6.965517  7.517241  8.068966  8.620690  9.172414
[15]  9.724138 10.275862 10.827586 11.379310 11.931034 12.482759 13.034483
[22] 13.586207 14.137931 14.689655 15.241379 15.793103 16.344828 16.896552
[29] 17.448276 18.000000
[1] 30
```

O simplement:

```
x2 <- seq(2,18,2); x2; length(x2)
```

```
[1]  2  4  6  8 10 12 14 16 18
[1] 9
```

També es poden definir com a repeticions d'un valor o d'un vector definit anteriorment:

```
x4 <- rep(1,5); x4; length(x4)
```

```
[1] 1 1 1 1 1
[1] 5
```

```
rep(x, length=8)
```

```
[1] 1 2 3 4 5 1 2 3
```

Fins i tot es poden definir mitjançant una fusió de les ordres anteriors:

```
x5 <- c(1:4, 8:10, seq(-7,5,by=2), rep(x,length=8)); x5; length(x5)
```

```
[1] 1 2 3 4 8 9 10 -7 -5 -3 -1 1 3 5 1 2 3 4 5 1 2 3
[1] 22
```

Manipulació de vectors

Si es vol accedir a un element específic d'un vector, s'introdueixen entre claudàtors el nom del vector i la posició de l'element. Per accedir a més d'un element, primer cal crear un vector amb les seves posicions.

```
x5[10] # Desè element del vector x5
```

```
[1] -3
```

```
x5[c(10,15,1)] # Desè, quinzè i primer element de x5
```

```
[1] -3  1  1
```

Es pot eliminar un o diversos elements d'un vector si s'introdueix la posició de l'element, precedida del signe menys (–) i entre claudàtors:

```
x6 <- x5[-10]; x6; length(x6)
```

```
[1] 1  2  3  4  8  9 10 -7 -5 -1  1  3  5  1  2  3  4  5  1  2  3
[1] 21
```

```
x7 <- x5[c(-8,-7,-2,-17)]; x7;
```

```
[1]  1  3  4  8  9 -5 -3 -1  1  3  5  1  2  4  5  1  2  3
```

```
x7 <- x5[-c(8,7,2,17)]; x7;
```

```
[1]  1  3  4  8  9 -5 -3 -1  1  3  5  1  2  4  5  1  2  3
```

Es pot inserir un nou element en un vector mitjançant la creació d'un nou vector utilitzant els elements del vector anterior:

```
x8 <- c(x6[1:4],20,x6[5:length(x6)]); x8; length(x8)
```

```
[1] 1 2 3 4  20  8 9 10 -7 -5 -1  1  3  5  1  2  3  4  5  1  2  3
[1] 22
```

Les operacions i funcions vistes anteriorment per a variables escalars es poden aplicar a vectors, amb l'excepció que les operacions es fan per a cada component del vector:

```
sin(c(0,30,45,60,90)*pi/180)
# Sinus de diversos angles donats en graus
```

```
[1] 0.0000000 0.5000000 0.7071068 0.8660254 1.0000000
```

```
exp(x6) # Exponencial d'un vector definit prèviament
```

```
[1] 2.718282e+00 7.389056e+00 2.008554e+01 5.459815e+01 2.980958e+03
[6] 8.103084e+03 2.202647e+04 9.118820e-04 6.737947e-03 3.678794e-01
[11] 2.718282e+00 2.008554e+01 1.484132e+02 2.718282e+00 7.389056e+00
[16] 2.008554e+01 5.459815e+01 1.484132e+02 2.718282e+00 7.389056e+00
[21] 2.008554e+01
```

```
x5>1
```

```
[1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
        FALSE FALSE
[13] TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

D'altra banda, també es pot accedir als elements d'un vector que compleixin una determinada condició. Per exemple, si es desitja conèixer el valor dels elements de x5 que són menors que 0, s'utilitza primer la funció which() per conèixer la posició dels elements que compleixen la condició:

```
ind <- which(x5<0); ind
```

```
[1]  8  9 10 11
```

```
x5[ind]
```

```
[1] -7 -5 -3 -1
```

O simplement:

```
x5[x5<0]
```

```
[1] -7 -5 -3 -1
```

A més de la funció length(), algunes de les funcions bàsiques per utilitzar vectors són les següents:

```
sum(x8) # Suma dels elements
```

```
[1] 74
```

```
range(x8) # Mínim i màxim
```

```
[1] -7 20
```

```
summary(x8) # Resum estadístic
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-7.000   1.000   3.000   3.364   4.750  20.000
```

```
sort(x8) # Organitza els elements de més petit a més gran
```

```
[1] -7 -5 -1 1 1 1 1  2  2  2  3  3  3  3  4  4  5  5  8  9 10 20
```

```
order(x8) # Mostra les posicions en organitzar-los de més petit a
          més gran
```

```
[1] 9 10 11 1 12 15 20 2 16 21 3 13 17 22 4 18 14 19 6  7  8  5
```

```
rev(x8) # Inverteix els elements del vector
```

```
[1] 3 2 1 5 4 3 2 1 5 3 1 -1 -5 -7 10 9 8 20 4 3 2 1
```

És molt freqüent en estadística tenir un conjunt de variables que descriuen a una sèrie d'individus, i que d'un individu concret (o diversos) no es disposi del valor d'una (o diverses) d'aquestes variables. *R* té en compte aquesta possibilitat; en aquests casos, el valor que manca apareix com a NA (de *non available*, no disponible) i algunes de les funcions bàsiques per tractar aquest tipus de dades són les següents:

```
v1 <- c(7,0,NA,8,5,6,NA,4);v1
```

```
[1]  7  0 NA  8  5  6 NA  4
```

```
is.na(v1) # Per saber on són aquests NA
```

```
[1] FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
```

```
sum(v1) # Com que falten valors, la suma no s'executa
```

```
[1] NA
```

```
sum(v1,na.rm=TRUE) # Fa l'operació sense considerar NA
```

```
[1] 30
```

Dins els vectors, també es poden emmagatzemar cadenes de caràcters. La sintaxi és similar; l'única diferència és que cada cadena ha d'anar entre cometes dobles:

```
v <- c("Jesús","Jaume","Xavier") ; v
```

```
[1] "Jesús"  "Jaume"  "Xavier"
```

Tips & Tricks!

- A les funcions en *R* s'hi poden agregar atributs.
- L'atribut (na.rm=TRUE) dona l'ordre que la funció s'executi sense tenir en compte les dades na.
- na significa *non available* (no disponible), rm significa *remove* (suprimeix) i TRUE significa cert. Aquest últim ha d'anar sempre en majúscules perquè, en cas contrari, no el reconeix.

Definició de matrius

Una matriu en *R* és un conjunt d'objectes ordenats per files i columnes. Una array en *R* és el mateix, tret del fet que pot tenir més de dues dimensions. En general, una matriu es pot crear de dues maneres: utilitzant la funció matrix o la funció array. Les dades que hi ha dins d'una matriu es manipulen de la mateixa manera que les que hi ha en vectors; la diferència és que ara s'ha de tenir en compte la posició de cada element en funció de les files i les columnes.

```
m1 <- matrix(1:20,nrow=5); m1
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    2    7   12   17
[3,]    3    8   13   18
[4,]    4    9   14   19
[5,]    5   10   15   20
```

```
m2 <- array(x5,dim=c(7,3)); m2
```

```
     [,1] [,2] [,3]
[1,]    1   -7    1
[2,]    2   -5    2
[3,]    3   -3    3
[4,]    4   -1    4
[5,]    8    1    5
[6,]    9    3    1
[7,]   10    5    2
```

```
m1[2,3] # Visualitza el valor de l'element de la fila 2 i la columna
          3 de m1
```

```
[1] 12
```

```
m1[2,c(1,3)] # Visualitza el valor dels elements de la fila 2,
               les columnes 2 i 3 de m1
```

```
[1]  2 12
```

```
m1[c(1:5),2] # Visualitza el valor de tots els elements de la columna
             2 de m1
```

```
[1]  6  7  8  9 10
```

També es poden visualitzar tots els elements d'una fila o d'una columna de la manera següent:

```
m2[2,] # Visualitza el valor de tots els elements de la fila 2 de m2
```

```
[1]  2 -5  2
```

```
m2[,3] # Visualitza el valor de tots els elements de la columna 3
       de m2
```

```
[1] 1 2 3 4 5 1 2
```

1.3.3. Estructures de dades

La manera més habitual d'emmagatzemar dades és utilitzar taules, data.frames en *R*. Un data.frame és com una matriu, una taula formada per files i columnes, amb la diferència que cada columna pot ser una variable de tipus diferent. En una taula poden coexistir columnes amb informació numèrica, entera, decimal; unes altres amb informació qualitativa de caràcters; unes altres lògiques, etc. El més freqüent és que aquestes taules tinguin dues dimensions (files i columnes), però en algun cas poden tenir més de dues dimensions. Per construir una estructura de tipus *data.frame*, s'utilitza la funció data.frames(v1,v2,...,v(n-1),vn), en què cada vector (vi) conté totes les dades de cada variable.

```
x <- c(2,2,1,2,1,1,1,2,2,1,1,2); n <- length(x);n
```

```
[1] 12
```

```
sex <- rep("Boy",n);sex
```

```
[1] "Boy" "Boy" "Boy" "Boy" "Boy" "Boy" "Boy" "Boy" "Boy" "Boy"
       "Boy" "Boy"
```

```
sex[x==2]="Girl"; sex
```

```
[1] "Girl" "Girl" "Boy"  "Girl" "Boy"  "Boy"  "Boy" "Girl"
        "Girl" "Boy"
[11] "Boy"  "Girl"
```

```
age <- c(3,6,4,2,8,9,5,4,4,7,1,10) ; age
```

```
[1]  3  6  4  2  8  9  5  4  4  7  1 10
```

```
table <- data.frame(age,sex); table
```

```
   age  sex             age  sex
1    3 Girl          7    5  Boy
2    6 Girl          8    4 Girl
3    4  Boy          9    4 Girl
4    2 Girl          10   7  Boy
5    8  Boy          11   1  Boy
6    9  Boy          12  10 Girl
```

El nom que s'associa a cada columna o variable dins de l'estructura és el nom que tenen els vectors. Per referir-nos a cada variable de l'estructura de dades per separat, s'utilitza el signe $ entre el nom del data.frame i el nom de la variable:

```
table$age
```

```
[1]  3  6  4  2  8  9  5  4  4  7  1 10
```

```
table$sex[4]
```

```
[1] "Girl"
```

Si no es vol haver d'utilitzar en tot moment el signe del dòlar, es pot fer ús de l'ordre attach(). Ara podrem accedir a qualsevol variable de la taula directament, únicament per mitjà del nom de la seva variable.

```
attach(table)
```

```
The following objects are masked _by_ .GlobalEnv:
age, sex
```

```
sex
```

```
[1] "Girl" "Girl" "Boy" "Girl" "Boy" "Boy" "Boy" "Girl" "Girl" "Boy"
[11] "Boy"  "Girl"
```

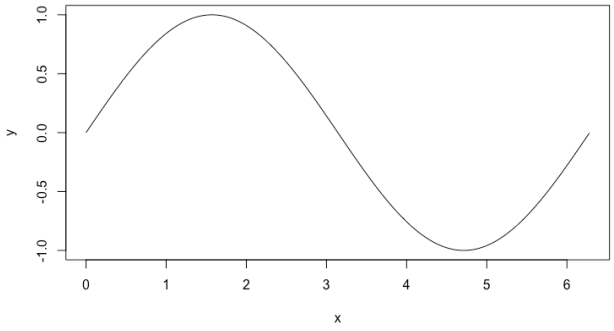
1.3.4. Funcions gràfiques bàsiques

Un altre gran avantatge de *R* és la seva capacitat gràfica. Els gràfics es poden exportar a diferents formats (PDF, EPS, JPG, etc.). Per veure una selecció de gràfics elaborats amb *R*, es pot executar un programa de demostració mitjançant la instrucció següent:

```
demo("graphics")
```

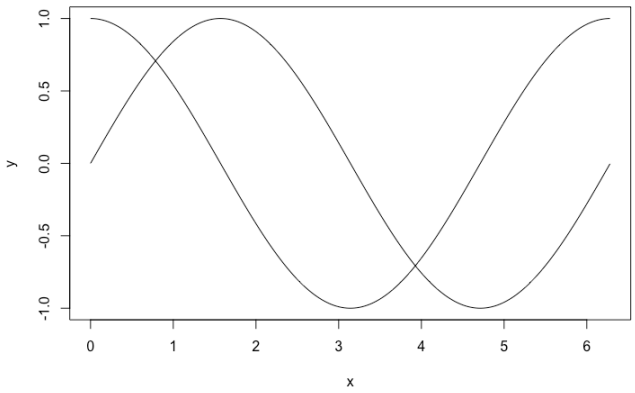
La funció bàsica i potser la més utilitzada per generar gràfics de sèries o dades és plot(). Prèviament, s'ha de definir el vector de les dades que seran representades en l'eix de les ordenades (eix y) i, si cal, el vector de les dades de l'eix de les abscisses (eix x). Com a atribut, s'especifica el tipus de gràfic (punts, línies, ambdós, etc.). Un exemple senzill és el següent:

```
x <- seq(0,2*pi,0.01); y <- sin(x); plot(x,y,type="l")
```



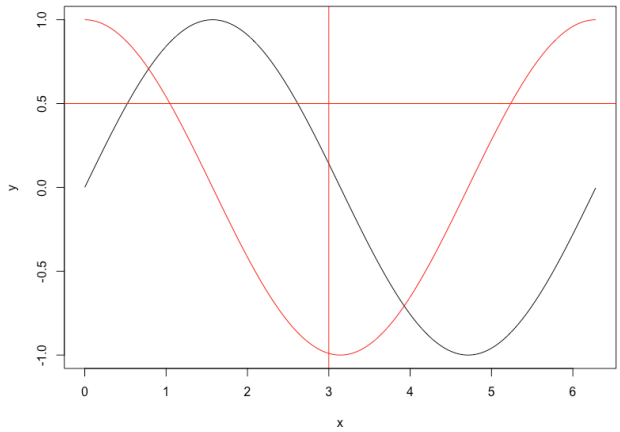
Es pot incloure la funció cosinus al gràfic que està actiu utilitzant la funció lines():

```
x <- seq(0,2*pi,0.01); z <- cos(x); lines(x,z,type="l")
```



També es poden generar múltiples gràfics en una sola finestra de la manera següent:

```
x=seq(0,2*pi,0.01)
y=sin(x); z=cos(x)
plot(x,y,type="l")
par(col='red')
lines(x,z)
abline(h=0.5); abline(v=3)
```



1.3.5. Desar i recuperar la sessió

Quan acabem una sessió de *R*, tenim l'opció de desar l'àrea de treball (*Environment*) o l'historial (*History*). L'àrea de treball inclou tots els objectes definits per l'usuari, que es van emmagatzemant en la memòria intermèdia mentre s'està treballant, però s'eliminen quan es tanca una sessió. D'altra banda, l'historial és el conjunt de totes les ordres que s'han utilitzat en la sessió. Si es vol desar un objecte en concret (per exemple, el vector x) en el fitxer MyData.Rdata, s'utilitza la instrucció següent:

```
save(x,file="MyData.Rdata")
```

Per desar tots els objectes existents en l'àrea de treball:

```
save.image("MyData.Rdata")
```

I per recuperar tots els objectes prèviament desats en un fitxer:

```
load("MyData.Rdata")
```

Si es desitja desar o carregar un historial d'ordres:

```
savehistory("MyData.Rdata"); loadhistory("MyData.Rdata")
```

Mitjançant la barra de menús també es poden desar i carregar tant els objectes de l'àrea de treball com l'historial.

Tips & Tricks!

- Per netejar la consola, usem [Ctrl+L].
- Per visualitzar els objectes emmagatzemats en l'àrea de treball, ls().
- Per eliminar un objecte, rm(name).
- Per eliminar tots els objectes, rm(list=ls()).
- Per visualitzar el directori de treball (*working directory*), getwd().
- Per ajustar el directori de treball a l'especificat, setwd("elmeudirectori").

1.3.6. Scripts o guions, la manera d'organitzar la sessió

Fins ara s'ha treballat directament en la consola de *R* o *Rstudio* i s'ha definit com es desen totes les instruccions que s'han executat (les correctes i les errònies). No obstant això, aquesta manera de desar una sessió no és la més aconsellable. Es recomana que la feina que es faci en qualsevol entorn de programació, sense que *R* en sigui l'excepció, es desi en forma de scripts o guions. Un script no és més que un document de text net que conté el conjunt d'instruccions o codis que es vol executar. S'hi poden registrar comentaris de cada instrucció o d'un conjunt d'instruccions. D'aquesta manera, el nostre codi queda desat d'una manera organitzada i clara per poder-la recuperar en una sessió futura.

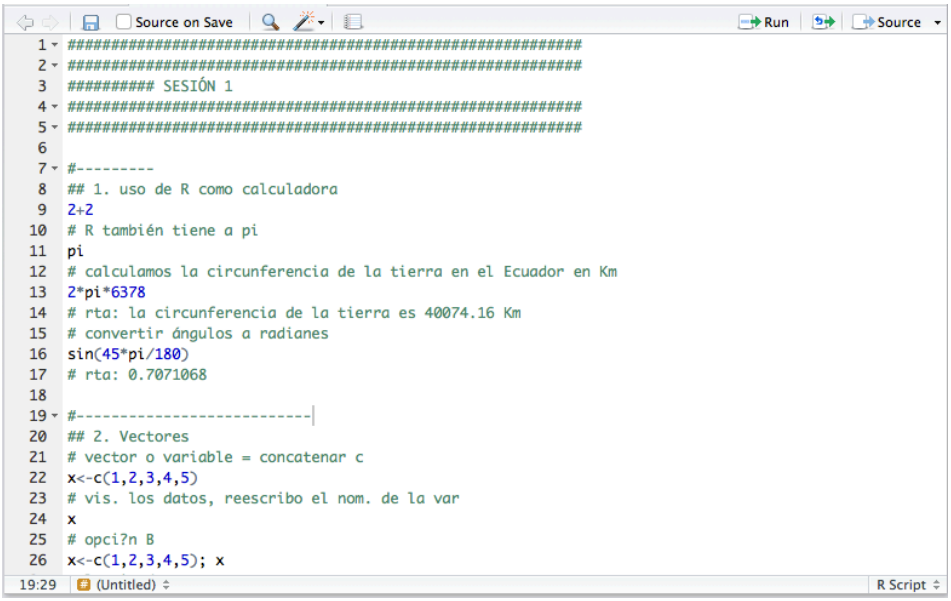


Fig. 1.8: Vista d'un script.

Per crear un nou script, seleccionem a la barra de menús "Fitxers", "Nou fitxer" i finalment "R script". També podem fer aquesta acció amb el teclat, prement simultàniament [Ctrl+Shift+N]. Aquest document es pot editar, modificar, desar i executar (totes les instruccions, part d'aquestes o només les línies desitjades). Un exemple de la sessió actual es pot veure a la figura 1.8.

Tips & Tricks!

- Per executar una línia d'instruccions des de la finestra *R Script* de *R-Studio*:
  - Es prem [Ctrl+Intro] al teclat tenint el cursor en qualsevol posició d'aquesta línia.
  - Es fa clic al botó "Executar" amb el ratolí.
  - Per executar tot l'script, es prem al teclat [Ctrl+A] i després [Ctrl+Intro].
- Per executar una línia d'instruccions des de la finestra *R Script* de *R-Commander*:
  - Es prem [Ctrl+R] al teclat tenint el cursor en qualsevol posició d'aquesta línia.
  - Es fa clic al botó "Executar" amb el ratolí.
  - Per executar tot l'script, es prem al teclat [Ctrl+A] i després [Ctrl+R].
- Tot el text precedit pel caràcter del coixinet, #, és ignorat per *R*; per tant, s'utilitza el coixinet per introduir comentaris.

1.3.7. Material extra

Als llocs web següents es poden trobar els fitxers necessaris per instal·lar *R*, *R-Commander* i *Rstudio*, així com diversos tutorials per ampliar la informació presentada en aquesta guia de pràctiques.

- https://www.r-project.org/
- https://www.rstudio.com/products/rstudio/features/
- http://www.rcommander.com/
- https://support.rstudio.com/hc/en-us

Igualment, hi ha una versió de *Rstudio* per treballar directament des d'internet:

- https://rstudio.cloud/

1.4. Exercicis que es proposen

1. Creeu un vector que contingui 12 valors: els quatre primers que siguin iguals a 3, els quatre següents que siguin iguals a 6 i els últims quatre que siguin iguals a 18.

2. Introduïu el vector $x = (3, 4, 30, 6, 85, 9)$:

   a) Reemplaceu la segona dada per 8.

   b) Introduïu el número 11 entre el cinquè i el sisè.

3. Creeu el vector X, que ha de contenir la informació següent:

$$\left[0, \frac{\pi}{16}, 2\frac{\pi}{16}, 3\frac{\pi}{16}, ..., 16\frac{\pi}{16}\right]$$

   a) Calculeu la suma de totes les seves dades: $\sum X_i$.

   b) Creeu un vector I a partir del vector X eliminant les dades emmagatzemades en les posicions 4, 9, 14.

   c) Calculeu $\sum \sin(X_i) - \sum \cos(I_i)$.

   d) Compareu els dos resultats anteriors i determineu quin dels dos és més gran.

4. Creeu el vector X, que ha de contenir 100 dades entre -1 i 1 igualment espaiades.

   a) Calculeu la suma de totes les seves dades: $\sum X_i$.

   b) Calculeu la suma de totes les seves dades: $\sum e^{X_i}$.

5. Al llarg d'un any, els imports de les factures mensuals del mòbil han estat: 23, 33, 25, 45, 10, 28, 39, 27, 15, 38, 34, 29. Escriviu un script en *R* que creï una taula de dades amb mesos i càrrecs. L'script ha de respondre automàticament les preguntes següents, fins i tot si canvia qualsevol valor.

   a) Quant heu gastat en total a l'any?

   b) En quin mes heu gastat menys diners? Quant ha estat?

   c) En quin mes heu gastat més diners? Quant ha estat?

   d) En quins mesos heu gastat més que la mitjana?

2
Estadística descriptiva

2.1. Introducció i objectius

Suposem que tenim les dades següents de 1384 pacients amb gammapatia monoclonal de significat incert (MGUS, per les seves sigles en anglès): ID del pacient (*id*), edat (*age*), sexe (*sex*), hemoglobina (*hgb*), creatinina (*creat*), mida del sèrum monoclonal (*mspike*), temps de progressió en la neoplàsia maligna PCM (*ptime*), ocurrència del PCM (*pstat*, 0=no, 1=sí), temps fins a la mort (*futime*) i ocurrència de la mort (*death*, 0=no, 1=sí). A continuació, es mostren els valors de les primeres 18 observacions.

	id	age	sex	dxyr	hgb	creat	mspike	ptime	pstat	futime	death
1	1	88	F	1981	13.1	1.3	0.5	30	0	30	1
2	2	78	F	1968	11.5	1.2	2.0	25	0	25	1
3	3	94	M	1980	10.5	1.5	2.6	46	0	46	1
4	4	68	M	1977	15.2	1.2	1.2	92	0	92	1
5	5	90	F	1973	10.7	0.8	1.0	8	0	8	1
6	6	90	M	1990	12.9	1.0	0.5	4	0	4	1
7	7	89	F	1974	10.5	0.9	1.3	151	0	151	1
8	8	87	F	1974	12.3	1.2	1.6	2	0	2	1
9	9	86	F	1994	14.5	0.9	2.4	57	0	57	0
10	10	79	F	1981	9.4	1.1	2.3	136	0	136	1
11	11	86	M	1972	11.8	1.0	2.3	2	0	2	1
12	12	89	F	1983	11.3	1.3	1.2	108	0	108	1
13	13	87	M	1968	11.2	1.1	1.3	10	0	10	1
14	14	80	F	1985	13.1	1.0	1.3	14	0	14	1
15	15	85	M	1979	13.0	1.1	1.0	18	0	18	1
16	16	90	F	1985	14.1	1.2	0.5	43	0	43	1
17	17	94	F	1975	11.0	1.1	0.7	34	0	34	1
18	18	86	M	1980	16.0	1.5	1.9	67	0	67	1

A simple vista, quina informació rellevant podem veure en el fitxer? Què en podem concloure? És evident que el primer pas quan es treballa amb dades és l'exploració inicial. Les dades s'han d'organitzar, representar d'alguna manera més amena i resumir. Per exemple, es pot representar gràficament (com es mostra a la figura) la proporció entre homes i dones, analitzar el rang de l'edat més comuna amb la malaltia, el valor mitjà d'hemoglobina, etc.

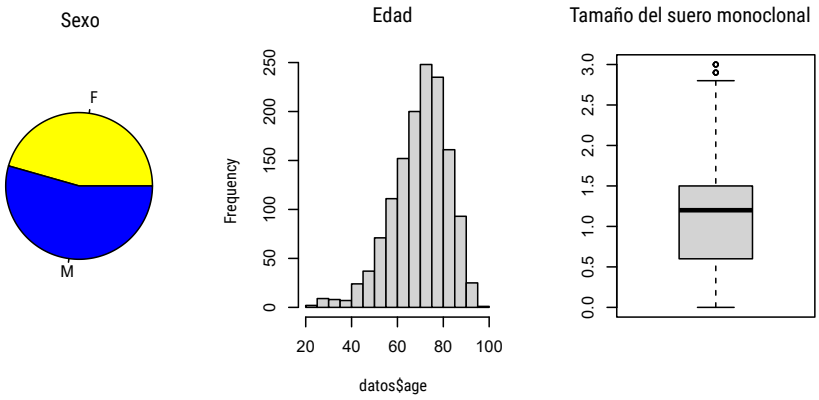


Fig. 2.1: Informació resumida i organitzada dels resultats de 1341 pacients amb MGUS.

En aquesta sessió es fa una introducció a les tècniques bàsiques per organitzar, representar i resumir un conjunt de dades. En l'estadística matemàtica, això es coneix com a *anàlisi exploratòria de dades* o *estadística descriptiva*. A més a més, es presenten les diferents maneres d'aplicar l'estadística descriptiva utilitzant *R*, *Rstudio* i *R-Commander* en un conjunt de dades per poder-les interpretar millor. En finalitzar aquesta sessió, l'estudiant ha de ser capaç de:

- Identificar les principals maneres de descriure, organitzar, representar i resumir un conjunt de dades.

- Construir taules de freqüències, representar-les gràficament, fer alguns càlculs estadístics importants (mitjana aritmètica, variància i moda) i interpretar tots aquests resultats en *R*, *Rstudio* i *R-Commander*.

2.2. Estadística descriptiva

L'estadística descriptiva és la disciplina de l'estadística que s'encarrega d'organitzar i resumir informació quantitativa per descriure les característiques principals d'un conjunt de dades. Sovint, el conjunt de dades inclou diferents *variables* (per exemple: velocitat, resistència, elasticitat, etc.). Per tant, el més usual és considerar les variables d'una en una, sense tenir en compte la possible correlació que hi hagi entre elles. Segons les seves característiques, es poden trobar *variables qualitatives* o *categòriques* (no necessiten nombres per expressar-se; per exemple: sexe, color, etc.) i *variables*

quantitatives o *numèriques* (sí que necessiten nombres per expressar-se; per exemple: edat, longitud, etc.). Per a cada variable, hi ha una sèrie d'observacions; les anotacions sobre quina modalitat (qualitatives) o quin valor (quantitatives) té cada observació es denominen *dades*. Aquestes dades es poden organitzar, resumir i representar mitjançant:

- **Taules:** Matrius on es desen les dades que corresponen a una determinada variable per a cada objecte. Per exemple, taules de freqüència.
- **Gràfics:** Representacions visuals de les taules que atorguen una visió més general i completa de les dades. Per exemple, gràfics de barres, histogrames, gràfics sectorials i polígons de freqüència.
- **Mides de tendència central:** Valors que pretenen proporcionar informació sobre el centre de la distribució de dades. Alguns exemples són la mitjana, la mediana i la moda.
- **Mides de variabilitat:** Valors que pretenen proporcionar informació sobre l'homogeneïtat dels valors entre ells. Alguns exemples són la desviació estàndard, la variància i els quartils.

2.2.1. Taula de freqüència

La manera més simple de presentar ordenadament dades categòriques és mitjançant una taula de freqüències. Aquesta taula indica el nombre de repeticions de cadascuna de les classes de la variable qualitativa. Es poden distingir els tipus de freqüències següents:

- **Freqüència absoluta** (n_i): És el nombre de repeticions que presenta una observació.
- **Freqüència relativa** (f_i): És la freqüència absoluta, dividida pel nombre total de dades.
- **Freqüència absoluta acumulada** (N_i): És la suma dels diferents valors de la freqüència absoluta prenent com a referència un individu donat.
- **Freqüència relativa acumulada** (F_i): És el resultat de dividir cada freqüència absoluta acumulada pel nombre total de dades.

Exemple 1: El conjunt de dades per al control de qualitat de l'aigua de diferents reactors és el següent, en què cada número representa el reactor que es va escollir com el millor:

1, 5, 3, 1, 2, 3, 4, 5, 1, 4, 2, 4, 4, 5, 1, 4, 2, 4, 2, 2

Reactor	Frequ. absoluta	Frequ. relativa	Frequ. abs. acumulada	Frequ. rel. acumulada
1	4	0.20	4	0.20
2	5	0.25	9	0.45
3	2	0.10	11	0.55
4	6	0.30	17	0.85
5	3	0.15	20	1.00

En *R*, la taula de freqüències es pot calcular de la manera següent:

```
datos_1 = c(1,5,3,1,2,3,4,5,1,4,2,4,4,5,1,4,2,4,2,2)
ni = table(datos_1) # Freqüència absoluta
fi = table(datos_1)/length(datos_1) # Freqüència relativa
Ni = cumsum(ni) # Freqüència absoluta acumulada
Fi = cumsum(fi) # Freqüència relativa acumulada
Tabla_Frec = cbind(ni,fi,Ni,Fi) # Taula de totes les freqüències
Tabla_Frec # Es visualitza la taula
```

```
   ni     fi   Ni     Fi
1   4   0.20    4   0.20
2   5   0.25    9   0.45
3   2   0.10   11   0.55
4   6   0.30   17   0.85
5   3   0.15   20   1.00
```

Exemple 2: Les resistències a la compressió de l'aliatge en lliures per polzada quadrada (psi) de 80 espècimens d'un nou aliatge d'alumini-liti sotmès a avaluació com a material possible per a elements estructurals d'aeronaus són:

105, 221,183, 186,121, 181,180, 143,167, 141,97, 154,153, 174,120, 168,176, 110,158, 133, 245, 228,174, 199,181, 158,156, 123,229, 146,163, 131,154, 115, 160, 208,158, 169, 148, 158, 207, 180, 190, 193,194, 133,150, 135,118, 149, 134, 178, 76,167, 184, 135, 218, 157, 101, 171, 165, 172, 199, 151, 142, 163, 145, 171, 160, 175, 149, 87, 160, 237, 196, 201, 200, 176, 150, 170

Quan els valors de la variable són molt nombrosos, convé agrupar les dades en intervals o classes per poder fer-ne una anàlisi i interpretació millor. Per construir una taula de freqüències amb dades agrupades, coneixent els intervals, s'han de determinar les freqüències corresponents a cada interval.

	ni	fi	Ni	Fi
70 <= x < 90	2	0.0250	2	0.0250
90 <= x < 110	3	0.0375	5	0.0625
110 <= x < 130	6	0.0750	11	0.1375
130 <= x < 150	14	0.1750	25	0.3125
150 <= x < 170	22	0.2750	47	0.5875
170 <= x < 190	17	0.2125	64	0.8000
190 <= x < 210	10	0.1250	74	0.9250
210 <= x < 230	4	0.0500	78	0.9750
230 <= x < 250	2	0.0250	80	1.0000

En *R*, la taula de freqüències amb dades agrupades es pot calcular de la manera següent:

```
datos_2=c(105,221,183,186,121,181,180,143,167,141,97,154,153,174,120,
          168,176,110,158,133,245,228,174,199,181,158,156,123,229,146,
          163,131,154,115,160,208,158,169,148,158,207,180,190,193,194,
          133,150,135,118,149,134,178,76,167,184,135,218,157,101,171,
          165,172,199,151,142,163,145,171,160,175,149,87,160,237,196,
          201,200,176,150,170)

breaks = seq(70,250,by=20);

breaks # Es crea el vector que conté els intervals
```

```
[1]  70  90 110 130 150 170 190 210 230 250
```

```
datos_2a = cut(datos_2, breaks, right=FALSE)
# Assigna cada valor a un interval

head(datos_2a, n=40) # Visualitza els primers 40 elements
```

```
[1] [90,110)  [210,230) [170,190) [170,190) [110,130) [170,190) [170,190)
[8] [130,150) [150,170) [130,150) [90,110)  [150,170) [150,170) [170,190)
[15] [110,130) [150,170) [170,190) [110,130) [150,170) [130,150) [230,250)
[22] [210,230) [170,190) [190,210) [170,190) [150,170) [150,170) [110,130)
[29] [210,230) [130,150) [150,170) [130,150) [150,170) [110,130) [150,170)
[36] [190,210) [150,170) [150,170) [130,150) [150,170)
9 Levels: [70,90) [90,110) [110,130) [130,150) [150,170) ... [230,250)
```

```
ni = table(datos_2a) # Freqüència absoluta

fi = table(Datos_2a)/length(Datos_2a) # Freqüència relativa

Ni = cumsum(ni) # Freqüència absoluta acumulada

Fi = cumsum(fi) # Freqüència relativa acumulada
```

```
Tabla_Frec = cbind(ni,fi,Ni,Fi) # Es crea una taula de totes
                                  les freqüències
Tabla_Frec # Es visualitza la taula
```

```
 ni        fi     Ni          Fi
[70,90)    2     0.0250      2 0.0250
[90,110)   3     0.0375      5 0.0625
[110,130)  6     0.0750      11 0.1375
[130,150) 14     0.1750      25 0.3125
[150,170) 22     0.2750      47 0.5875
[170,190) 17     0.2125      64 0.8000
[190,210) 10     0.1250      74 0.9250
[210,230)  4     0.0500      78 0.9750
[230,250)  2     0.0250      80 1.0000
```

Tips & Tricks!

- table() crea resultats tabulars de variables categòriques, és a dir, determina la freqüència absoluta de les dades.
- cumsum() calcula un vector els elements del qual són la suma acumulada del vector d'entrada.
- cbind() i rbind() combinen diversos objectes de *R* en un sol objecte: per columnes i per files, respectivament.
- cut() divideix el rang del vector *dades* en els intervals *breaks* i codifica els valors de les dades d'acord amb l'interval a què pertanyen.

2.2.2. Gràfics estadístics

Les distribucions de freqüències es poden presentar en taules com les anteriors o bé en gràfics. La representació gràfica s'utilitza per facilitar la comprensió dels resultats, però no afegeix cap informació extra respecte a la que contindria una taula de freqüències. No obstant això, com es diu popularment, una imatge val més que mil paraules. Hi ha diversos tipus de gràfics, cadascun apropiat a un tipus de variables. A continuació, es descriuen els més utilitzats i com s'elaboren en *R* utilitzant els dos exemples anteriors.

Diagrama de tija i fulles

És una eina que presenta una taula de dades en un format gràfic per ajudar a visualitzar la forma de la distribució. En la taula, cada dada es divideix, segons el seu valor, en una tija i una fulla. L'últim dígit de la dada representa la fulla, i els altres dígits, la tija. Els gràfics d'aquest tipus atorguen informació sobre la localització, la dispersió i els valors extrems de les nostres dades. El diagrama de tiges i fulles es calcula en *R*

mitjançant la funció stem(). La longitud del gràfic es pot modificar utilitzant l'atribut scale= , on 1 és el valor per defecte, 2 produeix un gràfic aproximadament el doble de llarg, etc. El gràfic de l'exemple 2 es genera de la manera següent:

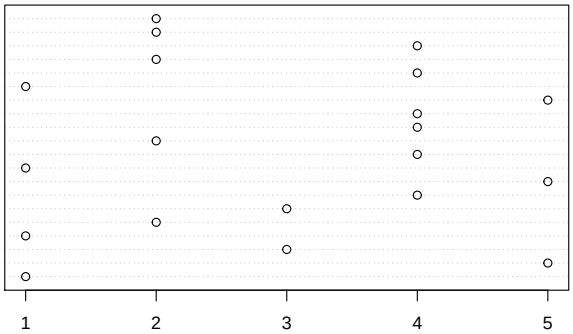
```
stem(datos_2,scale=2)
```

```
The decimal point is 1 digit(s) to the right of the |

  7 | 6                          16 | 0003357789
  8 | 7                          17 | 0112445668
  9 | 7                          18 | 0011346
 10 | 15                         19 | 034699
 11 | 058                        20 | 0178
 12 | 013                        21 | 8
 13 | 133455                     22 | 189
 14 | 12356899                   23 | 7
 15 | 001344678888               24 | 5
```

Gràfic de punts

Quan tenim una taula de freqüències petita de variables categòriques i els valors no disten gaire entre si, es tracta de representar les dades obtingudes d'una manera atractiva. En aquest gràfic, l'eix horitzontal representa els valors possibles de les dades i l'eix vertical correspon a la localització de cada dada dins de la llista. Cada dada es representa amb un punt i es col·loca damunt del valor que li correspongui i a una altura proporcional a l'ordre que té en el conjunt. A l'exemple 1, la primera dada (1) es representa amb un punt en la posició (1,1); la segona dada (5), un punt en la posició (5,2); el tercer, que és 3, en la posició (3,3); el quart, que és 1, en (1,4), i així successivament. El gràfic de punts es genera amb l'ordre dotchart(), tal com es mostra a continuació.

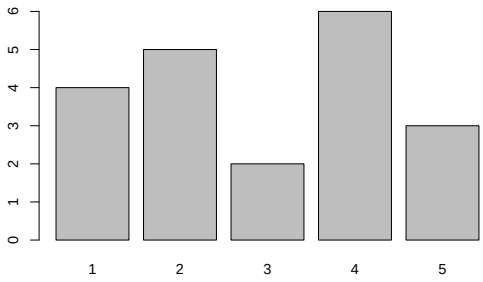
```
dotchart(datos_1)
```



Gràfic de barres

Aquest gràfic representa visualment la freqüència de variables categòriques mitjançant barres rectangulars d'una mateixa amplària. A cada categoria o classe de variable s'hi associa una barra l'altura de la qual representa la freqüència absoluta o la freqüència relativa d'aquesta classe. Per generar el gràfic de barres, s'utilitza l'ordre barplot(); ara bé, cal definir primer la taula de freqüències. Per a l'exemple 1, el gràfic de barres per a la freqüència absoluta és el següent:

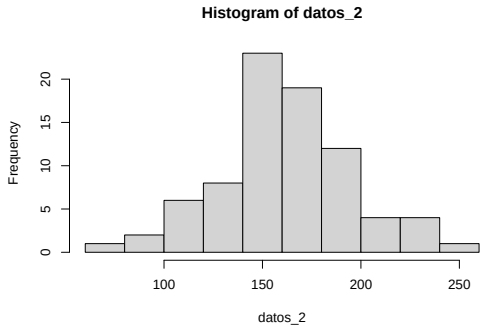
```
barplot(table(datos_1))
```



Histograma

És el gràfic apropiat per representar variables quantitatives amb un gran nombre de valors diferents. Les dades s'agrupen en intervals i es representen gràficament mitjançant rectangles juxtaposats les bases dels quals descansen sobre l'eix horitzontal i les altures del qual són tals que l'àrea de cada rectangle és proporcional a la freqüència de cada interval. Si tots els intervals tenen una longitud igual, llavors l'altura de cada rectangle és proporcional a la freqüència de l'interval. Per evitar confusions, la diferència principal amb el gràfic de barres és la inexistència d'espais entre rectangles. La funció hist() permet fer l'histograma d'unes dades i, a més, modificar la longitud dels intervals, si es vol. A diferència del gràfic de barres, la funció calcula automàticament la freqüència de l'interval. L'histograma de l'exemple 2 es genera de la manera següent:

```
h=hist(datos_2)
```



Si l'únic argument de la funció és el vector de dades, l'histograma es crea amb el nombre d'intervals (i, per tant, la seva longitud) calculats de manera automàtica. Si l'histograma es guarda en un objecte h = hist(), aquest objecte conté certa informació, com ara els límits dels intervals, la freqüència de cada interval, la seva densitat, el punt mitjà, etc.

```
h$breaks # Límits dels intervals
```

```
[1]  60  80 100 120 140 160 180 200 220 240 260
```

```
h$counts # Freqüència de cada interval
```

```
[1]  1  2  6  8 23 19 12  4  4  1
```

```
h$density # Densitat de cada interval
```

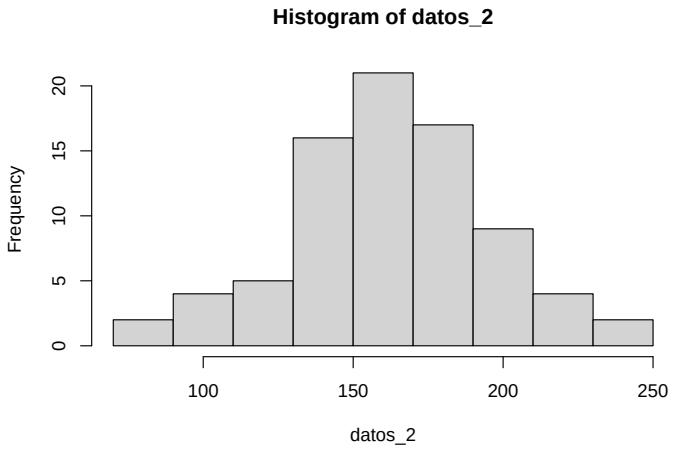
```
[1] 0.000625 0.001250 0.003750 0.005000 0.014375 0.011875
        0.007500 0.002500
[9] 0.002500 0.000625
```

```
h$mids # Punt central de cada interval
```

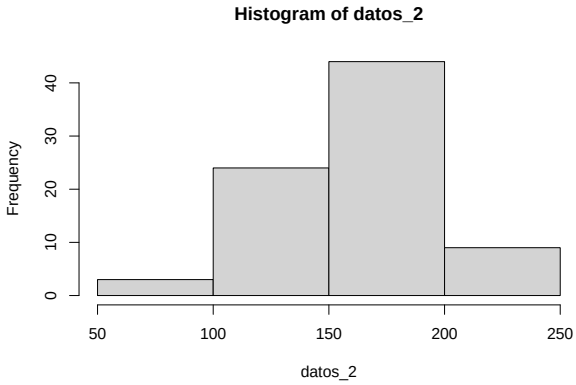
```
[1]  70  90 110 130 150 170 190 210 230 250
```

També es poden seleccionar els límits dels intervals o el nombre d'intervals en què es volen agrupar.

```
new_breaks = seq(70,250,by=20)
h1 = hist(datos_2,breaks=new_breaks)
```



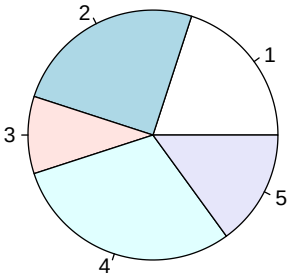

```
h2=hist(datos_2,breaks = 3)
```



Gràfic de sectors

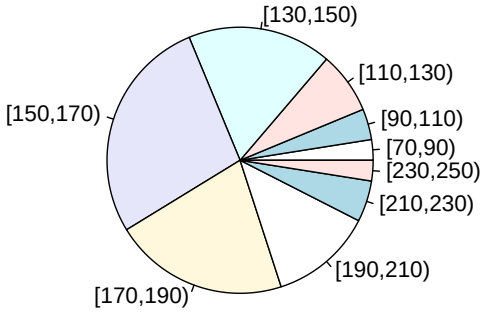
Aquest gràfic es representa com un cercle dividit en porcions, que són proporcionals a la freqüència relativa de cada categoria. La funció pie() permet crear el gràfic de sectors. Igual que en el gràfic de barres, cal definir prèviament la taula de freqüències. Per a l'exemple 1, el diagrama s'elabora de la manera següent:

```
pie(table(datos_1))
```



Si es vol crear el gràfic de sectors per a l'exemple 2, les dades s'han d'agrupar per poder visualitzar informació rellevant.

```
pie(table(datos_2a))
```



Tips & Tricks!

- Les funcions stem(), dotchart(), barplot(), hist() i pie() permeten resumir visualment les dades.
- Aquests gràfics es poden millorar definint-ne alguns atributs, com per exemple: col, main, names.arg, etc. Utilitzeu l'ajuda per saber-ne més coses.

2.2.3. Mesures de posició i tendència central

En certes ocasions és convenient resumir la informació d'un conjunt de dades numèriques en un sol valor per obtenir indicadors del comportament de la variable i poder fer comparacions. Les mesures de tendència central, també conegudes com a *mesures de posició* o *localització*, descriuen un valor al voltant del qual hi ha les observacions.

Mitjana

La mitjana, també coneguda com el valor mitjà, es defineix com la suma de tots els valors de cada observació (x_i), dividit pel nombre total d'observacions del conjunt de dades (N):

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

Si es disposa d'un conjunt de dades agrupades en què es coneix el valor mitjà de cada interval ($\bar{x}_i$) i el nombre de dades de cadascun (n_i), la mitjana és expressada per:

$$\bar{X} = \frac{x_1 n_1 + x_2 n_2 + x_3 n_3 + x_4 n_4 + \cdots + x_N n_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i n_i.$$

on $n_1 + n_2 + n_3 + n_4 + \cdots + n_n = N$. Per als exemples anteriors, les mitjanes es poden calcular d'acord amb la definició de la manera següent:

```
sum(datos_1)/length(datos_1)
```

```
[1] 2.95
```

```
sum(datos_2)/length(datos_2)
```

```
[1] 162.6625
```

No obstant això, la funció mean() calcula la mitjana directament.

```
mean(datos_1)
```

```
[1] 2.95
```

```
mean(datos_2)
```

```
[1] 162.6625
```

Mediana

La mediana és la dada que ocupa la posició central en la mostra ordenada de menor a major; és un punt que divideix la mostra ordenada en dos grups iguals (deixa el 50% dels valors per sota i l'altre 50% per damunt). Per calcular-la, s'ordenen les dades de més petit a més gran, i la dada central és la que ocupa la posició $\frac{N+1}{2}$, on N és el nombre total de dades. Si N és imparell, la mediana és la dada central mateixa; si N és parell, existeixen dues dades centrals, per la qual cosa la mediana és la mitjana de totes dues. Igualment, hi ha una funció que la calcula directament: median().

```
median(datos_1)
```

```
[1] 3
```

```
median(datos_2)
```

```
[1] 161.5
```

Moda

La moda és el valor amb freqüència absoluta més gran en les dades obtingudes. Indica quin és el valor més freqüent, però no quantes vegades es repeteix. Si hi ha més de dos valors que es repeteixin amb una freqüència més gran, es diu que les dades són *multimodals*. Es pot calcular la moda seguint les instruccions següents:

```
table(datos_1)
```

```
datos_1
1 2 3 4 5
4 5 2 6 3
```

```
# S'organitza la taula de freqüències de valor més gran
(el més freqüent) a més petit
freq_ord=sort(table(datos_1), decreasing = TRUE); freq_ord
```

```
datos_1
4 2 1 5 3
6 5 4 3 2
```

```
# Es pren el valor o valors que més es repeteixin
(el primer de la taula ordenada)
moda = names(freq_ord[1]); moda
```

```
[1] "4"
```

Quantils

Els quantils són valors de la llista de dades que la divideixen en parts iguals, és a dir, en intervals que comprenen el mateix nombre de valors. Els més usats són els percentils, els decils i els quartils. Els percentils són 99 valors que divideixen en cent parts iguals el conjunt de dades ordenades. Per exemple, el percentil d'ordre 15 deixa per sota el 15% de les observacions i per damunt queda el 85%. Els decils són els nou valors que divideixen el conjunt de dades ordenades en deu parts iguals; són un cas particular dels percentils. Els quartils són els tres valors que divideixen el conjunt de dades ordenades en quatre parts iguals; són també un cas particular dels percentils. En *R*, qualsevol d'aquests es calcula amb la funció quantile(), en què addicionalment s'ha d'especificar el quantil o els quantils desitjats (com un valor entre 0 i 1) de la manera següent:

```
quantile(datos_2,0.95) # Percentil d'ordre 95
```

```
   95%
221.35
```

```
quantile(datos_2,seq(0.1,0.9,by=0.1)) # Tots els decils
```

```
  10%   20%   30%   40%   50%   60%   70%   80%   90%
119.8 135.0 149.0 156.6 161.5 170.4 176.6 186.8 201.6
```

```
quantile(datos_2,seq(0.25,0.75,by=0.25)) # Tots els quartils
```

```
  25%   50%   75%
144.5 161.5 181.0
```

Finalment, el rang interquartílic és l'extensió coberta per la meitat central de les dades ordenades, excloent-ne la quarta part inicial (els que són inferiors al primer quartil) i la quarta part final (els que són superiors al tercer quartil). La funció IQR() calcula directament el rang interquartílic.

```
quantile(datos_2,0.75) - quantile(datos_2,0.25)
```

```
 75%
36.5
```

```
IQR(datos_2)
```

```
[1] 36.5
```

La mitjana, la mediana, el mínim, el màxim i els quartils es poden calcular directament mitjançant la funció summary().

```
summary(datos_2)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
76.0   144.5   161.5   162.7   181.0   245.0
```

2.2.4. Mesures de variabilitat i dispersió

Les mesures de posició donen una idea d'on se situa el centre de la distribució, però no ens diuen com és de dispers el conjunt de dades. Les mesures de dispersió o variabilitat descriuen com es troben de pròximes les dades entre elles o en relació amb alguna mesura de tendència central.

Rang

És l'interval entre el valor màxim i el valor mínim del conjunt de dades. És altament sensible als valors extrems, és a dir, és un paràmetre estadístic feble. Amb la funció range(), també s'obtenen el valor mínim i el màxim del conjunt de dades; per tant, per calcular el rang, n'hi ha prou de calcular-ne la diferència.

```
max(datos_1)-min (datos_1)
```

```
[1] 4
```

```
max(datos_2)-min(datos_2)
```

```
[1] 169
```

```
diff(range(datos_1))
```

```
[1] 4
```

```
diff(range(datos_2))
```

```
[1] 169
```

Variància i desviació típica

Aquestes mesures determinen com se situen de lluny les dades en relació amb la mitjana. Específicament, expressen "la distància intermèdia de cada punt respecte de la mitjana". La variància es calcula de la manera següent:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2.$$

on x_i és el valor de cada observació, $\overline{X}$ és la mitjana i N és el nombre total de dades. Noteu que les unitats de la variància s'expressen al quadrat; per tant, si tenim dades de longitud (en *mm*), en resulta un variància amb unitats de superfície (en mm^2), cosa que no té gaire sentit. Així doncs, disposem de la desviació estàndard o típica, que no és res més que l'arrel quadrada de la variància; d'aquesta manera, les unitats de la mesura de dispersió són les mateixes que les de les dades:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}.$$

Per a l'exemple 1, la variància i la desviació típica es poden calcular usant la definició de la manera següent:

```
sum((datos_1-mean(datos_1))2)/length(datos_1) # Variància
```

```
[1] 1.9475
```

```
sqrt(sum((datos_1-mean(datos_1))2)/length(datos_1))# Desviació típica
```

```
[1] 1.395529
```

En *R*, la variància i la desviació estàndard es poden calcular mitjançant les funcions var() i sd(), respectivament; no obstant això, aquestes funcions utilitzen $N-1$ (o $\sqrt{N-1}$) en el denominador, en lloc de N (o $\sqrt{N}$), per poder-les utilitzar com a estimadors no esbiaixats en inferència estadística. Aquestes mesures es coneixen com la variància i la desviació típica corregides. Per tant, per conèixer la variància i la desviació típica sense corregir, s'han de multiplicar pels factors $\frac{N-1}{N}$ i $\sqrt{\frac{N-1}{N}}$, respectivament.

```
var(datos_1) # Variància corregida
```

```
[1] 2.05
```

```
N = length(datos_1)
((N-1)/N)*var(datos_1) # Variància NO corregida
```

```
[1] 1.9475
```

```
sd(datos_1) # Desviació típica corregida
```

```
[1] 1.431782
```

```
sqrt((N-1)/N)*sd(datos_1) # Desviació típica NO corregida
```

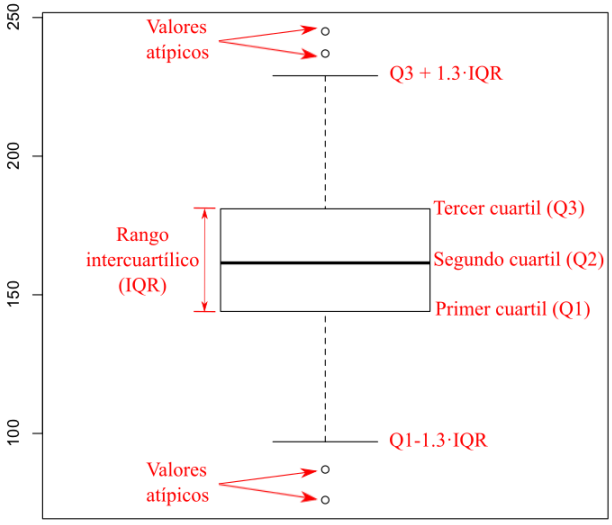
```
[1] 1.395529
```

2.2.5. Gràfic de caixa

Els diagrames de caixa són una presentació visual que descriu diverses característiques importants alhora, com ara la tendència central, la dispersió i la simetria. Per elaborar-los, es representen els tres quartils i els valors mínim i màxim de les dades sobre un rectangle, tot alineat horitzontalment o verticalment. Els valors amb dispersió fins a 1.3 vegades el rang interquartílic es representen com unes línies rectes o bigotis. Els valors fora d'aquest interval es representen mitjançant punts i es consideren valors extrems atípics.

boxplot() és la funció que s'utilitza per crear el gràfic. Igual com amb l'histograma, si es desa el gràfic de caixa en un objecte h = boxplot(), aquest objecte conté informació com ara els límits per considerar els valors atípics, els valors atípics, els quartils, etc.

```
bp = boxplot(datos_2); bp
```




```
$stats
      [,1]
[1,]  97.0
[2,] 144.0
[3,] 161.5
[4,] 181.0
[5,] 229.0

$n
[1] 80

$conf
        [,1]
[1,] 154.964
[2,] 168.036

$out
[1] 245  76  87 237

$group
[1] 1 1 1 1

$names
[1] "1"
```

Tips & Tricks!

- Les funcions mean(), median(), quantile(), IQR(), var(), sd() i boxplot() ens donen informació sobre la tendència central i la variabilitat de les dades.
- Recordeu que podeu consultar més informació sobre cada funció mitjançant la instrucció ?NomDeLaFunció; per exemple, ?boxplot.

2.3. Exercicis

Les dades següents es van extreure de la revista estatunidenca *Motor Trend* el 1974 i resumeixen el consum i deu aspectes de disseny i rendiment de 32 automòbils (models 1973-1974). Aquest conjunt de dades, que es denomina mtcars, conté 11 variables amb 32 observacions i està emmagatzemat en *R*. Per poder treballar amb les dades, només fa falta adjudicar un nom a l'objecte, com per exemple:

```
a = mtcars
```

Les variables són les següents:

mpg:	milles per galó de combustible
cyl:	nombre de cilindres
disp:	desplaçament
hp:	cavalls de potència
drat:	relació de l'eix posterior
wt:	pes (1000 lb)
qsec:	temps a 1/4 milla
vs:	V/S
am:	transmissió (0 = automàtic, 1 = manual)
gear:	nombre de marxes endavant
carb:	nombre de carburadors

Una vegada carregat el conjunt de dades, procedim a resoldre el qüestionari.

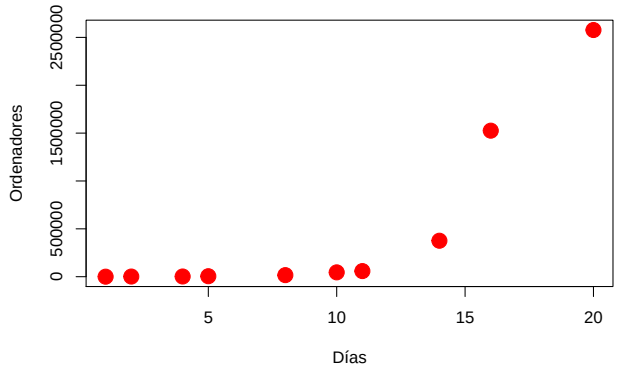
1. Determineu la mitjana, la mediana, la moda i la desviació estàndard de cadascuna de les variables. Es poden calcular per a totes les variables? Per a quines no? Justifiqueu la resposta.
2. Determineu quina variable presenta valors atípics. Com els heu trobat?
3. Creeu un gràfic de sectors per a cadascuna de les variables. De quines variables el gràfic no compleix l'objectiu d'impactar o ser més clar que la taula de dades?
4. Elaboreu l'histograma per a cadascuna de les variables usant 5 intervals. De nou, aquest gràfic és útil per a totes les variables? Justifiqueu la resposta.
5. Creeu un gràfic que inclogui el diagrama de caixes de totes les variables de manera que es puguin comparar.

3
Regressió lineal

3.1. Introducció i objectius

En la sessió 2, "Estadística descriptiva", hem après a resumir i a presentar un conjunt de dades en taules i gràfics. Aquests ajuden a interpretar-les amb vista a la presa de decisions. No obstant això, de vegades, quan es tenen les dades organitzades i representades en un gràfic, pot sorgir la necessitat d'estimar un valor de què no es disposa per diferents motius, sigui per pèrdua de la informació, per un error o fallada del sensor, per fallada tècnica, etc. Per exemple, el nombre de dies que han passat des que s'ha detectat un virus informàtic i el nombre d'ordinadors infectats en una determinada regió d'Europa estan registrats en la taula següent i representats en el seu diagrama de dispersió:

Dies	Ordinadors
1	255
2	1500
4	2105
5	5050
8	16300
10	45320
11	58570
14	375800
16	1525640
20	2577000



En aquesta sessió es podrà donar resposta a una sèrie de preguntes que poden sorgir, com ara: Quants ordinadors es van infectar al cap de 12 dies? Quina quantitat d'ordinadors estaran infectats al cap de 22 dies? El cas dels ordinadors infectats al cap de 12 dies és un esdeveniment del passat que no ha estat mesurat. En canvi, estimar la quantitat d'ordinadors que estaran infectats al cap de 22 dies correspon a un esdeveniment futur immediat (no gaire llunyà). Les dues qüestions es poden resoldre aplicant la tècnica de la regressió lineal.

Aquesta tècnica és un model matemàtic bàsic de regressió i d'anàlisi predictiva d'ús habitual. La idea general de la regressió és resumir i estudiar la relació entre dues variables contínues, trobar la relació matemàtica lineal que hi ha entre elles. No obstant això, un bon observador pot inferir que, per exemple, la relació entre els dies i el nombre d'ordinadors *no* és lineal. Aquesta "complicació" es pot resoldre fàcilment per aplicar després la tècnica de la regressió lineal.

Ara bé, generalment els conjunts de dades amb què habitualment es treballa en estadística solen ser extensos o simplement s'obtenen d'algun equip de mesurament. Aleshores, el més còmode és que estiguin desades en fitxers típics de fulls de càlcul o com a fitxers de text. Per a aquesta tercera sessió, a més d'aplicar els conceptes de regressió lineal, també s'utilitza un conjunt de dades desat en un fitxer extern. Per tant, a la primera part de la sessió es detalla el procediment per importar-les als entorns amb què es pot treballar: *R-Console*, *R-Commander* i *Rstudio*. Addicionalment, s'estudien les dues mesures que descriuen l'encert de la relació trobada en la regressió lineal. Finalment, es presenten les instruccions bàsiques per desenvolupar la tècnica utilitzant *R-Console*, *R-Commander* i l'entorn *Rstudio*. En finalitzar aquesta sessió, l'estudiant ha de ser capaç de:

- Importar dades utilitzant la consola *R*, així com el paquet *R-Commander* i l'entorn *Rstudio*.
- Comprendre el concepte del criteri de mínims quadrats.
- Interpretar la intercepció i el pendent d'una equació de regressió estimada.
- Calcular, entendre i interpretar el coeficient de determinació i el coeficient de correlació.
- Obtenir l'estimació de la línia recta de regressió i els coeficients de determinació i de correlació utilitzant *R*.
- Estimar els valors que manquin o desconeguts a partir de la relació trobada.

3.2. Importar dades a R-Console, Rstudio i R-Commander

Generar una estructura de dades en *R-Console* pot ser treballós segons el volum de les dades. Per això, en alguns casos és més fàcil o còmode crear-les i/o editar-les prèviament utilitzant un full de càlcul. D'altra banda, pot ser que moltes dades les tinguem ja en altres programes o simplement poden procedir directament d'algun programa d'adquisició de dades. A continuació s'explica breument la manera d'importar dades utilitzant *R-Console*, així com el paquet *R-Commander* i l'entorn *Rstudio*.

3.2.1. Importar dades amb R-Console

Les dades contingudes en arxius externs es poden importar directament des de la consola mitjançant la funció read.table(file,header,sep,dec,...), en què s'ha d'especificar en primera instància el nom del fitxer. Depenent de la manera en què les dades estiguin organitzades dins del fitxer, s'han de definir els atributs heater, sep, dec:

- **file** és el nom de l'arxiu on estan desades les dades.
- **header** és un tipus de dada lògica que indica si l'arxiu té a la primera fila els noms de les variables.
- **sep** és el caràcter que separa les magnituds entre si.
- **dec** és el caràcter que separa la part sencera de la part decimal d'un número qualsevol (normalment és un punt o una coma).

Per exemple, quan obrim el fitxer coches.txt en qualsevol editor de text, es pot apreciar que el nom de les variables consta a la primera fila del fitxer, les dades estan separades per tabulacions i el caràcter decimal està definit pel punt, tal com es mostra a la figura.

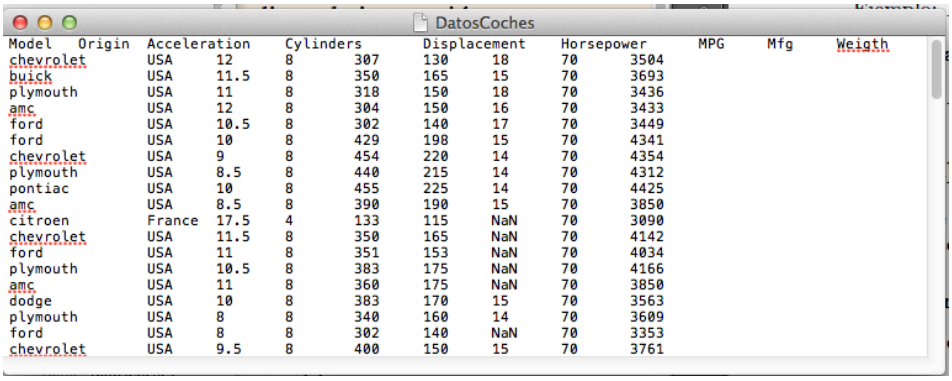


Fig. 3.1: Vista del fitxer "coches.txt".

Per importar-ne les dades i desar-les en una estructura de dades (data.frame) anomenada Datos, s'executa:

```
Datos = read.table("datos_Sesion3/coches.txt",
                   header=TRUE, sep="\t", dec=".")
head(Datos)
```

```
         Model  Origin Acceleration Cylinders Displacement
1 chevrolet     USA             12.0         8          307
```

```
2 buick          USA             11.5          8          350
3 plymouth       USA             11.0          8          318
4 amc            USA             12.0          8          304
5 ford           USA             10.5          8          302
6 ford           USA             10.0          8          429

  Horsepower MPG Mfg Weight
1        130  18  70   3504
2        165  15  70   3693
3        150  18  70   3436
4        150  16  70   3433
5        140  17  70   3449
6        198  15  70   4341
```

3.2.2. Importar dades amb Rstudio

Per importar dades des de *Rstudio*, hem d'anar a Tools > Import Dataset > From Local File. Aquí seleccionem el fitxer que volem importar. En el nostre cas, importem el fitxer"coches2".txt.

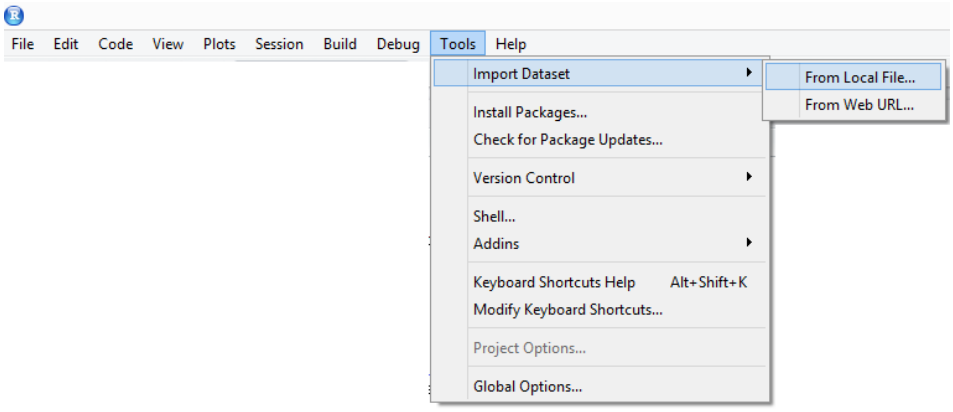


Fig. 3.2: Importar un fitxer de dades des de *Rstudio*.

Després de seleccionar el fitxer, a la finestra emergent *Import Dataset* s'han d'indicar diverses característiques que concerneixen el fitxer (igual que en *R-Console*): si el nom de les variables està inclòs a la primera fila del fitxer o no, com estan separades les dades, el caràcter utilitzat per separar els nombres decimals, etc.

Una vegada feta amb èxit la importació de les dades, a la cantonada superior dreta, a la pestanya de *Environment*, es pot observar que, entre les variables actuals a l'espai de treball, hi ha el fitxer "Dades". Noteu que s'especifica que és una estructura de dades (data.frame) de 100 observacions de 9 variables.

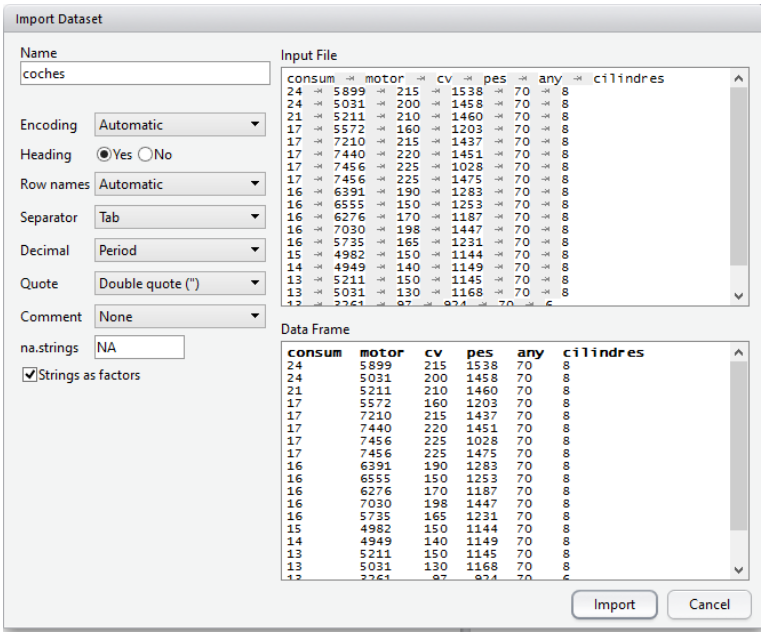


Fig. 3.3: Selecció de característiques del fitxer de dades des de *Rstudio*.

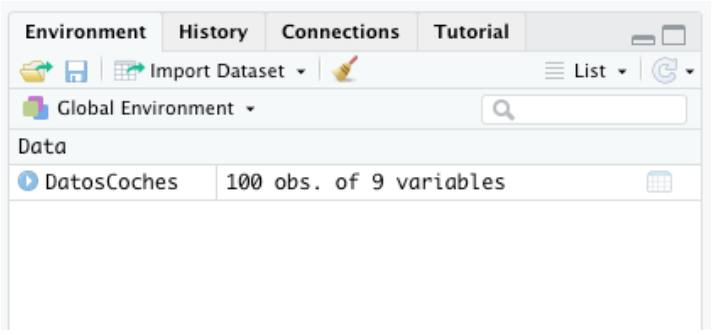


Fig. 3.4: Vista de les variables en l'entorn de treball en *Rstudio*.

3.2.3. Importar dades amb R-Commander

Si *R-Commander* no està en execució, n'hi ha prou de carregar-lo des de *R-Console* o *Rstudio*.

```
library(Rcmdr)
Loading required package: splines
Loading required package: RcmdrMisc
Loading required package: car
```

Per importar dades des de *R-Commander*, hem d'anar a *Datos* > *Importar datos* i seleccionar si l'origen de les dades és un arxiu de text, el porta-retalls o un URL.

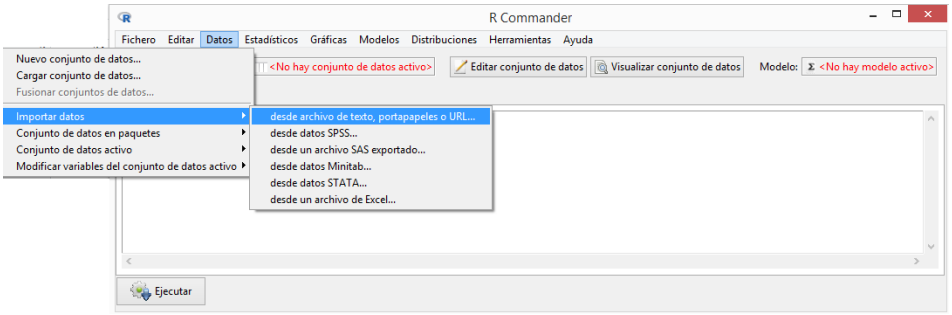


Fig. 3.5: Importar un fitxer de dades des de *R-Commander*.

De la mateixa manera que en *R-Console* i *Rstudio*, s'han d'indicar les característiques del fitxer de dades.

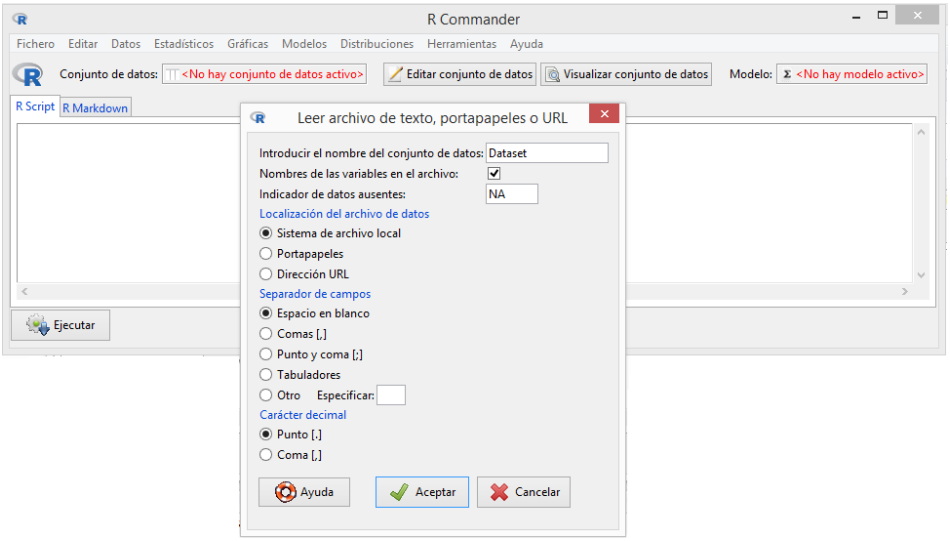


Fig. 3.6: Selecció de característiques del fitxer de dades des de *R-Commander*.

Noteu que, a la finestra principal, sota el menú, la casella *Conjunto de datos* ara està activada amb el nom que li hem donat en importar el fitxer, en el nostre cas "Dataset". D'altra banda, a la pestanya *R Script* apareix la línia d'instruccions que es poden executar des de *R-Console* o *Rstudio*.

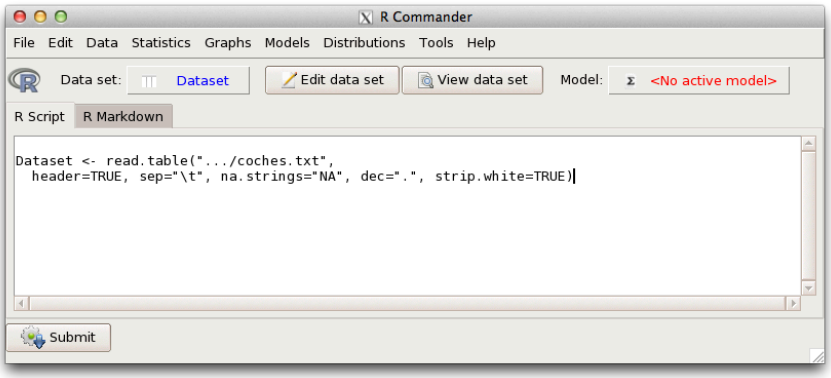


Fig. 3.7: Vista de la finestra *R Script* en *R-Commander* després d'haver importat un fitxer de dades.

Tips & Tricks!

- Per executar una línia d'instruccions des de la finestra *R Script* de *Rstudio*:
  - Premeu [Ctrl+Intro] al teclat tenint el cursor en qualsevol posició d'aquesta línia.
  - Cliqueu el botó *Executar* amb el ratolí.
  - Per executar tot l'script, premeu [Ctrl+A] al teclat i després [Ctrl+Intro].
- Per executar una línia d'instruccions des de la finestra *R Script* de *R-Commander*:
  - Premeu [Ctrl+R] al teclat tenint el cursor en qualsevol posició d'aquesta línia.
  - Cliqueu el botó *Executar* amb el ratolí.
  - Per executar tot l'script, premeu [Ctrl+A] al teclat i després [Ctrl+R].

3.3. Regressió lineal

La regressió lineal és una eina estadística que aporta l'habilitat d'estimar la relació matemàtica entre una variable dependent (o resposta, normalment y) i una variable independent (o predictor, normalment x). Té per objectiu principal utilitzar la informació obtinguda sobre un fenomen per predir-ne el comportament en el futur. Aquesta informació sol estar organitzada per parelles de valors observats i es representa gràficament en un núvol de punts o diagrama de dispersió.

Exemple 1: La relació entre la mida del lot a la fàbrica d'un cert producte i les hores de treball necessàries és expressada a la taula següent.

Lots	Hores
1	10
2	20
3	15
4	40
5	25

3.3.1. Model de regressió lineal simple

La regressió lineal simple consisteix a trobar la línia recta ($\hat{y} = mx + b$) que s'ajusti millor a través dels punts. La línia que s'ajusta millor es denomina *línia de regressió* o *recta de regressió*. A la figura 3.8, es pot apreciar el gràfic de dispersió de les dades de l'exemple 1 (punts en vermell). La línia diagonal blava és la línia de regressió i representa la predicció en y per a cada valor possible de x. Les línies verticals des dels punts fins a la línia de regressió representen els errors de predicció ($y_i - \hat{y}_i$). Com es pot veure, el punt en $x = 1$ està molt prop de la línia de regressió, de manera que el seu error de predicció és petit. En canvi, el punt en $x = 4$ està molt més lluny de la línia de regressió i, per tant, el seu error de predicció és més gran.

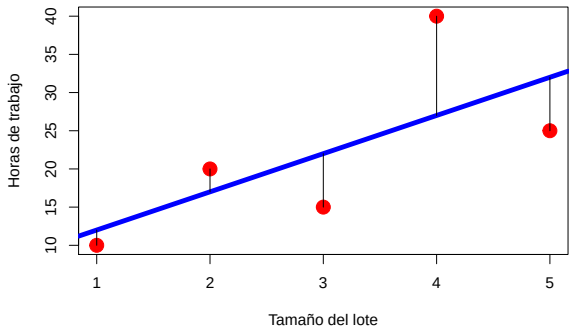


Fig. 3.8: Diagrama de dispersió i línia de regressió.

La línia recta que s'ajusta millor a les dades és aquella per a la qual els n errors de predicció (un per cada punt de dades) són tan petits com sigui possible en sentit general. Una manera d'aconseguir aquest objectiu és invocar el *criteri dels mínims quadrats*, que consisteix a minimitzar la suma dels errors de predicció al quadrat. És a dir, s'han de buscar els valors de m (pendent) i b (intercepció) tals que la suma del quadrat dels errors de predicció $Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ sigui la més petita possible. Per tant, s'ha de minimitzar l'equació:

$$Q = \sum_{i=1}^{n} (y_i - (mx_i + b))^2.$$

Si derivem i i igualem a zero, obtenim:

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

Per a l'exemple 1, tenim:

x	y	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
1	10	-2	4	-12	24
2	20	-2	1	-2	2
3	15	0	0	-7	0
4	40	1	1	18	18
5	25	2	4	3	6
$\bar{x} = 3$	$\bar{y} = 22$		$\sum_{i=1}^{5}(x_i - \bar{x})^2 = 10$		$\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = 50$

Per tant, $m = \frac{50}{10} = 5$ i $b = 22 - 5 \times 3 = 7$. D'aquesta manera, la recta de regressió és expressada per: $\hat{y} = 5x + 7$.

3.3.2. Model de regressió exponencial

En altres casos, la línia que uneix els valors obtinguts no s'aproxima a una recta, sinó a una funció exponencial. En altres paraules, s'assembla a una funció de tipus $y = \alpha e^{\beta x}$. Per tant, la regressió consisteix a trobar els valors de α i β que s'ajustin millor a les dades. En aquesta mena de regressions, també podem obtenir el valor del coeficient de determinació R^2, que segueix el mateix criteri que per a la regressió lineal (com més a prop estigui d'1, més precisa serà l'aproximació).

Encara que no ho sembli, aquesta aproximació no és més difícil que la lineal. Gràcies a les propietats dels logaritmes neperians, una relació exponencial es pot convertir en una relació lineal d'una manera molt senzilla:

$$\ln(y) = \ln\left(\alpha e^{\beta x}\right) = \ln(\alpha) + \ln\left(e^{\beta x}\right) = \ln(\alpha) + \beta \ln(e^x) = \ln(\alpha) + \beta x.$$

La solució al problema inicial seria la regressió lineal entre x i $\ln(y)$.
Reprenent l'exemple de la introducció d'aquesta sessió sobre el registre del nombre de dies que han passat des que s'ha detectat un virus informàtic i el nombre d'ordinadors infectats, es pot observar en el diagrama de dispersió que la relació entre les dues variables no és lineal, sinó que té un comportament exponencial.

Dies	Ordinadors
1	255
2	1500
4	2105
5	5050
8	16300
10	45320
11	58570
14	375800
16	1525640
20	2577000

En crear el diagrama de dispersió entre la variable Dies i el logaritme per a la variable Ordinadors, log(Ordinadors), s'observa que la relació té una tendència lineal.

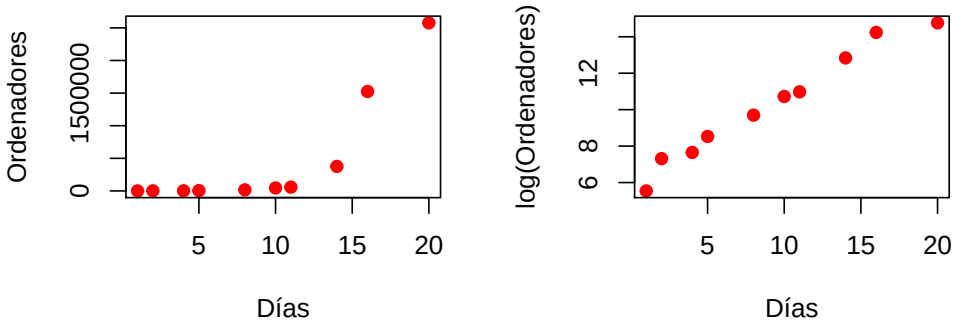


Fig. 3.9: Diagrama de dispersió: dies vs.ordinadors (esq) ; dies vs. log(Ordinadors).

Per tant, seguint el criteri de mínims quadrats, s'obté:

$$m = \beta = \frac{\sum_{i=1}^{10}(x_i - \bar{x})(y_i^* - \bar{y}^*)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{171.0931}{354.9} = 0.4821,$$

$$b = \ln(\alpha) = \bar{y}^* - m\bar{x} = 10.2269 - 0.4821 \times 9.1 = 5.8399,$$

$$\alpha = e^b = e^{5.8399} = 343.7072,$$

on $\bar{y}^*$ és el logaritme de la variable Ordinadors. D'aquesta manera, la línia de regressió és expressada per: $\ln(y) = 0.4821 + 5.8399x$, o per $y = 343.7072e^{0.4821x}$.

3.3.3. Avaluar l'exactitud del model de regressió

Hi ha diverses maneres d'avaluar en quina mesura el nostre model s'ajusta a les dades; la qualitat d'ajust i la de la regressió les determina normalment el *coeficient de determinació* (R^2) o el *coeficient de Pearson* (R). Aquests nombres característics de cada regressió indiquen com s'ajusta de bé la línia a les dades. Per exemple, $R^2 = 0.85$ vol dir que el 85% de la variació total en y es pot explicar per la relació lineal entre x i y. En conseqüència, com més s'acosti a 1, millor s'ajustarà als valors. En aquest cas, la línia passa exactament per cada punt i és capaç de detallar tota la variació. Com més lluny estigui dels punts, pitjor serà l'aproximació. El coeficient de determinació és la relació entre la variabilitat explicada per la regressió i la variabilitat total. Es calcula mitjançant la fórmula següent:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

on $\hat{y}_i$ és l'estimació del valor de y_i. En l'exemple donat:

x	y	$\hat{y} = 5x + 7$	$(\hat{y} - \bar{y})^2$	$(y - \bar{y})^2$
1	10	12	100	144
2	20	17	25	4
3	15	22	0	49
4	40	27	25	324
5	25	32	100	9
$\bar{x} = 3$	$\bar{y} = 22$		$\sum_{i=1}^{5}(\hat{y}_i - \bar{y})^2 = 250$	$\sum_{i=1}^{5}(y_i - \bar{y})^2 = 530$

Per tant, $R^2 = \frac{250}{530} = 0.4717$ i $R = \pm\sqrt{R^2} = \sqrt{0.4717} = 0.6868$, on el signe ve determinat pel pendent de la recta de regressió.

3.4. Regressió lineal amb R-Console o Rstudio

A continuació, es descriu detalladament el procediment per elaborar una regressió lineal utilitzant *R*. Les dades que s'utilitzaran estan desades en el fitxer coches.txt, que es pot descarregar des d'Atenea.

3.4.1. Carregar dades

Tal com s'ha explicat anteriorment, s'importen les dades des de *Rstudio*, *R-Commander* o *R-Console* mitjançant la funció read.table(). S'observa que hi ha dades que no estan disponibles, NaN, i, si es vol fer alguna operació (per exemple, una mitjana), el resultat serà també NaN. Aquest problema es pot evitar agregant un atribut a la funció:

```
Datos = read.table("datos_Sesion3/coches.txt",
                   header=TRUE,sep="\t",na.strings="NA",dec=".")
mean(Datos$MPG)
```

```
[1] NaN
```

```
mean(Datos$MPG, na.rm = TRUE)
```

```
[1] 23.71809
```

Si resulta enutjós haver d'invocar la variable per mitjà del nom de l'estructura, Datos, més el símbol del dòlar, $, més el nom propi de la variable, MPG, es pot utilitzar la funció attach() per vincular totes les variables de l'estructura de dades a la ruta de cerca de *R*, és a dir, les variables es poden invocar només pels seus noms.

```
attach(Datos)
```

```
The following objects are masked from Datos (pos = 3):

    Acceleration, Cylinders, Displacement, Horsepower,
    Mfg, Model, MPG, Origin, Weight

The following objects are masked from Datos (pos = 4):

    Acceleration, Cylinders, Displacement, Horsepower,
    Mfg, Model, MPG, Origin, Weight
```

```
mean(MPG)
```

```
[1] NaN
```

```
mean(MPG, na.rm = TRUE)
```

```
[1] 23.71809
```

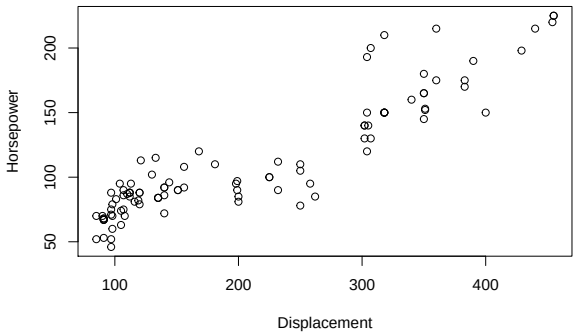
Tips & Tricks!

- La funció attach() permet invocar les variables d'una estructura de dades només pel seu nom.
- A les funcions en *R* s'hi poden agregar atributs.
- L'atribut na.rm=TRUE dona l'ordre que la funció s'executi sense tenir en compte les dades no disponibles.
- na significa *non available* (no disponible), rm, *remove* (suprimeix), i TRUE, veritable. Aquest últim ha d'anar sempre en majúscules perquè, en cas contrari, no és reconegut.

3.4.2. Diagrama de dispersió

Un diagrama de dispersió permet observar clarament si hi ha cap relació entre les variables que estem estudiant. Aquesta dispersió es pot obtenir mitjançant la funció plot(). Per veure'n un exemple, farem un diagrama de dispersió dels cavalls de potència (*Horsepower* (y)) i el desplaçament (*Displacement* (x)).

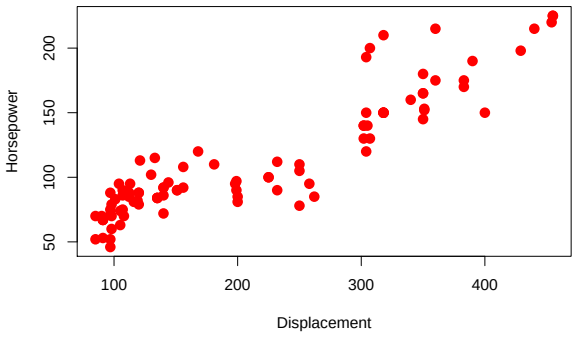
```
plot(Horsepower Displacement)
```



Tips & Tricks!

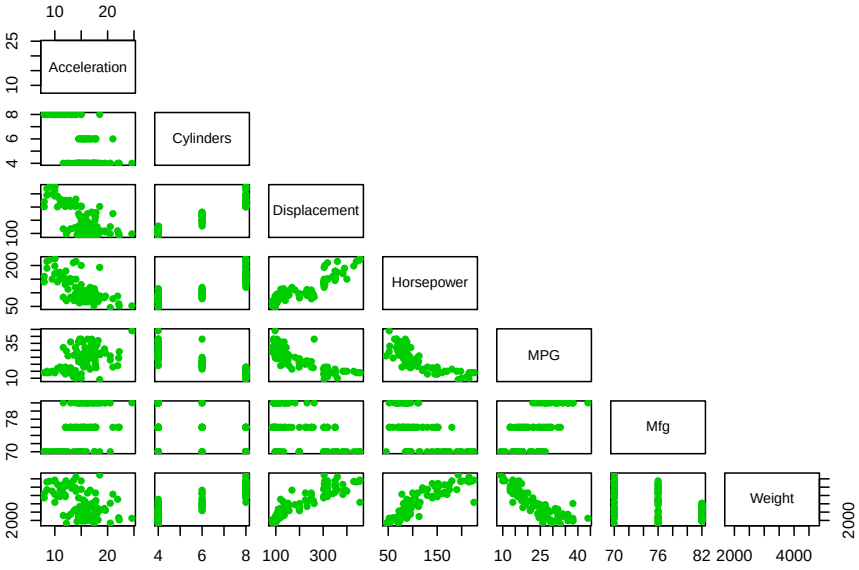
- Amb la funció plot(), es generen gràfics. Hi ha diferents atributs que permeten millorar-ne la visualització, per exemple:
  - Assignant un valor a pch s'escull el símbol per representar els punts (nombres enters entre 0 i 25 o símbols comuns com *, ., o, O, 0, +, –, |, %, #).
  - lwd defineix el gruix de les línies en el gràfic.
  - col asigna un color als punts (en aquest cas, per al color vermell, es pot indicar red o 2).
  - cex determina el factor pel qual es multiplica la mida del punt original.
- plot(x,y) i plot(y~x) generen el mateix gràfic, x en l'eix de les abscisses i y en l'eix de les ordenades del pla XY.

```
plot(Displacement,Horsepower,pch=16,col="red",cex=1.5)
```



Si es desitja veure, de manera general, els diagrames de dispersió entre totes les combinacions de les variables d'una estructura de dades, s'utilitza la funció pairs().

```
pairs(Datos[3:9], upper.panel = NULL, pch = 16, col="green3")
```



Tips & Tricks!

Amb la funció pairs(), es crea una matriu gràfica de la correlació entre totes les variables numèriques del conjunt de dades.

- L'atribut upper.panel = NULL mostra la matriu inferior dels gràfics i n'evita la duplicitat.
- També és possible crear el gràfic en color amb l'opció bg = c() llistant els colors triats separats per comes i entre " " .

3.4.3. Model lineal dels mínims quadrats

El model dels mínims quadrats implementat en *R* consisteix a aproximar una sèrie de valors amb un polinomi del mínim grau possible. Si els valors s'assemblen a una recta, la funció de la línia de regressió serà de la forma $y = mx + b$, de primer grau. Aquest polinomi s'aconsegueix mitjançant la funció lm(y ~ x). Si es vol utilitzar el model després, per exemple per estimar algun valor, representar gràficament els punts i la seva recta, etc., aquest s'ha d'assignar a un objecte amb un nom triat per l'usuari. Per exemple, volem buscar la relació existent entre les variables Displacement i Horsepower.

```
model1=lm(Horsepower Displacement); model1
```

```
Call:
lm(formula = Horas ~ Lotes)

Coefficients:
(Intercept)        Lotes
          7            5
```

Interpretant els resultats obtinguts, es dedueix que la recta de regressió és $y = 34.8870 + 0.3706x$, on x =Displacement i y =Horsepower. L'objecte model1 desa tots els paràmetres del model. Per exemple, una altra manera de conèixer els valors del pendent i la intercepció de la recta de regressió és mitjançant les instruccions següents:

```
model1$coef[1]
```

```
(Intercept)
   34.88703
```

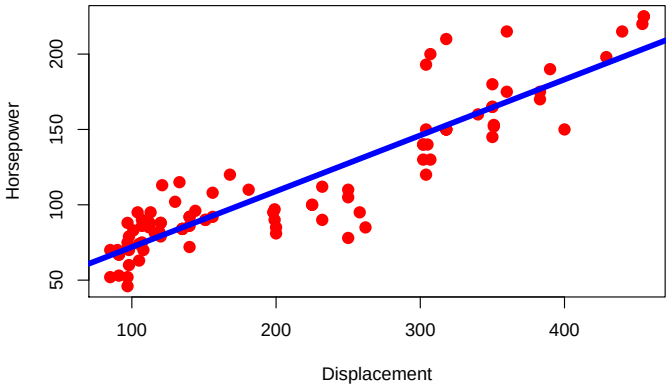
```
model1$coef[2]
```

```
Displacement
   0.3706237
```

3.4.4. Afegir la recta de regressió

Amb la funció abline(), s'afegeix la recta de regressió creada amb la funció lm() al diagrama de dispersió. La primera entrada que cal indicar és el nom de la variable que conté el model, en aquesta ocasió "model1", i els paràmetres addicionals de visualització.

```
plot(Displacement,Horsepower,pch=16,col="red",cex=1.5)
abline(model1,col="blue",lwd=5)
```



3.4.5. Coeficients de determinació (R^2) i de correlació (R)

Una vegada calculat el model lineal, el coeficient de determinació es visualitza utilitzant la funció summary(). Amb aquesta ordre, s'obtenen els paràmetres més importants del model. No obstant això, per a aquest coeficient ens interessa solament multiple R-squared.

```
summary(model1)
```

```
Call:
lm(formula = Horsepower ~ Displacement)

Residuals:
    Min      1Q  Median      3Q     Max
-49.543 -11.425  -0.704   9.604  57.255

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.88703    3.96235   8.805 5.09e-14 ***
Displacement  0.37062    0.01677  22.094  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.63 on 97 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8342, Adjusted R-squared:  0.8325
F-statistic: 488.2 on 1 and 97 DF,  p-value: < 2.2e-16
```

Per tant, el coeficient R quadrat múltiple = 0.8342.
El coeficient de correlació de Pearson, que és l'arrel quadrada del coeficient de determinació, també es pot calcular mitjançant la funció cor().

```
sqrt(0.8342)
```

```
[1] 0.9133455
```

```
cor(Horsepower,Displacement,use="na.or.complete")
```

```
[1] 0.9133645
```

3.4.6. Estimació de valors indeterminats

Una de les aplicacions principals de la regressió és la possibilitat d'estimar el valor de la variable dependent per a un valor determinat de la variable independent. Aquest recurs és molt útil, ja que permet conèixer aproximadament el comportament de les dues variables en situacions sense dades. Per fer l'estimació n'hi ha prou de reemplaçar en la fórmula de la recta els valors del pendent, la intercepció i la variable dependent. Per exemple, si es vol estimar quin serà el valor que manca dels cavalls de potència (observació 77), sabent que aquest cotxe té un valor de desplaçament de 151, n'hi ha prou d'executar:

```
34.8870+0.3706*151
```

```
[1] 90.8476
```

Ara bé, si es vol minimitzar l'error per arrodoniment i, a més, estimar diversos valors, el millor que es pot fer és crear una funció utilitzant els valors directament del model de la manera següent:

```
f1 <- function(x) {model1$coef[1] + model1$coef[2]*(x)}
```

Així, cada vegada que es vulgui fer una estimació només caldrà executar la funció creada:

```
Displacement_pred = 151
Horsepower_pred=  f1 (Displacement_pred);Horsepower_pred
```

```
(Intercept)
   90.85121
```

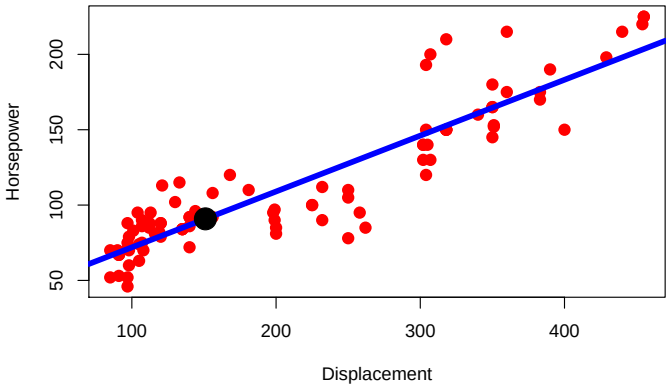
o utilitzar la funció predict().

```
Displacement_pred = 151
Horsepower_pred=predict(model1,data.frame(Displacement=Displacement_pred));
Horsepower_pred
```

```
       1
90.85121
```

Finalment, aquesta estimació es pot representar en el gràfic mitjançant un punt negre (per exemple), amb la instrucció següent:

```
plot(Displacement,Horsepower,pch=16,col="red",cex=1.5)
abline(model1,col="blue",lwd=5)
points(Displacement_pred,Horsepower_pred, col ="black", pch=20, cex
=4)
```



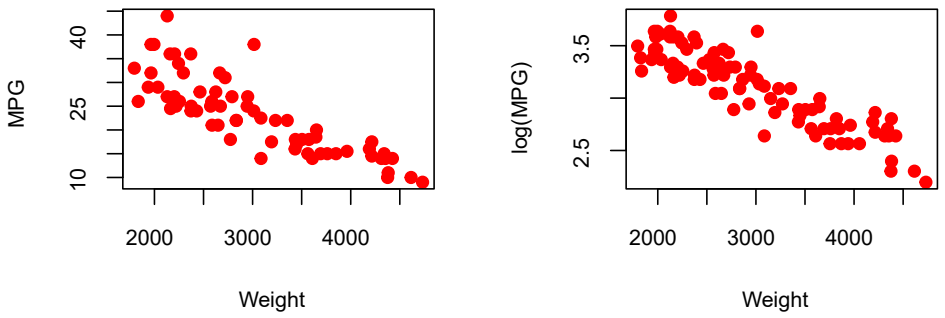
Tips & Tricks!

- lm(y~x) crea un model lineal entre les variables x i y.
- abline() agrega una línia recta al gràfic actual; els paràmetres més usuals són:
  - a=A, b=B defineixen la intercepció (A) i el pendent de la recta (B); també es pot indicar el nom de la variable que conté un model lineal.
  - h=H per definir una línia horitzontal en $y = H$.
  - v=V per definir una línia vertical en $x = V$.
- summary() mostra els resultats del model lineal.
- predict() genera la predicció per a valors nous. Aquests valors s'han d'organitzar en un *data.frame* i la variable independent ha de tenir el mateix nom que s'hagi utilitzat en la creació del model.
- points() agrega punts a un gràfic existent.

3.4.7. Regressió exponencial

Com ja s'ha vist, si dues variables tenen una relació exponencial, aquesta es pot linealitzar per mitjà de logaritmes. Utilitzant el fitxer coches.txt, la relació entre les variables pes (Weight(y)) i consum (MPG (x)) es pot observar mitjançant el diagrama de dispersió. Es pot intuir que la relació tendeix a ser més semblant a una d'exponencial que a una de lineal.

```
par(mfrow=c(1,2))
plot(Weight,MPG,xlab="Weight", ylab="MPG",col="red",pch=20,cex=1.5)
plot(Weight,log(MPG),xlab="Weight", ylab="log(MPG)",col="red",
pch=20,cex=1.5)
```




```
model2=lm(log(MPG) Weight); summary(model2)
```

```
Call:
lm(formula = log(MPG) ~ Weight)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41639 -0.13593  0.00205  0.11741  0.55351

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.300e+00  6.447e-02   66.69   <2e-16 ***
Weight      -4.033e-04  2.101e-05  -19.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1624 on 92 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.8002, Adjusted R-squared:  0.798
F-statistic: 368.4 on 1 and 92 DF,  p-value: < 2.2e-16
```

D'aquestes dades, obtenim que $\ln(\text{MPG}) = 4.3 - 0.0004033 \times \text{Weight}$, o que $\text{MPG} = e^{4.3} \cdot e^{-0.0004033 \times \text{Weight}}$. A més a més, R quadrat múltiple $= 0.8002$ i R quadrat ajustat $= 0.798$.

Finalment, volem estimar els valors de la variable MPG que no han estat registrats (NaN), és a dir, les observacions: 11, 12, 13, 14, 15 i 18. Per fer-ho, utilitzem l'equació de la regressió i els valors del pes de les observacions: 11, 12, 13, 14, 15 i 18.

```
obs = which(MPG=="NaN"); obs
```

```
[1] 11 12 13 14 15 18
```

```
Weight_pred = Weight[obs]
Log_MPG_pred = predict(model2,data.frame(Weight=Weight_pred));
Log_MPG_pred
```

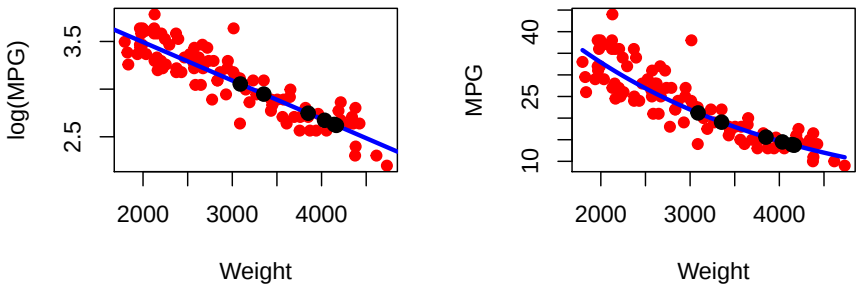
```
       1        2        3        4        5        6
3.053833 2.629596 2.673149 2.619917 2.747350 2.947774
```

```
MPG_pred = exp(Log_MPG_pred); MPG_pred
```

```
       1        2        3        4        5        6
21.19644 13.86816 14.48551 13.73459 15.60123 19.06347
```

Representant gràficament les dades existents, el resultat de la regressió i l'estimació dels valors no registrats, tenim:

```
par(mfrow=c(1,2))
plot(Weight,log(MPG),xlab="Weight", ylab="log(MPG)",
     col="red",pch=20,cex=1.5)
abline(model2,col="blue",lwd=3)
points(Weight_pred,Log_MPG_pred, col = "black", pch= 20, cex = 2)
plot(Weight,MPG,xlab="Weight", ylab="MPG",col="red",pch=20,cex=1.5)
curve(exp(4.3)*exp(-0.0004033*x),add=T,col="blue",lwd=3)
points(Weight_pred,MPG_pred, col = "black", pch= 20, cex = 2)
```

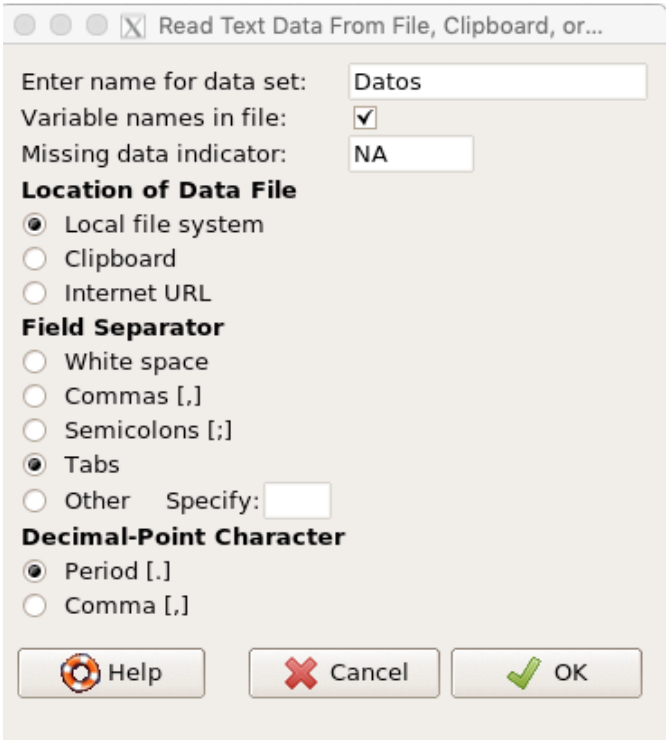


3.5. Regressió lineal amb R-Commander

Igual com a l'apartat anterior, s'utilitza el fitxer "coches.txt", disponible a Atenea.

3.5.1. Carregar dades

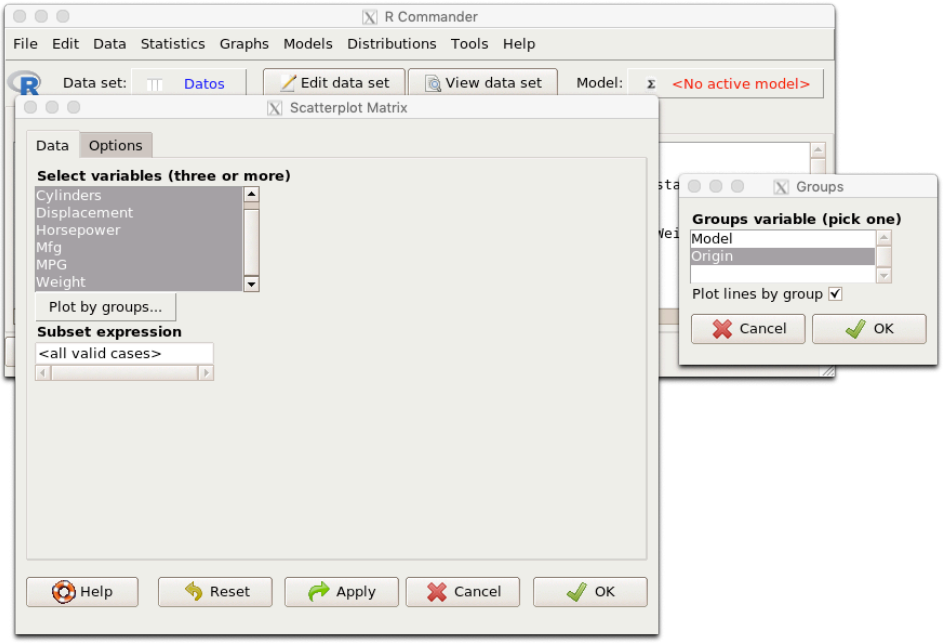
Es carreguen les dades del fitxer, que té les característiques següents:



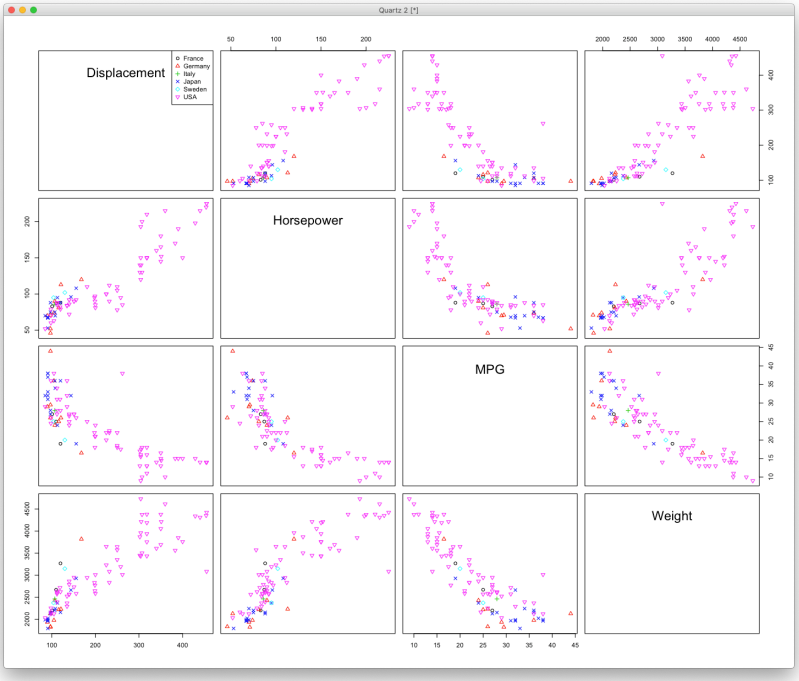
Es recomana visualitzar les dades per verificar que s'han importat correctament.

3.5.2. Diagrames de dispersió

Per visualitzar les relacions existents entre diferents variables en forma d'una matriu de gràfics, se selecciona *Gráficos* > *Matriz de diagramas de dispersión* a la barra de menús. Apareix la finestra següent, en què es poden triar les variables del futur gràfic i, si es vol, discriminar-les per grups d'acord amb les variables qualitatives existents. També es poden canviar diverses opcions de visualització per mitjà de la pestanya *Opciones*. Per exemple, farem una matriu amb les variables: Displacement, Horsepower, MPG, Weight, però agrupades d'acord amb l'origen.

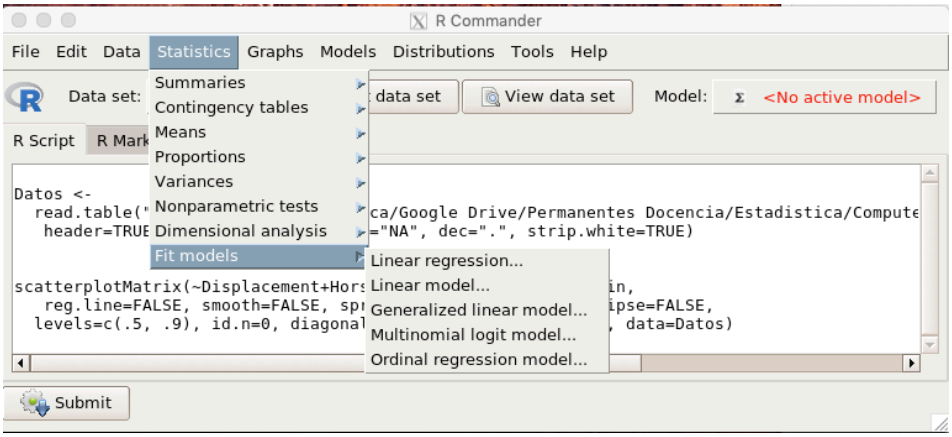


La figura resultant és:

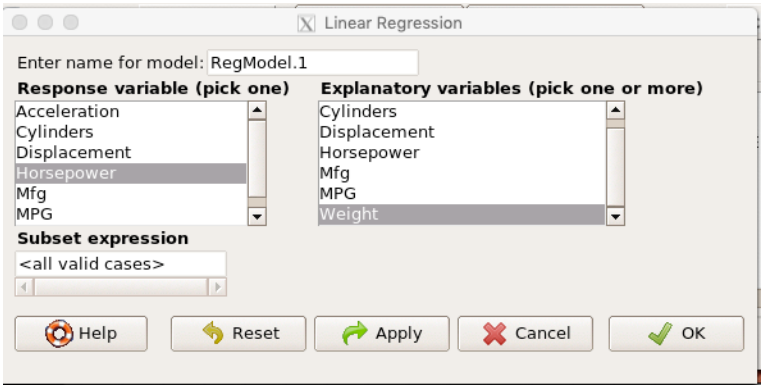


3.5.3. Model lineal dels mínims quadrats

Ara busquem la recta de regressió entre els cavalls de potència (Horsepower (y)) i el pes (Weight (x)) per poder estimar el valor que manca de la variable Horsepower (observació 77) utilitzant el valor del pes d'aquesta observació. Se selecciona *Estadística* > *Ajuste de modelos* > *Regresión lineal* a la barra de menús.



Es defineix el nom del model, se selecciona Horsepower com a variable independent i Weight com a variable dependent.



Quan fem clic a OK, s'observa que a la finestra principal, sota el menú, la casella *Model* ara es mostra activada amb el nom que li hem donat en crear el model lineal, en el nostre cas RegModel.1. D'altra banda, a la pestanya de *R Script* apareix la línia d'instruccions que es poden executar des de *R-Console* o *Rstudio*.

```
RegModel.1 <- lm(Horsepower Weight, data=Datos) summary(RegModel.1)
```

```
Call:
lm(formula = Horsepower ~ Weight, data = Datos)

Residuals:
    Min      1Q  Median      3Q     Max
-61.610 -12.510   0.872   9.627 109.312

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -35.592850   8.656252  -4.112 8.22e-05 ***
Weight        0.049022   0.002776  17.657  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.29 on 97 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7627, Adjusted R-squared:  0.7603
F-statistic: 311.8 on 1 and 97 DF,  p-value: < 2.2e-16
```

3.5.4. Afegir la recta de regressió

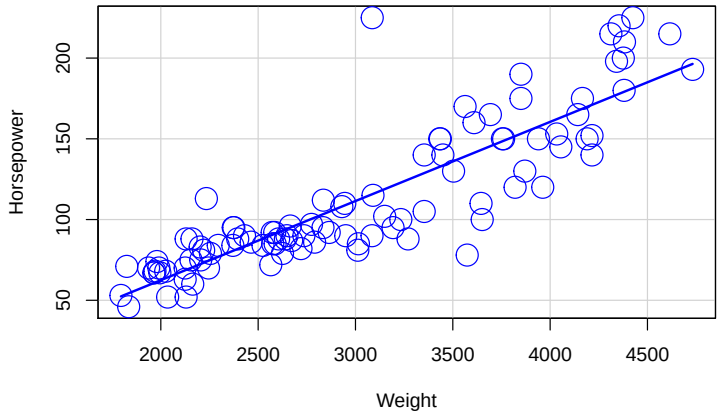
Per afegir la recta de regressió, només cal fer el gràfic de dispersió entre les variables i seleccionar, dins la pestanya Opcions, la recta de regressió. També es poden modificar altres opcions de visualització.

El resultat a la finestra de sortida i a la figura és el següent:

```
library(car)
```

```
Loading required package: carData
```

```
scatterplot(Horsepower Weight, reg.line=lm, smooth=FALSE,
spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,
levels=c(.5, .9), cex=2.5, data=Datos)
```



3.5.5. Estimació de valors indeterminats

Finalment, l'estimació dels valors indeterminats es fa de la mateixa manera que en *R-Console* o *Rstudio*.

```
obs = which(Horsepower=="NaN"); obs
```

```
[1] 77
```

```
Weight_pred = Weight[obs]
Horsepower_pred = predict(RegModel.1,data.frame(Weight=Weight_pred));
Horsepower_pred
```

```
       1
113.1877
```

3.6. Exercicis

1. Tenim les altures següents (en cm) d'un conjunt de persones:

   78, 181, 168, 183, 164, 181, 174, 176, 174, 176, 181, 168, 164, 174, 171

   A aquest mateix conjunt de persones, se les pesa i s'obtenen els pesos següents:

   82, 89, 68, 91, 65, 80, 79, 81, 80, 79, 82, 69, 67, 80, 78

Feu un estudi de regressió lineal, calculant la recta de regressió i el coeficient de determinació, i traceu el diagrama de dispersió. Què podeu deduir a partir del valor de R?

2. Utilitzant les dades del fitxer reg.txt, analitzeu si és apropiat elaborar models de regressió lineal entre les variables x1-y1, x2-y2, x3-y3, x4-y4. Per a cada parell de variables, podeu seguir el mètode següent:

   a) Diagrama de dispersió. Us sembla apropiat un model lineal per descriure cada conjunt de dades?

   b) Ajust d'un model lineal. Comenteu els resultats.

   c) En cas que un model lineal no sigui apropiat, què es podria fer per ajustar un model que pugui predir la variable y en funció de la variable x?

3. Volem estudiar la resistència d'unes peces de ciment en funció de la seva edat en dies. Utilitzant les dades contingudes en el fitxer "cemento.txt", proposeu un model que relacioni la resistència amb el temps d'assecatge. Quina resistència tindran al cap de 5 dies? I al cap de 50? Utilitzeu el coeficient de determinació per justificar aquests valors.

4. Volem estudiar l'evolució del nivell màxim anual del mar (en cm) a Venècia. Les dades de què disposem corresponen als anys 1931-1981 i estan contingudes en el fitxer venecia.txt (dades reals, publicades a Smith R. L., "Extreme value theory based on the r largest annual events", *Journal of Hydrology*, 86 (1986)). Feu un estudi de regressió i comenteu l'evolució del màxim anual del nivell de la mar a Venècia.

5. Es vol estudiar l'evolució de la producció mundial de petroli del 1880 al 1973. Les dades es troben al fitxer petroleo.txt.

6. La hidròlisi d'un cert èster té lloc en un mitjà àcid segons un procés cinètic de primer ordre. Partint d'una concentració inicial de 30 mm de l'èster, s'han mesurat concentracions a diferents temps i s'han obtingut els resultats registrats en el fitxer ester.txt. Quina estimeu que ha estat la concentració al cap de 70 segons de començar el procés? Expliqueu els resultats.

7. Estudiant els incendis forestals, deduïm que pot existir una relació entre la quantitat de pluja caiguda durant els mesos d'estiu (en mm) i el nombre d'incendis declarats. Recopilem la informació dels últims deu anys i obtenim les dades següents:

Pluja	Incendis
97	521
27	863
93	712
175	163
38	138
192	811
28	534
182	442
61	963
77	313

A la vista d'aquestes dades, podríem fer una previsió de quin serà el nombre aproximat de focs que es declararan amb una pluja de 120 mm? I de 10 mm? Expliqueu les conclusions.

4
Variables aleatòries discretes i distribucions de probabilitat

4.1. Introducció i objectius

Suposem que, en una inspecció de fabricació, tenim un lot de 100 peces i una d'aquestes està contaminada. Si s'inspecciona una peça d'aquest lot, quina probabilitat hi ha que aquesta peça estigui contaminada? Si se n'inspeccionen dues, quina probabilitat hi ha que cap estigui contaminada? Quina probabilitat hi ha que s'hagi d'inspeccionar cinc peces fins a trobar la contaminada? El fet d'inspeccionar una peça, dues o unes quantes fins a trobar la contaminada es denomina *experiment aleatori*. El resultat de l'experiment s'assigna a una variable. Si els resultats possibles no poden prendre valors dins d'un mínim conjunt numerable, la variable es denomina *variable aleatòria discreta*. Si un experiment aleatori es repeteix diverses vegades, s'espera que durant tot el temps els resultats estiguin condicionats per les seves probabilitats. Si aquestes probabilitats segueixen un comportament específic, l'experiment es pot classificar dins d'uns certs models de distribució.

Aquesta sessió se centra a conèixer com es representen les probabilitats d'un experiment aleatori de variable aleatòria discreta, simular repeticions d'un experiment i comparar els resultats amb les probabilitats assignades, i, finalment, descriure els models més freqüents utilitzant els mateixos termes emprats per descriure les dades recollides (esperança i variància). En finalitzar aquesta sessió, l'estudiant ha de ser capaç de:

- Representar gràficament una distribució de variable aleatòria discreta usant *R*.
- Simular la repetició de diferents experiments aleatoris discrets i comparar el resultat d'aquests experiments amb les probabilitats estudiades prèviament.
- Calcular i interpretar el valor esperat i la variància d'una variable aleatòria discreta.
- Reconèixer i aplicar correctament les distribucions de probabilitat discretes més comunes en enginyeria.

4.2. Variables aleatòries discretes (VAD)

Normalment, expressem el resultat d'un experiment aleatori amb un simple número, però això no sempre és possible, com en els casos de llançar una moneda, atrapar una pilota, etc. En aquests casos, no és raonable suggerir que es faci una anàlisi quantitativa. Així doncs, cal assignar un nombre real a cadascun dels successos elementals de l'espai mostral. Una variable aleatòria X es pot definir com la funció que transforma els resultats de l'espai mostral Ω en punts sobre la recta real $\mathbb{R}$. En altres paraules, és una funció el domini de la qual és Ω i el rang $\mathbb{R}$:

$$\mathbf{X} : \Omega \to \mathbb{R}.$$

Una variable aleatòria és una variable aleatòria discreta (VAD o DRV, per les seves inicials en anglès) si el conjunt dels seus possibles resultats és comptable. Això és, si l'espai mostral conté un nombre finit de possibilitats o una seqüència inacabada amb tants elements com nombres enters hi ha. Dit amb més rigor, es determina una VAD com la variable que hi ha entre dos valors observables, en què hi ha almenys un valor no observable.

Però una variable aleatòria el conjunt de valors possibles de la qual és un interval complet de nombres és no discreta, és a dir, si l'espai mostral conté un nombre infinit de possibilitats igual al nombre de punts en un segment de línia, es denomina *variable aleatòria contínua*.

En la majoria de problemes pràctics, les variables aleatòries contínues representen les dades mesurades, com ara els pesos, les temperatures, la distància o els períodes de vida, mentre que les variables aleatòries discretes representen dades comptables, com el nombre d'elements defectuosos en un mostreig, els objectes o el nombre d'accidents mortals a l'any en un punt concret de l'autopista.

Exemple: Si llancem 3 monedes, o una mateixa moneda 3 vegades (experiment aleatori), podem recollir el resultat com un trio; per exemple, si C és "cara" i $+$ és "creu", el conjunt

$$\Omega = \{(CCC), (CC+), (C+C), (+CC), (++C), (+C+), (C++), (+++)\}$$

conté tots els resultats possibles. Aquest espai mostral no està representat per nombres. No obstant això, es poden definir diferents variables aleatòries depenent del que es vulgui analitzar. Per exemple, les variables aleatòries X_1, X_2 i X_3 es poden definir tal com es veu a la taula següent, on el valor de X_1 indica el nombre de "cares", X_2 pren com a valor 1 si s'ha obtingut exactament una "cara" i X_3 pren com a valor 1 si l'últim llançament és "cara" i 0 la resta de casos.

Ω	CCC	$CC+$	$C+C$	$+CC$	$++C$	$+C+$	$C++$	$+++$
X_1	3	2	2	2	1	1	1	0
X_2	0	0	0	0	1	1	1	0
X_3	1	0	1	1	1	0	0	0

4.2.1. Funció de densitat

La funció de densitat (també coneguda com a distribució de probabilitat, funció de probabilitat o funció de massa de probabilitat), associada a una variable aleatòria discreta X definida sobre l'espai mostral Ω, és l'aplicació f que assigna a cada element x_i de $X(\Omega)$ la probabilitat que la variable X prengui aquest valor x_i:

$$f : X(\Omega) \to \mathbb{R},$$
$$x_i \to f(x_i).$$

Aquesta funció té les propietats següents:

- El seu valor sempre és un nombre real positiu entre zero i u: $0 \leq f(x) \leq 1$.
- La probabilitat que X prengui un valor exacte està definida per la seva funció de densitat:
$$P(X = x) = f(x).$$
- La suma total dels seus valors és 1: $\sum_{x_i} f(x_i) = 1$.
- La probabilitat que X prengui un valor en l'interval $[a, b]$ és la suma de les probabilitats en aquest interval: $P(a \leq X \leq b) = \sum_{x_i=a}^{b} f(x_i)$.

Exemple 1: Considerem la variable aleatòria discreta X_1, que representa el "Resultat obtingut" en llançar un dau normal de sis cares. Així, X_1 pot prendre els valors següents: 1, 2, 3, 4, 5, 6. D'acord amb la teoria de la probabilitat, la probabilitat que X_1 sigui 1 és expressada per:

$$P(X_1 = 1) = f_1(1) = \frac{Nm.\text{ esdeveniments favorables}}{Nm.\text{ esdeveniments totals}} = \frac{1}{6}.$$

De la mateixa manera, $f_1(2) = f_1(3) = f_1(4) = f_1(5) = f_1(6) = \frac{1}{6}$. Per tant, la funció de densitat es pot representar per mitjà de la taula següent:

X_1	1	2	3	4	5	6
$f_1(x)$	1/6	1/6	1/6	1/6	1/6	1/6

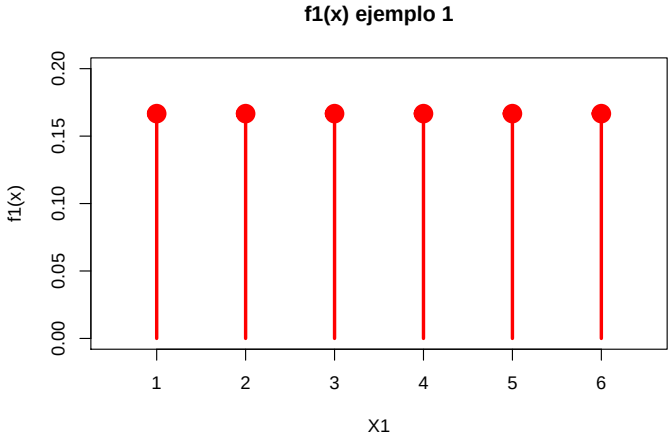
Exemple 2: Ara considerem un dau trucat en què la probabilitat que surti un número és proporcional al valor d'aquest. Això significa que la probabilitat d'obtenir un 1 és 1α, la d'obtenir un 2 és 2α, per a 3 és 3α i el mateix amb les altres possibilitats. Considerem la variable aleatòria discreta X_2, que representa el "Resultat obtingut" en llançar el dau trucat. Partint del fet que $\Sigma p_k = 1$, podem calcular $\alpha = 1/21$; per tant, $f_2(1) = 1/21, f_2(2) = 2/21, f_2(3) = 3/21, f_2(4) = 4/21, f_2(5) = 5/21, f_2(6) = 6/21$. D'aquesta manera, la seva funció de densitat és:

X_2	1	2	3	4	5	6
$f_2(x)$	1/21	2/21	3/21	4/21	5/21	6/21

La funció de densitat es representa gràficament mitjançant línies verticals. Aquestes línies s'estenen des de la línia de base al llarg de l'eix x, on es representen els valors possibles de la variable aleatòria ($X = k$). La probabilitat d'ocurrència de cada valor, $f(x) = P(X = k)$, s'indica mitjançant punts gruixuts al final de la línia. La funció de densitat de l'exemple 1, el dau normal, es representa gràficament de la manera següent:

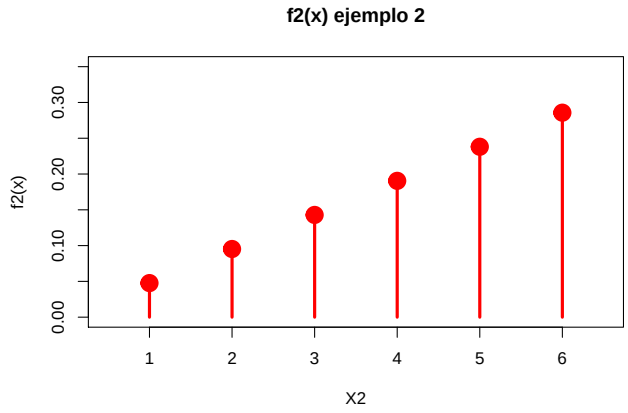
```
rm(list=ls()) # Elimina tots els objectes de l'espai de treball
graphics.off() # Elimina els gràfics creats amb anterioritat
```

```
x1 = 1:6 # Resultats possibles
f1 = rep(1/6,6) # Probabilitat d'ocurrència de cada resultat
                #Creació de les línies verticals
plot(x1, f1, type="h", col="red", lwd=3, main="f1(x) ejemplo 1",
xlab="X1", ylab="f1(x)", xlim=c(0.5,6.5), ylim=c(0,0.20))
# Es creen els punts i es desa el gràfic complet en un objecte
per a usos posteriors
points(x1, f1, col="red", lwd=10); gra.fx.ej1 = recordPlot()
```



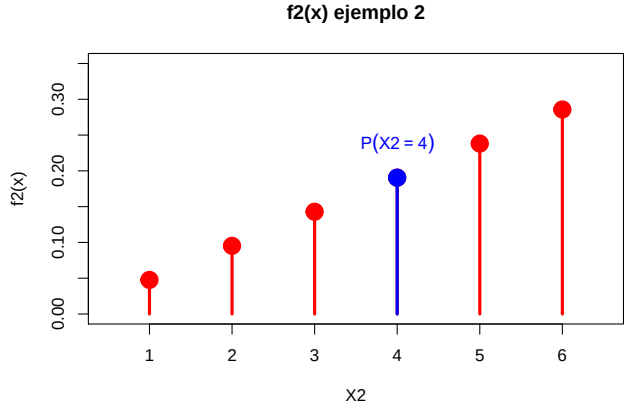
Si considerem l'exemple 2, el dau trucat, la seva funció de densitat és:

```
x2 = 1:6 # Resultats possibles
f2 = x2/21 # Probabilitat d'ocurrència de cada resultat
           # Creació de les línies verticals
plot(x2, f2, type="h", col="red", lwd=3, main="f2(x) ejemplo 2",
xlab="X2", ylab="f2(x)", xlim=c(0.5,6.5), ylim=c(0,0.35))
# Es creen els punts i es desa el gràfic complet en un objecte
per a usos posteriors
points(x2, f2, col="red", lwd=10); gra.fx.ej2 = recordPlot()
```



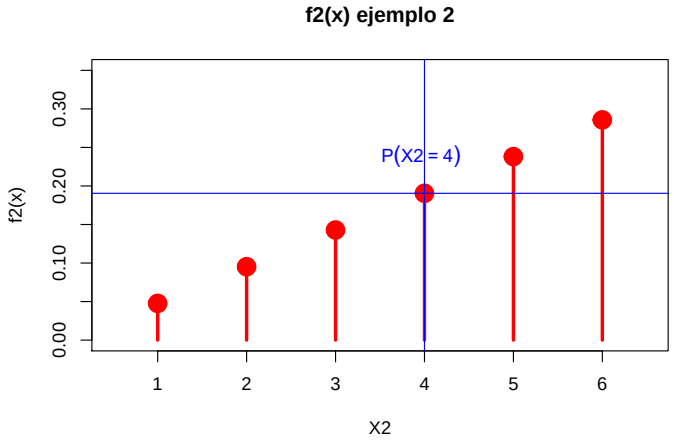
Finalment, si es vol destacar en el gràfic una probabilitat específica, per exemple $P(X = 4)$, es pot optar per canviar el color de la línia i el punt en $X = 4$:

```
gra.fx.ej2 # Es recupera el gràfic creat anteriorment
lines(4, f2[4], type="h", col="blue", lwd=3)
points(4, f2[4], col="blue", lwd=10)
text(4, f2[4]+0.02, expression(P(X2 == 4)), pos=3, col="blue")
```



o, per agregar una línia vertical en $x = 4$ i una d'horitzontal en $f_2(4)$ amb la funció abline():

```
gra.fx.ej2 # Es recupera el gràfic creat anteriorment
abline(v=4, col="blue")
abline(h=f2[4], col="blue")
text(4, f2[4]+0.02, expression(P(X2 == 4)), pos=3, col="blue")
```



4.2.2. Funció de distribució

La funció de distribució o funció de probabilitat acumulada associada a una variable aleatòria discreta X, definida sobre un espai mostral, és l'aplicació F que assigna a cada element x_i de $X(\Omega)$ la probabilitat que la variable X prengui qualsevol valor menor o igual que x_i:

$$F : X(\Omega) \to \mathbb{R},$$
$$x_i \to F(x_i) = P(X \leq x_i) = \sum_{x_i \leq x} P(X = x_i).$$

Per tant:

$$P(a < X \leq b) = F(b) - F(a), \qquad \forall a \leq b.$$

D'aquesta manera, la funció de densitat i la funció de distribució de l'exemple 1 són:

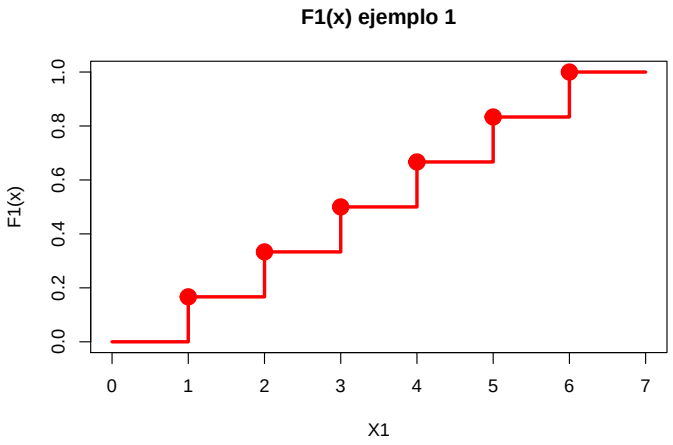
X_1	1	2	3	4	5	6
$f_1(x)$	1/6	1/6	1/6	1/6	1/6	1/6
$F_1(x)$	1/6	2/6	3/6	4/6	5/6	1

I per a l'exemple 2 són:

X_2	1	2	3	4	5	6
$f_2(x)$	1/21	2/21	3/21	4/21	5/21	6/21
$F_2(x)$	1/21	3/21	6/21	10/21	15/21	1

La funció de distribució es representa gràficament mitjançant línies horitzontals esglaonades, que indiquen la probabilitat $F(x) = P(X \leq x)$, on x pot ser qualsevol valor real. Els punts gruixuts al començament de cada línia horitzontal indiquen el valor que pren la funció per als valors de la variable aleatòria X. La funció de distribució de l'exemple 1, el dau normal, es representa gràficament de la manera següent:

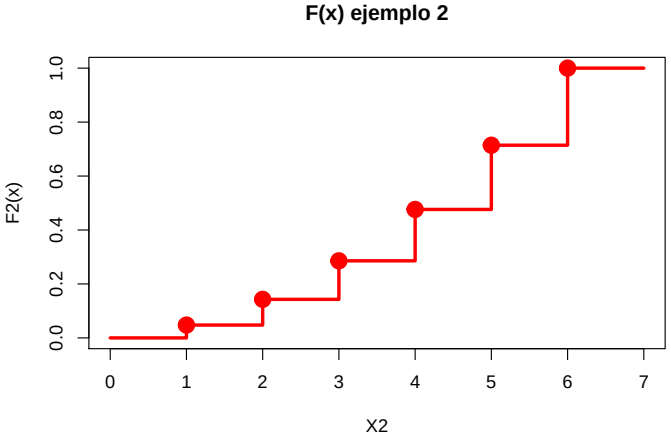
```
F1 = cumsum(f1) # Es genera un vector amb la suma acumulada
plot(c(0,x1,7), c(0,F1,1), type="s", col="red", lwd=3,
main="F1(x) ejemplo 1", xlab="X1", ylab="F1(x)")
points(x1, F1, col="red", lwd=8); gra.Fx.ej1 = recordPlot()
```



Noteu que hem agregat els elements 0 i 7 en l'eix x i les seves imatges, $F_1(0) = 0$ i $F_1(7) = 1$, per completar el gràfic.

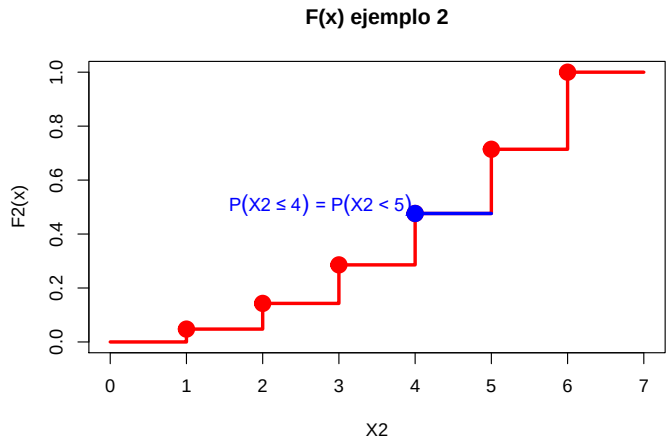
Si considerem l'exemple 2, el dau trucat, la seva funció de densitat és:

```
F2 = cumsum(f2) # Es genera un vector amb la suma acumulada
plot(c(0,x2,7), c(0,F2,1), type="s", col="red", lwd=3,
main="F(x) ejemplo 2", xlab="X2",ylab="F2(x)")
points(x2, F2, col="red", lwd=8); gra.Fx.ej2 = recordPlot()
```



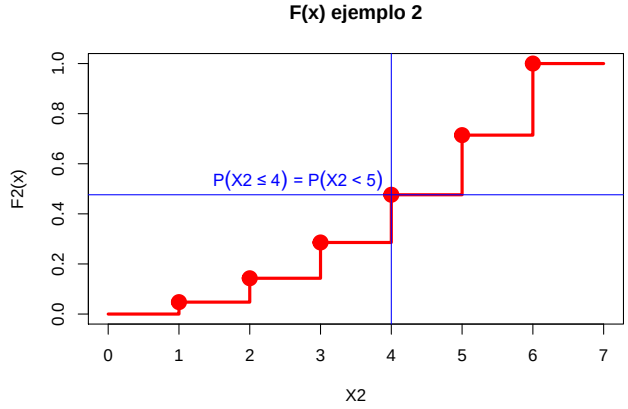
Finalment, si es vol destacar en el gràfic una probabilitat específica, per exemple $P(X \leq 4)$, es pot optar per canviar el color de la línia i el punt en $X = 4$.

```
gra.Fx.ej2
lines(c(4,5), rep(F2[4],2), type="s", col="blue", lwd=3)
points(4, F2[4], col="blue", lwd=8)
text(4, F2[4]+0.02,
expression(P(X2 <= 4) == P(X2<5)), pos=2, col="blue")
```



o, per agregar una línia vertical en $x = 4$ i una d'horitzontal en $F(4)$:

```
gra.Fx.ej2
abline(v=4, col="blue")
abline(h=F2[4], col="blue")
text(4, F2[4]+0.02,
expression(P(X2 <= 4) == P(X2<5)), pos=2, col="blue")
```



Tips & Tricks!

- rm(list = ls()) elimina tots els objectes (variables, dades, funcions) carregats en l'espai de treball.
- graphics.off() elimina o neteja la finestra de gràfics.
- plot() representa gràficament les dades especificades eliminant els gràfics que s'hagin creat prèviament. Les opcions type, col, lwd, xlim, ylim, main, xlab i ylab configuren el tipus de gràfic, el color de les línies, el gruix de la línia, els límits, el títol i les etiquetes dels eixos x i y, respectivament. Vegeu l'ajuda de *R* per saber-ne més detalls.
- El paràmetre type= especifica el tipus de gràfic desitjat: p per a punts, l per a línies, b per a punts i línies, c per a punts buits units per línies, o per a punts i línies superposats, s i S per a esglaons i h per a línies verticals de tipus histograma. Finalment, n no produeix punts ni línies.
- points() agrega punts al gràfic elaborat prèviament.
- lines() agrega línies al gràfic elaborat prèviament. lines(x,y,type=h) crea el mateix gràfic que plot(x,y,type=h), amb la diferència que, amb el primer, el gràfic que estigui prèviament elaborat no s'elimina. Aquesta propietat és útil si es vol superposar dos tipus de gràfics diferents, per exemple amb histograma.
- abline() agrega una línia recta al gràfic actiu; aquesta línia pot ser horitzontal en un valor donat, h=valor, o vertical, v=valor.
- text(x,y,expression()) situa un text donat en la posició (x, y) d'una gràfic elaborat prèviament.
- recordPlot() permet guardar un gràfic en un objecte de *R* per a usos posteriors.

4.2.3. Mesures característiques de les VAD

Aquests paràmetres o característiques importants quantifiquen la tendència central i la variabilitat o dispersió de la variable aleatòria discreta. De fet, conèixer aquestes quantitats, deixant a part la distribució completa, pot donar-nos una idea de la naturalesa del sistema.

Valor esperat

Des del punt de vista de la *freqüencial de probabilitat*, el valor esperat representa la quantitat mitjana que "esperem" com a resultat final d'un experiment aleatori repetit moltes vegades. El valor esperat $E(X)$ és justament la mitjana ponderada d'una variable aleatòria discreta X:

$$\mu = E[X] = \sum_{i=1}^{k} x_i f(x_i).$$

D'aquesta manera, el valor esperat en l'exemple 1 és:

$$\mu_{X_1} = E[X_1] = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = 3.5.$$

I en l'exemple 2 és:

$$\mu_{X_2} = E[X_2] = 1\cdot\frac{1}{21} + 2\cdot\frac{2}{21} + 3\cdot\frac{3}{21} + 4\cdot\frac{4}{21} + 5\cdot\frac{5}{21} + 6\cdot\frac{6}{21} = 4.33.$$

Variància

Aquest paràmetre mesura la dispersió o *scatter* dels possibles valors de X. La variància és la mitjana (esperada) de la distància al quadrat (o desviació) de la mitjana:

$$\sigma^2 = V(X) = Var(X) = E[(X-\mu)^2] = E[X^2] - \mu^2 =$$
$$= \sum_{i=1}^{k}(x_i-\mu)^2 f(x_i) = \sum_{i=1}^{k} x_i^2 f(x_i) - \mu^2.$$

Per tant, la variància de l'exemple 1 és:

$$\sigma^2_{X_1} = V(X_1) = (1-3.5)^2\cdot\frac{1}{6} + (2-3.5)^2\cdot\frac{1}{6} + (3-3.5)^2\cdot\frac{1}{6} + (4-3.5)^2\cdot\frac{1}{6} +$$
$$+ (5-3.5)^2\cdot\frac{1}{6} + (6-3.5)^2\cdot\frac{1}{6} = 2.91667.$$

I la de l'exemple 2 és:

$$\sigma^2_{X_2} = V(X_2) = (1-4.33)^2 \cdot \frac{1}{21} + (2-4.33)^2 \cdot \frac{2}{21} + (3-4.33)^2 \cdot \frac{3}{21} + \ldots$$
$$+ (6-4.33)^2 \cdot \frac{6}{21} = 2.222.$$

Tots aquests resultats es poden validar o verificar si podem repetir l'experiment moltes vegades. Per exemple, si llancem el dau n vegades, obtenim n_1 vegades el número 1; n_2 vegades, el número 2; n_3 vegades, el número 3, i així successivament. La freqüència relativa d'obtenir un valor i és expressada per $f_i = n_i/n$. Si n és molt gran, segurament la freqüència serà $f_i = f(i)$, aproximadament. D'altra banda, la mitjana i la variància de les repeticions seran pròximes a μ i σ^2, respectivament.

4.2.4. Ús de sample() per generar simulacions

Fer un experiment aleatori coneixent-ne l'espai mostral es pot considerar com fer una selecció aleatòria (mostreig) del conjunt de l'espai mostral. Tenint en compte l'anterior, amb *R* es poden simular observacions o experiments d'una VAD per mitjà de la funció sample().

Si el vector de dades x conté els elements del conjunt del qual estem fent el mostreig (espai mostral), llavors la instrucció sample(x) reorganitza el contingut de x en una seqüència aleatòria mentre manté tots els valors numèrics intactes. En altres paraules, es pot considerar que es fan tantes simulacions de l'experiment com elements en l'espai mostral.

La mida de la mostra (o nombre de simulacions) es pot especificar agregant l'atribut size =. Per defecte, el mostreig es fa sense reemplaçament; tanmateix, això es pot canviar amb l'atribut replace =. Hi ha opcions més avançades a l'hora de fer un mostreig, com per exemple especificar la probabilitat de cada element de ser seleccionat; per defecte, cada valor de l'espai mostral és igualment probable. D'aquesta manera, per mitjà de la instrucció sample(x, size=n, replace=T, prob=p), se seleccionen n elements del vector x (amb reemplaçament), les probabilitats del qual s'especifiquen en el vector p.

Considerant l'exemple 1 del llançament d'un dau normal de 6 cares, un resultat pot simular-se així:

```
x = 1:6 sample(x, size=1, prob=)
```

```
[1] 1
```

Cada vegada que es fa l'experiment, és a dir, cada vegada que s'executa la instrucció sample(), el resultat és diferent, a causa de l'aleatorietat.

```
sample(x, size=1, prob=)
```

```
[1] 3
```

Per simular que llencem el dau 100 vegades:

```
fair_die = sample(x, 100, replace=T, prob=); fair_die
```

```
 [1] 4 1 3 5 4 5 5 1 2 1 6 6 4 1 3 5 1 3 5 5 2 3 3 1 4 5 5 1 3 2
   6 4 1 6 5 1 4
[38] 3 6 2 2 4 4 3 1 3 3 3 1 5 3 3 6 5 5 5 5 5 2 5 1 1 1 5 5 1 5
   1 2 3 1 4 2 4
[75] 2 4 6 6 3 1 4 3 2 1 1 3 5 6 6 4 1 2 4 6 6 5 6 5 6 1
```

Com ja s'ha dit, els resultats obtinguts són diferents cada vegada que s'executa la instrucció. Això és així perquè s'han generat nombres pseudoaleatoris a partir d'un nombre donat, denominat *llavor* (*seed*). Si aquesta llavor no s'estableix a priori, *R* utilitza el rellotge del sistema, cosa que explica l'aleatorietat del procés. Si es vol reproduir els mateixos resultats en la simulació, es pot assignar el valor de la llavor mitjançant la instrucció set.seed(k) just abans de la simulació, on *k* és un nombre real. Per exemple:

```
set.seed(1)
fair_die = sample(x, 100, replace=T, prob=); fair_die
```

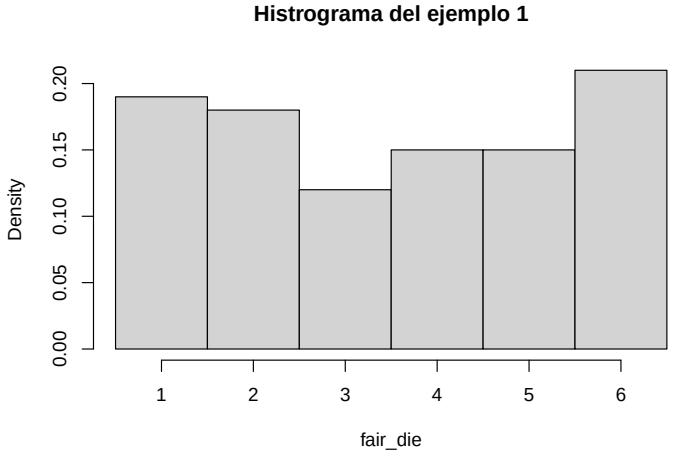
```
 [1] 1 4 1 2 5 3 6 2 3 3 1 5 5 2 6 6 2 1 5 5 1 1 6 5 5 2 2 6 1
   4 1 4 3 6 2 2 6
[38] 4 4 4 2 4 1 6 1 4 1 6 2 3 2 6 6 2 5 2 6 6 6 1 3 3 6 4 6 3
   1 4 5 1 1 6 4 5
[75] 5 4 6 5 4 4 1 5 5 6 1 1 3 6 2 2 3 6 2 4 3 5 2 2 1 3
```

```
set.seed(1)
fair_die = sample(x, 100, replace=T, prob=); fair_die
```

```
 [1] 1 4 1 2 5 3 6 2 3 3 1 5 5 2 6 6 2 1 5 5 1 1 6 5 5 2 2 6 1
   4 1 4 3 6 2 2 6
[38] 4 4 4 2 4 1 6 1 4 1 6 2 3 2 6 6 2 5 2 6 6 6 1 3 3 6 4 6 3
   1 4 5 1 1 6 4 5
[75] 5 4 6 5 4 4 1 5 5 6 1 1 3 6 2 2 3 6 2 4 3 5 2 2 1 3
```

Una manera de resumir els resultats obtinguts de la simulació o el mostreig és per mitjà de la taula de freqüència relativa i la seva representació gràfica per mitjà de l'histograma (o diagrama de barres).

```
h_fair_die = hist(fair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE,
main="Histograma de l'exemple 1")
```

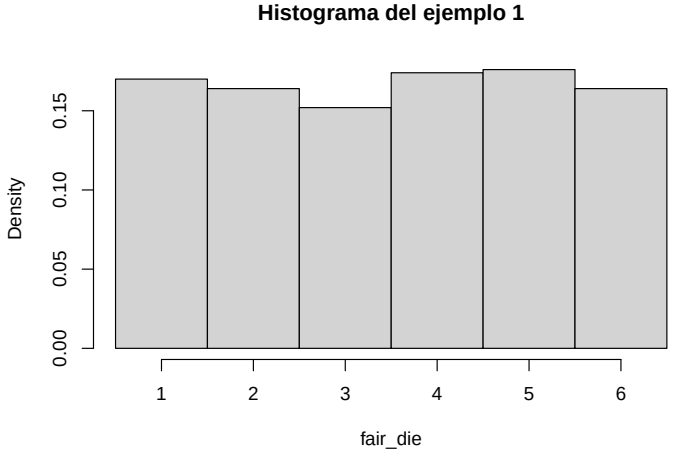



```
data.frame(Result=h_fair_die$mids, Frequency=h_fair_die$density)
```

```
     Result  Frequency
1         1       0.19
2         2       0.18
3         3       0.12
4         4       0.15
5         5       0.15
6         6       0.21
```

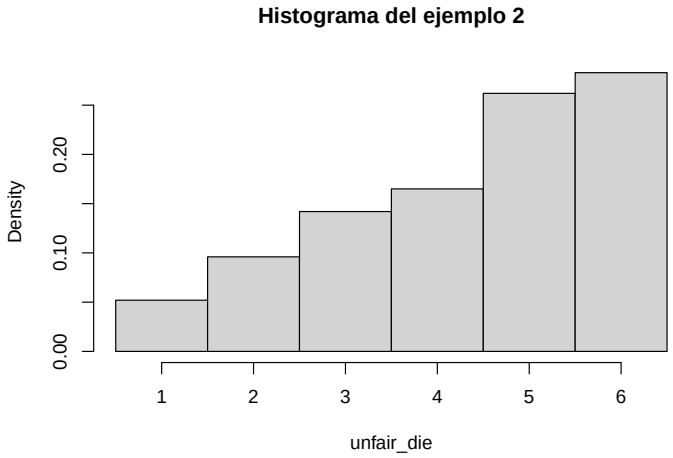
En simular 1000 llançaments i representar gràficament la seva taula de freqüències relativa, s'obté:

```
set.seed(1)
fair_die = sample(x, 1000, replace=T, prob=)
h_fair_die = hist(fair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE,
main="Histograma de l'exemple 1"); gra.hist.ej1 = recordPlot()
```



Si considerem ara l'exemple 2 (el dau trucat), la simulació de 1000 llançaments d'aquest dau i la representació gràfica de la seva freqüència relativa es poden fer de la manera següent:

```
x=1:6; f=x/21;
set.seed(1)
unfair_die = sample(x, 1000, replace=T, prob=f)
h_unfair_die = hist(unfair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE,
main="Histograma de l'exemple 2");
gra.hist.ej2 = recordPlot()
```

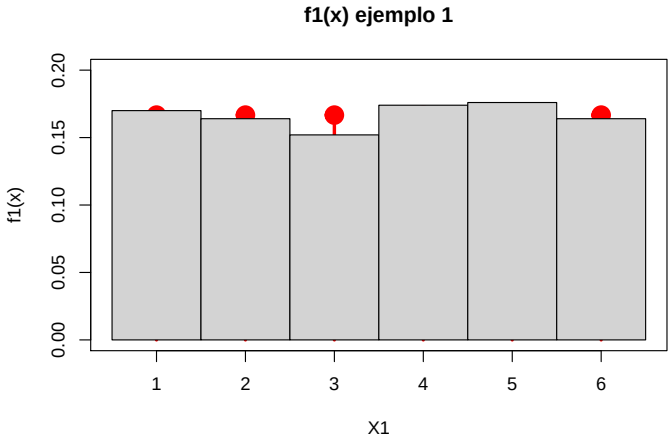


Noteu que els resultats no tenen la mateixa probabilitat d'ocórrer, raó per la qual s'ha de definir prèviament el vector de probabilitats.

4.2.5. Validació dels experiments simulats i la seva distribució de probabilitat

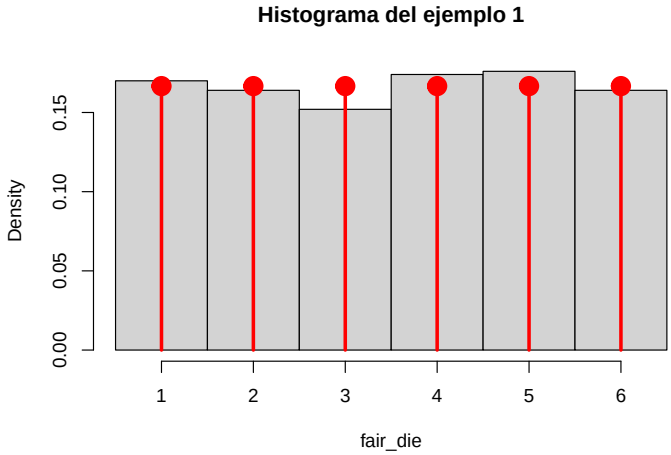
Per comparar la distribució d'una VAD amb les simulacions fetes, el més senzill és crear un gràfic en què se superposin la funció de densitat i l'histograma dels resultats de les simulacions (taula de freqüències). Per defecte, les funcions plot() i hist() esborren la figura que s'hagi carregat prèviament. No obstant això, la funció hist() té una opció que permet elaborar el gràfic sense esborrar l'anterior: add=TRUE. Aquesta opció no està disponible en la funció plot(). D'aquesta manera, podríem pensar que la comparació en l'exemple 1 es pot fer fent primer la funció de densitat i posteriorment l'histograma de la manera següent:

```
gra.fx.ej1
hist(fair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE, add=T)
```



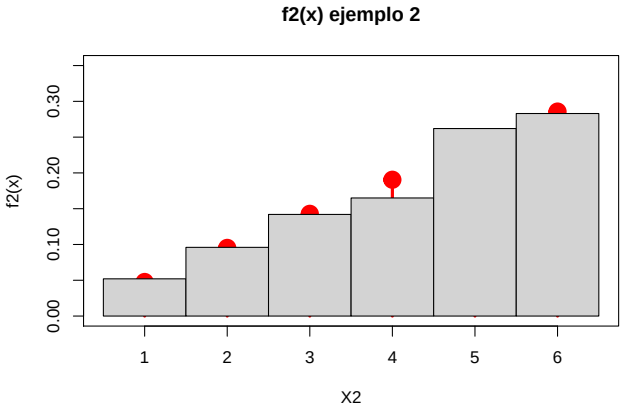
Una altra manera de fer gràfiques en *R* és mitjançant la funció lines(), que agrega línies a un gràfic existent d'una manera molt similar a com ho fa la funció plot(), però sense esborrar els gràfics anteriors. Per tant, es pot fer primer l'histograma i després la funció de densitat.

```
gra.hist.ej1
x = 1:6
f = rep(1/6,6)
lines(x, f, type="h", col="red", lwd=3)
points(x, f, col="red", lwd=10)
```



La comparació de la funció de densitat de l'exemple 2 i dels resultats de simulació és:

```
gra.fx.ej2
hist(unfair_die, breaks=seq(0.5,6.5,by=1), freq=FALSE, add=T)
```



D'altra banda, també es pot comparar la mitjana i la variància dels experiments simulats amb el valor esperat i la variància de la VAD; aquests valors han de ser similars. Recordem que, per a l'exemple 1, el valor esperat és 3.5 i la variància és 2.91667. La mitjana i la variància de les dades simulades de l'exemple són:

```
mean(fair_die)
```

```
[1] 3.514
```

```
var(fair_die)
```

```
[1] 2.936741
```

Per a l'exemple 2, el valor esperat és 4.33 i la variància és 2.222. La mitjana i la variància de les dades simulades en aquest exemple són:

```
mean(unfair_die)
```

```
[1] 4.338
```

```
var(unfair_die)
```

```
[1] 2.276032
```

Tips & Tricks!

- sample(x,n,replace=T,prob=) selecciona una mostra amb reemplaçament de n elements del vector x en què la probabilitat de selecció és la mateixa per a tots els elements. En altres paraules, s'estan simulant n experiments equiprobables de l'espai mostral x.
- sample(x,n,replace=T,prob=p) simula n experiments de l'espai mostral x tenint en compte la probabilitat d'ocurrència definida pel vector p.
- Utilitzeu set.seed() per assegurar-vos que els resultats de qualsevol funció que inclou aleatorietat siguin reproduïbles.

4.3. Les distribucions de probabilitat discretes més habituals

Hi ha moltes situacions pràctiques en la ciència i l'enginyeria en què les distribucions de probabilitat i les seves propietats s'utilitzen per resoldre problemes importants. En algunes d'aquestes situacions, la naturalesa de la distribució i fins i tot una bona estimació de l'estructura de la probabilitat poden ser determinants per a les dades històriques o d'estudi a llarg termini, i fins i tot per a grans quantitats de dades ja previstes. No obstant això, no totes les funcions de probabilitat es deriven de grans quantitats de dades històriques. Hi ha nombroses situacions la naturalesa de les quals suggereix un tipus concret de distribució. Aquestes (també anomenades *distribucions estàndard*) s'utilitzen a tot el món en problemes de la vida real, perquè l'escenari científic que dona lloc a cadascun d'aquests és recognoscible i ocorre en la pràctica de manera general. En aquesta sessió, s'analitzen i apliquen les distribucions de probabilitat discretes més típiques que s'utilitzen en enginyeria.

Distribució binomial

Molts experiments consisteixen en la repetició de l'assaig i per tant s'obtenen dos possibles resultats, que poden marcar-se com a reeixit o fallit (assaig dicotòmic). Si la probabilitat d'èxit (p) és la mateixa en cada assaig i aquests són independents, es denominen *assajos de Bernoulli*. Si una variable discreta aleatòria (VAD) indica el nombre d'èxits en n assajos de Bernoulli, amb una probabilitat d'èxit p, diem que aquesta variable segueix una distribució binomial amb els paràmetres n i p. La seva notació és $X \hookrightarrow B(n, p)$. La seva funció de densitat, la seva mitjana i la seva variància són expressades per:

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{per a} \qquad x = 0, 1, 2, ..., n.$$

$$E(X) = np, \qquad V(X) = np(1-p).$$

Distribució geomètrica

Si una VAD indica el nombre d'assajos de Bernoulli necessaris fins al primer succés, amb una probabilitat d'èxit p, diem que aquesta variable segueix una distribució geomètrica amb un paràmetre p i la seva notació és $X \hookrightarrow G(p)$. La seva funció de densitat, la seva mitjana i la seva variància són expressades per:

$$f(x) = P(X = x) = p(1-p)^{x-1} \qquad \text{per a} \qquad x = 1, 2, ...$$

$$E(X) = \frac{1}{p}, \qquad V(X) = \frac{(1-p)}{p^2}.$$

Distribució binomial negativa

Si una VAD indica el nombre d'assajos de Bernoulli necessaris fins a obtenir r èxits, amb una probabilitat d'èxit p, podem dir que aquesta variable segueix la distribució binomial negativa amb els paràmetres r i p. La seva notació és $X \hookrightarrow NB(r, p)$. La seva funció de densitat, la seva mitjana i la seva variància són expressades per:

$$f(x) = P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad \text{per a} \qquad x = r, r+1, r+2, \ldots$$

$$E(X) = \frac{r}{p}, V(X) = r\frac{(1-p)}{p^2}.$$

Distribució hipergeomètrica

Si una VAD indica el nombre d'èxits en n assajos dependents dicotòmics, amb una probabilitat d'èxit que canvia en cada assaig (població que consisteix en N èxits i $N-k$ fracassos), podem dir que aquesta variable segueix la distribució hipergeomètrica amb paràmetres n, N i k. La seva notació és $X \hookrightarrow H(n, N, k)$. La seva funció de densitat, la seva mitjana i la seva variància són expressades per:

$$f(x) = P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \quad \text{per a} \quad x = max\{0, n+k-N\}, \ldots, min\{k, n\},$$

$$p = \frac{k}{N}, E(X) = np, V(X) = np(1-p)\frac{N-n}{N-1}.$$

Distribució de Poisson

Els experiments que tenen com a resultat el nombre d'esdeveniments que ocorren durant un interval de temps donat o en una regió específica es denominen *assajos de Poisson*. Si una VAD indica el nombre d'assajos de Poisson, amb una freqüència d'ocurrència mitjana λ, podem dir que aquesta variable segueix la distribució de Poisson amb paràmetre λ. La seva notació és $X \hookrightarrow P(\lambda)$ i la seva funció de densitat, la seva mitjana i la seva variància són expressades per:

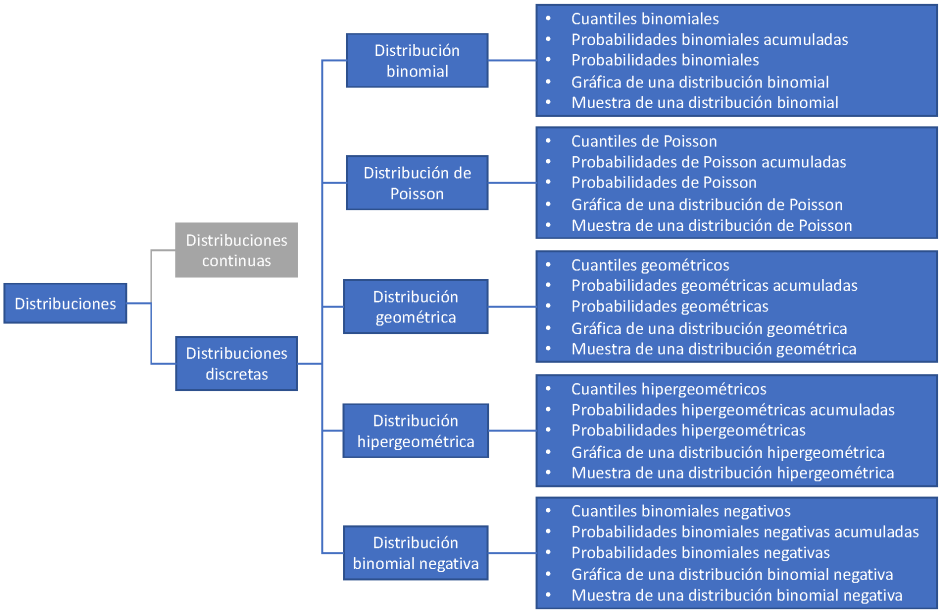
$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{per a} \qquad x = 0, 1, 2, \ldots$$

$$E(X) = \lambda, \qquad V(X) = \lambda.$$

Donada una funció de distribució discreta usada habitualment en enginyeria (binomial, geomètrica, binomial negativa, hipergeomètrica o de Poisson) amb els seus respectius paràmetres, utilitzant *R* i/o *R-Commander* es poden obtenir molt fàcilment la probabilitat d'un esdeveniment elemental, les probabilitats acumulatives, els quantils i els gràfics de distribucions estadístiques estàndard (que poden usar-se, per exemple, com a substituts de les taules estadístiques); a més, es poden generar mostrejos o simulacions d'aquestes distribucions.

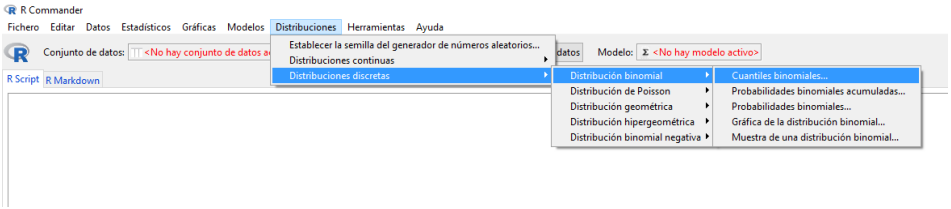
Si es treballa amb *R-Commander*, a la barra superior tenim, a *Distribuciones* > *Distribuciones discretas*, totes les distribucions estudiades. L'arbre del submenú complet per a distribucions de probabilitat discretes es mostra a la figura següent. La major part d'opcions del menú ens porten a diferents quadres de diàleg. Les opcions del menú estan inactives (en gris) si no es poden aplicar al context actual.

Si no es disposa de *R-Commander* o si es prefereix usar *R-Console* o *Rstudio*, o simplement s'està creant un script per a ús posterior, també s'indiquen les instruccions per a cada cas. D'altra banda, per no ser repetitius, en aquesta guia només s'expliquen totes les opcions concernents a la distribució binomial; les altres distribucions tenen les mateixes opcions, amb la diferència que els paràmetres són diferents per a cada distribució. Per exemple, la distribució binomial té com a paràmetres n i p, mentre que la distribució geomètrica només té p i la hipergeomètrica té N, n i k.



Donada una VAD X que segueix una distribució binomial amb els paràmetres n i p, és a dir, $X \hookrightarrow B(n, p)$, es poden calcular les probabilitats binomials (funció de densitat) i

les probabilitats acumulades, elaborar els gràfics de les funcions de densitat i de distribució, calcular els quantils binomials i fer un mostreig o simulació d'experiments binomials, tal com es mostra a la figura.

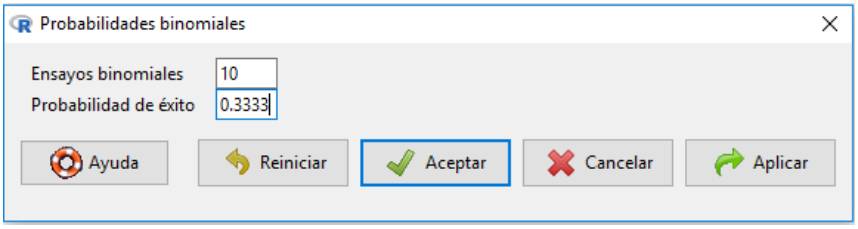


4.3.1. Probabilitats elementals

Calculeu la funció de densitat de la distribució donada, és a dir, la probabilitat que la variable X prengui cadascun dels valors de l'espai mostral, $P(X = x)$.

Exemple: Un terç de la població d'una comunitat té una malaltia concreta. Se'n trien 10 persones de manera aleatòria. Quina és la probabilitat que 4 persones en un mostreig aleatori de 10 persones d'aquesta comunitat pateixin la malaltia?

Si definim la VAD X, que denota el nombre de persones que pateixen la malaltia d'un mostreig aleatori de 10 persones, on la probabilitat d'èxit és 1/3, llavors $X \hookrightarrow B(10, 1/3)$. Per tant, $P(X = x)$ s'obté en introduir els paràmetres següents en el quadre de diàleg que es genera en executar *Distribuciones* > *Distribuciones discretas* > *Distribución binomial* > *Probabilidades binomiales...*



A la finestra d'instruccions, es genera el codi següent amb la seva respectiva resposta en la finestra de resultats:

```
local({  .Table <- data.frame(Probability=dbinom(0:10, size=10,
prob=0.3333))
rownames(.Table) <- 0:10 print(.Table) })
```

```
    Probability
0  1.735020e-02
1  8.673800e-02
2  1.951312e-01
3  2.601359e-01
4  2.275848e-01
5  1.365304e-01
6  5.687914e-02
7  1.624874e-02
8  3.046183e-03
9  3.384140e-04
10 1.691816e-05
```

Per tant, la probabilitat que 4 persones pateixin la malaltia és $P(X = 4) = 0.2275848$. Noteu que els valors de la funció de densitat ($f(x)$) es calculen mitjançant la funció dbinom(x, size=n, prob=p), on x és el vector de valors de X, del qual es vol calcular la probabilitat, n és el nombre total d'experiments i p és la probabilitat d'èxit. D'aquesta manera, si es prefereix usar directament les ordres en *R-Console* o *Rstudio*, es poden calcular només les probabilitats desitjades seguint la instrucció següent:

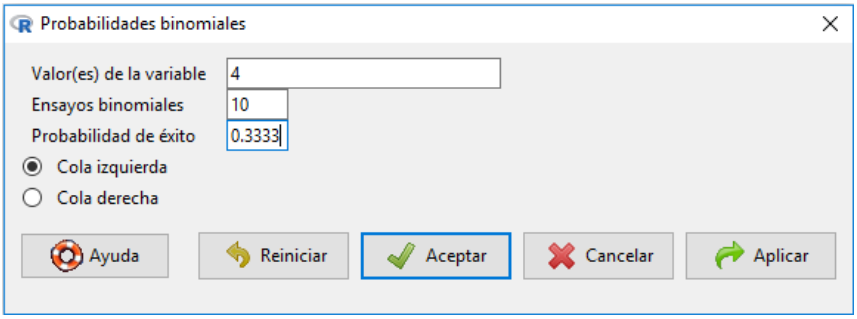
```
dbinom(4, size=10, prob=0.3333)
```

```
[1] 0.2275848
```

4.3.2. Probabilitats acumulades

Calcula la funció de distribució acumulada de la distribució donada: la probabilitat que la variable X sigui com a màxim x, $P(X \leq x)$. A més, $P(X > x)$ també es pot calcular seleccionant l'opció *Cola derecha* en lloc de *Cola izquierda*. Seguint l'exemple anterior, quina la probabilitat hi ha que 4 persones o menys en un mostreig aleatori de 10 persones d'aquesta comunitat pateixin la malaltia?

Per calcular $P(X \leq 4)$, s'introdueixen els paràmetres següents en el seu corresponent quadre de diàleg:



I es genera el codi següent i el seu resultat:

```
pbinom(c(4), size=10, prob=0.3333, lower.tail=TRUE)
```

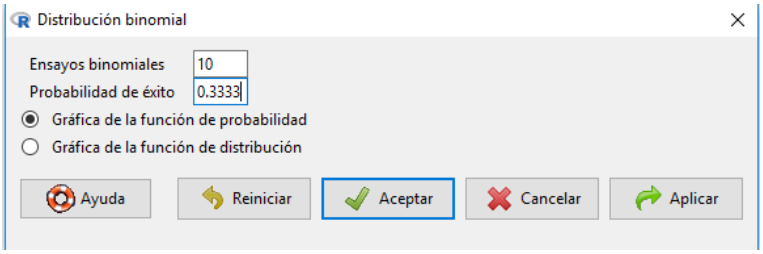
```
[1] 0.7869402
```

Per tant, $P(X \leq 4) = 0.7875542$. Si se selecciona *Cola derecha*, la probabilitat calculada seria: $P(X > 4) = P(X \geq 5) = 1 - P(X \leq 4)$.

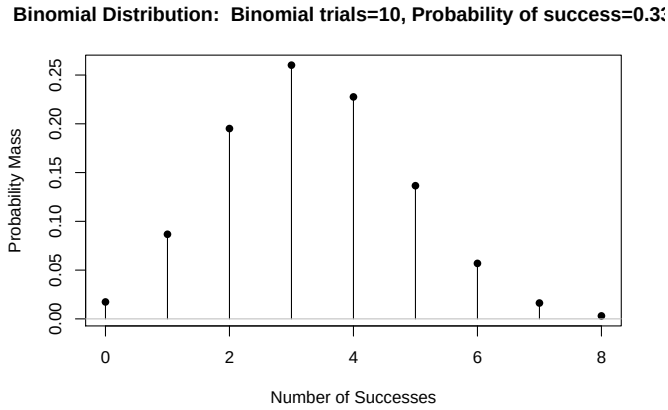
Noteu que les probabilitats acumulades es calculen mitjançant la funció pbinom(x, size=n, prob=p, lower.tail=), on, a més de x, size i prob, s'ha de definir la cua lower.tail, TRUE per a l'esquerra i FALSE per a la dreta.

4.3.3. Gràfic d'una distribució

Aquesta opció permet generar la representació gràfica de la funció de densitat o la funció de distribució de la distribució donada. La funció de densitat de l'exemple es genera de la manera següent:

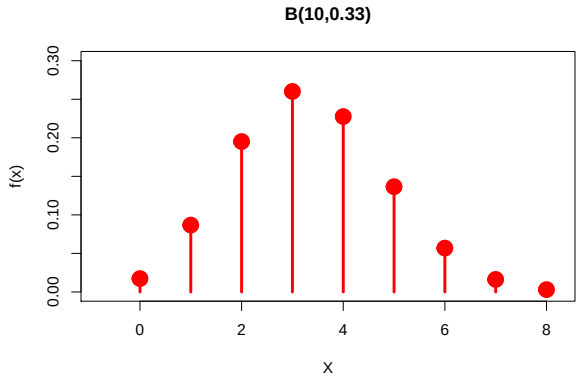



```
local({  .x <- 0:8 plotDistr(.x, dbinom(.x, size=10, prob=0.3333
), xlab="Number of Successes", ylab="Probability Mass",
main="Binomial Distribution: Binomial trials=10,
      Probability of success=0.3333",
discrete=TRUE) })
```



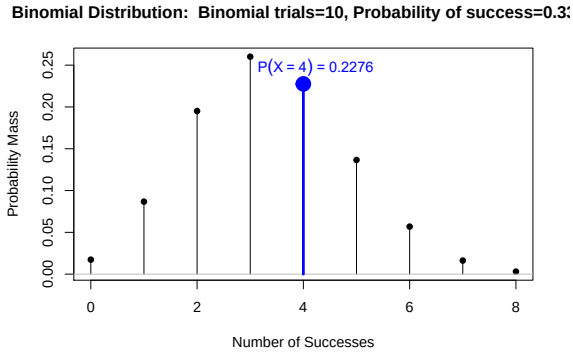
Per fer el gràfic directament des de *R-Console* o *Rstudio*, es poden utilitzar les instruccions explicades al principi de la guia de la manera següent:

```
x = 0:8 # Resultats possibles
f = dbinom(x, size=10, prob=0.3333)
#Probabilitat d'ocurrència de cada resultat
plot(x, f, type="h", col="red", lwd=3, main="B(10,0.33)",
xlab="X", ylab="f(x)", xlim=c(-0.8,8.2), ylim=c(0,0.30))
#Línies verticals
points(x, f, col="red", lwd=10); gra.fx.binom = recordPlot() #Punts
```

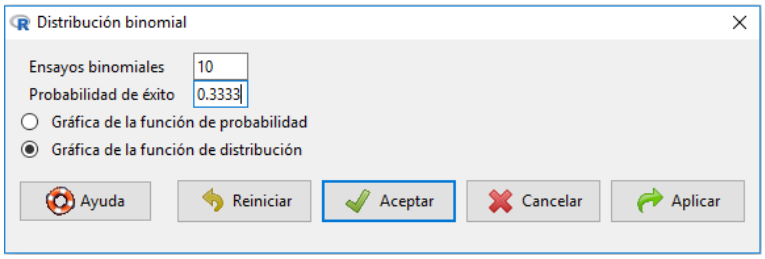


Independentment de com s'hagi generat el gràfic de la funció de densitat, és possible destacar alguna probabilitat específica. Continuant amb l'exemple, la probabilitat que 4 persones en un mostreig aleatori de 10 persones pateixin la malaltia es pot representar agregant el codi següent:

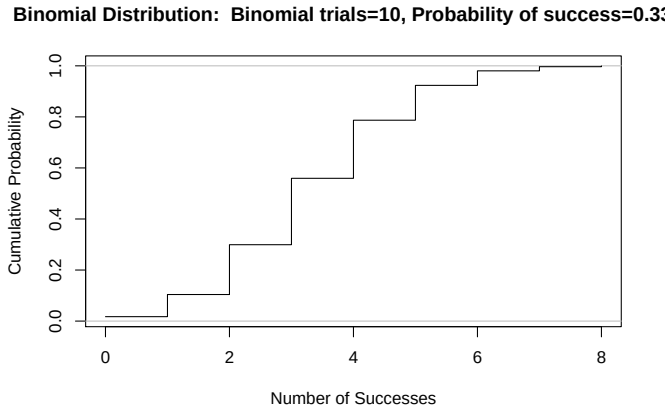
```
x = 4 f_4 = dbinom(4, size=10, prob=0.3333) # Càlcul de P(X=4)=f(4)
lines(x, f_4, type="h", col="blue", lwd=3) # Agrega la línia en X=4
points(x, f_4, col="blue", lwd=10) # Agrega el punt en (4,f(4))
text(x, f_4,expression(P(X==4)), pos=3, col="blue")
```



D'altra banda, el gràfic de la funció de distribució es pot generar introduint els paràmetres següents en el quadre de diàleg:

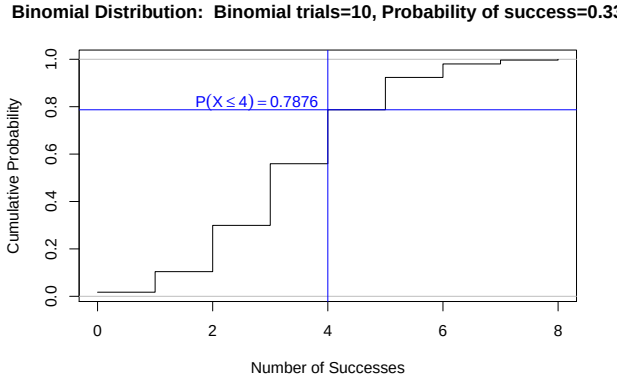



```
local({  .x <- 0:8 plotDistr(.x, pbinom(.x, size=10, prob=0.3333),
xlab="Number of Successes", ylab=Çumulative Probability",
main="Binomial Distribution: Binomial trials=10,
      Probability of success=0.3333",
discrete=TRUE, cdf=TRUE) })
```



De la mateixa manera, es pot representar la probabilitat que 4 persones o menys en un mostreig aleatori de 10 persones d'aquesta comunitat pateixin la malaltia:

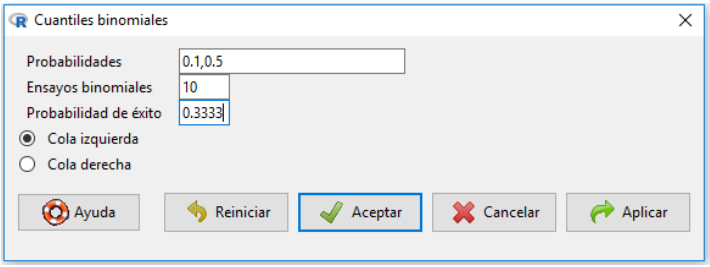
```
x = 4 f_4 = pbinom(4, size=10, prob=0.3333)
abline(v=x, col="blue")
abline(h=f_4, col="blue")
text(x, f_4,expression(P(X<=4)== P(X<5)),
pos=2, col="blue")
```



4.3.4. Quantils

El p-èsim quantil es defineix com el valor més petit de x, de manera que $F(x) = P(X \leq x) \geq p$. Aquest es calcula seleccionant l'opció *Cola izquierda*. També es pot calcular el valor més petit de x tal que $P(X > x) \geq p$ seleccionant l'opció *Cola derecha*. Continuant amb l'exemple, es determinarà el desè quantil, la mitjana i el valor de x tal que $P(X > x) = 0.15$.

El desè quantil i la mediana indiquen el valor de x tal que $P(X \leq x) = 0.1$ i $P(X \leq x) = 0.5$, respectivament. Per tant, si seleccionem la primera opció del submenú de *Distribución binomial*, ens apareix la finestra següent:




```
qbinom(c(0.1,0.5), size=10, prob=0.3333, lower.tail=TRUE)
```

```
[1] 1 3
```

Aleshores, el desè quantil és 1, és a dir $P(X \leq 1) \geq 0.1$, i la mediana és 3, és a dir, $P(X \leq 3) \geq 0.5$. Per calcular el valor de x tal que $P(X > x) = 0.15$, el paràmetre *Probabilidades* es configura en 0.15 i se selecciona l'opció *Cola derecha*; d'aquesta manera, es generen l'ordre i la sortida següents:

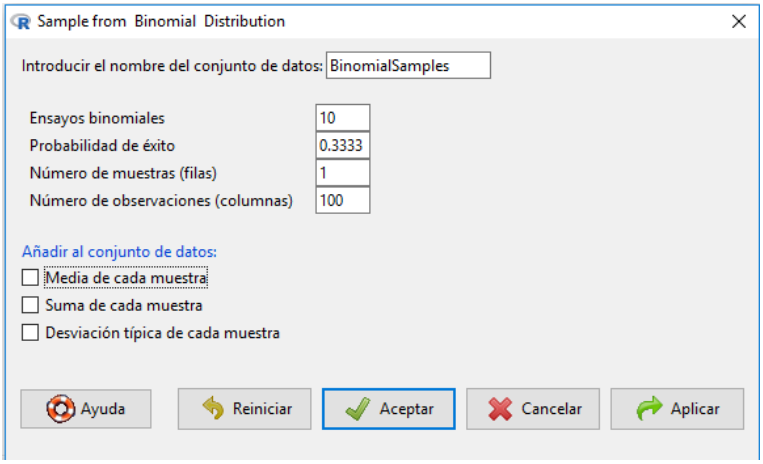
```
qbinom(c(0.15), size=10, prob=0.3333, lower.tail=FALSE)
```

```
[1] 5
```

la qual cosa indica que, $P(X > 5) \leq 0.15$.

4.3.5. Mostreig

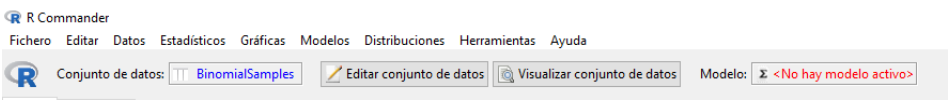
El mostreig simula escenaris aleatoris per a la distribució donada. Per a l'exemple previ, si volem simular una observació de 100 mostres, cada observació o experiment consisteix a seleccionar 10 persones de la població i comptar quantes pateixen la malaltia. Els paràmetres que hem d'introduir en el quadre de diàleg són els següents:



També s'hi poden afegir algunes mesures extres, com la mitjana, la suma de cada mostra i la desviació típica.

```
BinomialSamples <- as.data.frame(matrix(rbinom(1*100, size=10,
                  prob=0.3333), ncol=100))
rownames(BinomialSamples) <- "sample"
colnames(BinomialSamples) <- paste("obs", 1:100, sep="")
```

Una vegada hem simulat les dades, aquestes ja apareixen en el conjunt de dades de *R-Commander*. En conseqüència, el nom del conjunt de dades apareix al botó que hi ha al costat de l'extrem esquerre a la finestra principal. Les dades simulades es desen en el vector anomenat BinomialSamples i es poden veure fent clic a *Visualizar conjunto de datos*.



Tips & Tricks!

En cas de no disposar o que no s'estigui utilitzant *R-Commander*, es poden introduir directament en *R-Console* o *Rstudio* les funcions següents:

- dbinom() calcula la funció de distribució d'una distribució binomial, és a dir, $f(x) = P(X = x)$ per a tots els possibles valors de x.
- pbinom() calcula la probabilitat acumulada d'una distribució binomial, és a dir, $F(x) = P(X \leq x)$ o $1 - F(x) = P(X > x)$ per a un valor de x donat.
- qbinom() calcula el p-èsim quantil d'una distribució binomial, és a dir, el valor de x tal que $F(x) = P(X \leq x) \geq p$ per a un valor de p donat.
- rbinom() fa un mostreig o simulació d'un nombre determinat d'experiments d'una distribució binomial.

Noteu que la primera lletra indica: funció de distribució (d), probabilitat (p), quantil (q) o mostra aleatòria (r). La resta de lletres indiquen la distribució: binomial (binom), geomètrica (geom), binomial negativa (nbinom), hipergeomètrica (hyper) o Poisson (pois). D'aquesta manera, si el que es vol és calcular un quantil d'una distribució geomètrica, la funció corresponent és qgeom(); per a la distribució binomial negativa, és qnbinom(); per a la hipergeomètrica, és qhyper(), i per a la de Poisson, és qpois().

4.4. Exercicis proposats

1. Es llancen dos daus de sis cares. Un només té nombres imparells i l'altre, parells. Defineix la variable aleatòria X com la suma dels nombres que han sortit.

   a) Determineu la funció de densitat y convertiu-la en gràfic.

   b) Simuleu 1.000 llançaments i mostreu la seva freqüència relativa.

   c) Compareu (en el mateix gràfic) la funció de densitat i la freqüència relativa de les simulacions.

2. Segons *Chemical Engineering Progress* (novembre de 1990), aproximadament el 30% de tots els problemes de canonades a les plantes químiques es deuen a errors de l'operador.

   a) Amb les 20 fallades en les canonades següents, determineu la probabilitat que:

   - Almenys 10 es deguin a un error de l'operador.
   - No més de 4 siguin per un error de l'operador.
   - Exactament 5 siguin per un error de l'operador.

b) Quin és el nombre esperat d'errors causats per un operador que poden ocórrer en els 20 problemes en les canonades següents? I la variància?

c) Simuleu una observació de 1000 mostres (cada mostra consisteix a comptar quantes fallades són causades per un error de l'operador en els 20 problemes en les canonades següents) i compareu els resultats.

3. Durant la Segona Guerra Mundial, es van llançar 535 bombes sobre el sud de Londres. Aquesta àrea ha estat dividida en una quadrícula de 576 petits quadrats de 0,25 metres quadrats cadascun. Assumint que l'objectiu és aleatori (cada bomba impacta en un sol lloc alhora, cada lloc té la mateixa probabilitat de ser impactat i els impactes són successos independents), trobeu:

a) El gràfic de les funcions de densitat i de distribució de la variable que indica el nombre de bombes que impacten en un quadrat en particular.

b) Quina probabilitat hi ha que exactament 2 bombes impactin en una zona en particular?

c) Quina probabilitat hi ha que una zona concreta sigui bombardejada (hi caigui almenys una bomba)?

d) El gràfic de la funció de densitat de la variable que indica el nombre de zones que reben exactament 2 impactes.

e) Quantes zones s'espera que sofreixin exactamente dos impactes?

f) El gràfic de la funció de densitat de la variable que indica el nombre de zones que han de ser inspeccionades per trobar-ne 10 que hagin estat bombardejades.

g) Quantes zones han de ser inspeccionades per trobar-ne 10 que hagin estat bombardejades?

5
Variables aleatòries contínues i distribucions de probabilitat

5.1. Introducció i objectius

La sessió anterior estava centrada en l'anàlisi de probabilitats de variables aleatòries discretes (VAD) i l'aplicació de les seves distribucions de probabilitat més habituals en enginyeria. L'objectiu d'aquesta sessió és mostrar, analitzar i aplicar el concepte de la variable aleatòria contínua (VAC), simular repeticions d'un experiment i comparar-ne els resultats amb la seva distribució. A més, es descriuen els models de distribució continus més utilitzats. En finalitzar aquesta sessió, l'estudiant ha de ser capaç de:

- Representar gràficament una distribució de variable aleatòria contínua usant *R*.
- Simular la repetició de diferents experiments aleatoris continus i comparar el resultat d'aquests experiments amb les probabilitats estudiades prèviament.
- Calcular i interpretar el valor esperat i la variància d'una variable aleatòria contínua.
- Reconèixer i aplicar correctament les distribucions de probabilitat contínua més habituals en enginyeria.

5.2. Variables aleatòries contínues (VAC)

Com ja s'ha descrit en la sessió prèvia, una variable aleatòria $\mathbf{X}$ es pot definir com la funció que assigna un nombre real a cada sortida d'un espai mostral Ω. En altres paraules, és una funció de domini Ω i rang $\mathbb{R}$. Una variable aleatòria és una variable aleatòria contínua (VAC o CRV, per les seves inicials en anglès) si el seu conjunt de sortides no es pot explicar. Això és, un espai mostral conté un nombre infinit de possibilitats igual al nombre de punts en un segment de línia. Pren valors en una escala contínua, normalment valors que són precisament els mateixos que conté un espai mostral continu. En els problemes pràctics, aquestes variables representen dades mesurades, com ara tots els possibles pesos, altures, temperatures, distàncies, etc.

5.2.1. Funció de densitat

La funció de densitat de probabilitat (o, simplement, funció de densitat) d'una variable aleatòria contínua X, definida en l'espai mostral Ω, és l'aplicació de f tal que:

$$f : X(\Omega) \to \mathbb{R},$$
$$x_i \to f(x_i).$$

Aquesta funció té les propietats següents:

- El seu valor sempre és un nombre real positiu o zero: $0 \leq f(x)$.
- L'àrea total sota la seva corba és 1: $\int_{-\infty}^{\infty} f(x)dx = 1$.
- La probabilitat que X prengui un valor en l'interval $[a, b]$ és l'àrea sota la corba de la funció de densitat en aquest interval: $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx, \qquad \forall a \leq b.$
- Tenint en compte que la probabilitat que X prengui un valor exacte és zero ($P(X = x) = 0$),

$$f(x) \neq P(X = x)$$
$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$

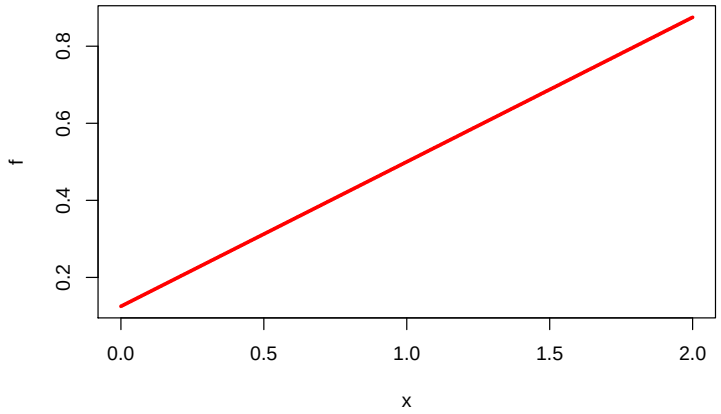
Exemple: Considerem X la variable aleatòria que indica la càrrega dinàmica en un pont (en newtons), la funció de densitat de la qual és expressada per:

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \leq x \leq 2 \\ 0 & \text{resta.} \end{cases}$$

La funció de densitat d'una VAC es pot representar gràficament per mitjà d'una corba contínua. En *R*, es pot fer executant les instruccions següents:

```
rm(list=ls()) # Elimina tots els objectes de l'espai de treball
graphics.off() # Elimina els gràfics creats anteriorment
```

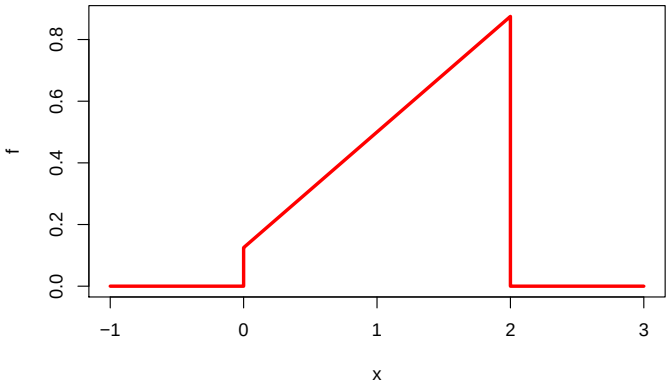
```
x = seq(0,2,0.0001) # Resultats possibles
f = 1/8+3/8*x # Funció de densitat
plot(x, f, type="l", col="red", lwd=3)
```



Noteu que, a diferència del gràfic de funció de densitat de VAD, el vector de resultats possibles o espai mostral (x) conté molts valors entre 0 i 2 (i tendeix a infinit) espaiats igualment. D'altra banda, la funció plot() utilitza el paràmetre type=l per especificar que es dibuixarà una línia que uneix tots els punts, al contrari que type=h, que dibuixa línies verticals, com en el cas de VAD.

Si es vol representar gràficament la funció de probabilitat incloent-hi també alguns valors de X en què la probabilitat és zero, es pot executar el codi següent:

```
LowLim = 0 # Valor mínim que pren X
UppLim = 2 # Valor màxim que pren X
delta = 0.001 # Resolució, diferència entre valors de X

# Valors parcials de X
x.less0 = seq(LowLim-1, LowLim, delta)  #-1 < X < 0
x.less2 = seq(LowLim, UppLim, delta) # 0 < X < 2
x.greater2 = seq(UppLim, UppLim+1, delta) # 2 < X < 3

# Valors parcials de la funció de densitat
fx.less0 = rep(0, length(x.less0))  # f(x) per a -1 < X < 0
fx.less2 = 1/8+3/8*x.less2 # f(x) per a 0 < X < 2
fx.greater2 = rep(0, length(x.greater2)) # f(x) per a 2 < X < 3

# Vectors finals
x = c(x.less0, x.less2, x.greater2) # -1 < X < 3
f = c(fx.less0, fx.less2, fx.greater2) # f(x) per a -1 < X < 3

# Representació gràfica
plot(x, f, type="l", col="red", lwd=3);
gra.fx = recordPlot()
```

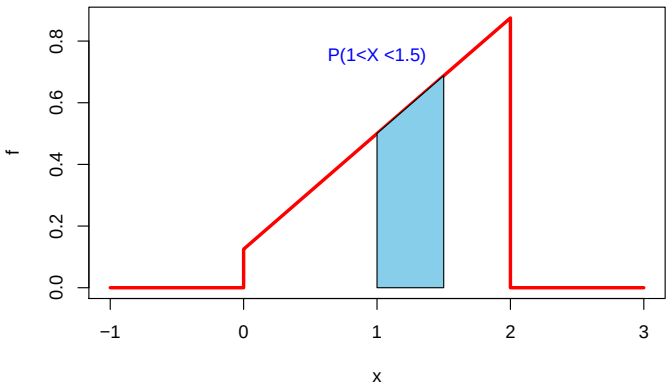


Finalment, si es vol destacar en el gràfic una probabilitat específica, per exemple $P(1 < X < 1.5)$, és a dir:

$$P(1 < X < 1.5) = \int_1^{1.5} f(x)dx = \int_1^{1.5} \left(\frac{1}{8} + \frac{3}{8}x\right) dx = 0.297.$$

Es pot agregar un polígon al gràfic de la funció de densitat especificant-ne els vèrtexs.

```
x.i = 1 # Límit inferior
x.f = 1.5 # Límit superior
x.x = seq(x.i, x.f, delta) # Valors de X dins dels límits
coord.x = c(x.i, x.x, x.f) # Vector de vèrtexs en la coordenada x
coord.y = c(0, 1/8+3/8*x.x, 0) # Vector de vèrtexs en la coordenada y
polygon(coord.x, coord.y, col="skyblue") # Àrea sota la corba
text(x.i, 0.7, "P(1 < X < 1.5)", pos=3, col="blue")
```



5.2.2. Funció de distribució

La funció de distribució o funció de probabilitat acumulada associada a una variable aleatòria contínua X, definida en un espai mostral Ω, és l'aplicació F que assigna a cada element x_i de $X(\Omega)$ la probabilitat que la variable X prengui qualsevol valor menor o igual que x_i:

$$F : X(\Omega) \to \mathbb{R}$$

$$x_i \to F(x_i) = P(X \leq x_i) = \int_{-\infty}^{x} f(u)du.$$

Per tant:

$$P(a \leq X \leq b) = F(b) - F(a), \qquad \forall a \leq b$$

$$f(x) = F'(x).$$

En l'exemple, la funció de distribució és expressada per:

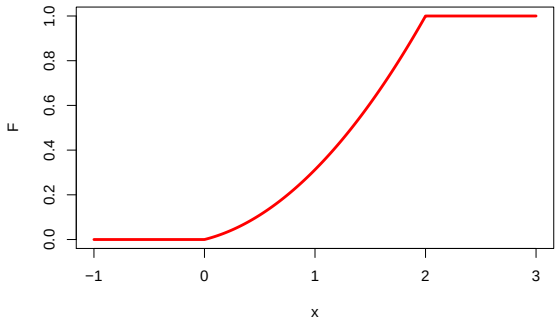
$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{0}^{x} \left(\frac{1}{8} + \frac{3}{8}x\right)dx = \frac{1}{8}x + \frac{3}{16}x^2 \qquad \text{per a} \qquad 0 \leq x \leq 2,$$

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{8}x + \frac{3}{16}x^2 & 0 \leq x \leq 2 \\ 1 & 2 < x. \end{cases}$$

La funció de distribució es representa gràficament mitjançant una corba contínua que indica la probabilitat $F(x) = P(X \leq x)$, on x pot ser qualsevol valor real. Noteu que aquesta funció comença en 0 i acaba en 1. La funció de distribució de l'exemple es representa gràficament de la manera següent:

```
LowLim = 0 # Valor mínim que pren X
UppLim = 2 # Valor màxim que pren X
delta = 0.001 # Resolució, diferència entre valors de X

# Valors parcials de X
x.less0 = seq(LowLim-1, LowLim, delta) # -1 < X < 0
x.less2 = seq(LowLim, UppLim, delta) # 0 < X < 2
x.greater2 = seq(UppLim, UppLim+1, delta) # 2 < X < 3

# Valors parcials de la funció de distribució
Fx.less0 = rep(0, length(x.less0)) # F(x) per a -1 < X < 0
Fx.less2 = 1/8*x.less2+3/16*x.less2^2 # F(x) per a 0 < X < 2
Fx.greater2 = rep(1, length(x.greater2)) # F(x) per a 2 < X < 3
```

```
# Vectors finals
x = c(x.less0, x.less2, x.greater2) # -1 < X < 3
F = c(Fx.less0, Fx.less2, Fx.greater2) # F(x) per a -1 < X < 3
# Representació gràfica
plot(x, F, type="l", col="red", lwd=3)
```

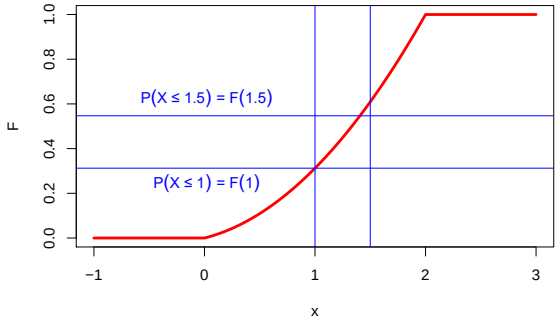


Finalment, si es vol destacar en el gràfic una probabilitat específica, per exemple la probabilitat que la càrrega dinàmica estigui entre 1 i 1.5, $P(1 < X < 1.5)$, que és expressada per:

$$P(1 < X < 1.5) = F(1.5) - F(1) = \\ = \left[\frac{1}{8}(1.5) + \frac{3}{16}(1.5)^2\right] - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right] = 0.297.$$

S'agreguen dues línies horitzontals: una a $F(1.5)$ i una altra a $F(1)$.

```
a = 1; b = 1.5
Fa = 1/8* a+3/16*a^2;Fb = 1/8*a +3/16*b^2
abline(v=b,col="blue")
abline(h=Fb,col="blue")
text(0,Fb, expression(P(X <= 1.5) == F(1.5)), pos=3, col="blue")
abline(v=a,col="blue")
abline(h=Fa,col="blue")
text(0,Fa, expression(P(X <= 1) == F(1)), pos=1, col="blue")
```



Tips & Tricks!

La diferència entre les definicions de VAD i VAC està en el vector de l'espai mostral x=seq(). Un vector amb molts valors (acostant-se a l'infinit) de x en què la distància entre valors consecutius és molt petita pot considerar-se un vector d'espai mostral d'una VAC. Com més gran és aquest vector, millor és l'aproximació. No obstant això, el càlcul computacional també és més gran. Per tant, s'aconsella seleccionar una longitud de vector apropiada, de tal manera que pugui considerar-se VAC però sense sobrecarregar el treball de l'ordinador.

La instrucció polygon(coord.x, coord.y) dibuixa un polígon els vèrtexs del qual s'especifiquen en els vectors coord.x i coord.y.

5.2.3. Mesures característiques de les VAC

Aquests paràmetres o característiques, com succeeix en el cas de les VAD, quantifiquen la tendència central i la variabilitat o dispersió de la variable aleatòria contínua. De fet, coneixent aquestes quantitats, a part de la distribució completa, podem tenir una idea de la naturalesa del sistema.

Valor esperat

Aquest paràmetre explica com "esperem" que la variable es comporti de mitjana a llarg termini (és el que també es denomina *teoria freqüencial de la probabilitat*). El valor esperat $E(X)$ de la VAC X és expressat per:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

D'aquesta manera, el valor esperat de la càrrega dinàmica del pont de l'exemple (X) és expressat per:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{2} x\left(\frac{1}{8} + \frac{3}{8}x\right)dx = \frac{5}{4} = 1.25.$$

Variància

Aquest paràmetre mesura la dispersió dels possibles valors de X. La variància és la mitjana (esperada) de la distància al quadrat (o desviació) de la mitjana:

$$\sigma^2 = Var(X) = E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2 =$$
$$= \int_{-\infty}^{\infty} (x - E[X])^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - E[X]^2.$$

Per tant, la variància de la càrrega dinàmica de l'exemple (X) és:

$$V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{0}^{2} \left(x - \frac{5}{4}\right)^2 \left(\frac{1}{8} + \frac{3}{8}x\right) dx = \frac{13}{48} = 0.2708.$$

5.2.4. Ús de sample() per generar simulacions

Amb *R*, també es poden generar observacions d'una VAC utilitzant la funció sample() (com ja s'ha explicat en la sessió anterior). El vector x, que conté els valors que es vol mostrejar (espai mostral), ha de ser prou gran per representar el nombre més gran possible de valors de l'interval continu. El vector de probabilitats amb què es regeix cada element de l'espai mostral es genera usant la funció de densitat de la VAC.

Considerant l'exemple, un experiment o observació es pot simular de la manera següent:

```
x = seq(0,2,0.01)
f = 1/8+3/8*x sample(x, size=1, prob=f)
```

```
[1] 0.57
```

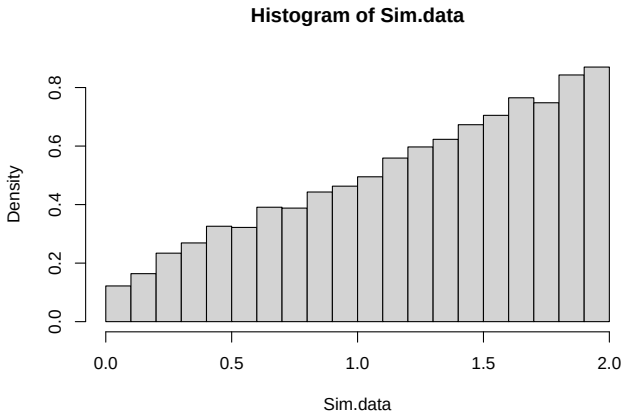
Per simular 100 observacions:

```
sample(x, size=100, replace=T, prob=f)
```

```
[1] 0.68 1.28 1.72 1.07 1.92 1.76 0.60 1.71 1.69 1.70 0.92 0.66
    1.64 0.80 1.87
[16] 1.28 1.62 1.97 1.96 0.33 1.99 0.31 0.54 1.59 0.75 0.49 1.37
     1.68 0.53 1.75
[31] 0.38 1.82 1.71 1.44 1.82 1.41 1.22 0.42 1.75 1.92 1.21 1.76
     1.31 0.84 1.31
[46] 1.58 0.88 0.96 1.54 1.73 1.12 1.46 1.36 1.66 0.59 1.20 0.13
     1.70 1.71 0.84
[61] 1.99 1.94 1.57 1.84 1.73 0.48 1.13 1.79 1.58 1.74 1.86 1.61
     1.22 1.75 1.01
[76] 1.40 0.63 1.65 1.00 1.99 0.81 1.49 1.18 1.27 0.97 1.39 0.93
     1.91 1.95 1.65
[91] 0.50 0.01 1.62 1.34 1.68 1.54 0.35 1.99 1.14 1.63
```

Aquesta simulació no és útil, a causa de la gran quantitat de resultats possibles que no estan inclosos en el vector x. Una bona simulació d'una VAC hauria d'incloure milers de resultats possibles, no solament 200. Per incrementar la longitud del vector x, i que la separació entre valors sigui només de 0.0001, simular 1000 observacions i veure la densitat dels resultats (histograma), es pot executar el codi següent:

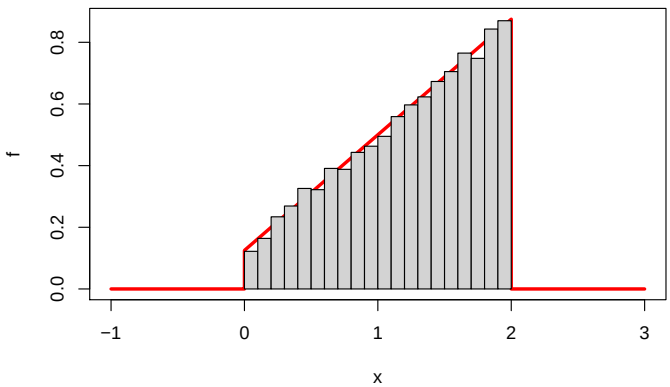
```
x = seq(0,2,0.0001)
f = 1/8+3/8*x set.seed(10)
Sim.data = sample(x, size=10000, replace=T, prob=f)
hist(Sim.data, freq=FALSE); gra.his = recordPlot()
```



5.2.5. Validació dels experiments simulats i la seva distribució de probabilitat

La comparació de la distribució d'una VAC amb les simulacions realitzades es fa de la mateixa manera que amb una VAD, superposant la funció de densitat i l'histograma dels resultats de les simulacions (taula de freqüències).

```
gra.fx hist(Sim.data, freq=FALSE, add=T)
```



D'altra banda, també es poden comparar la mitjana i la variància dels experiments simulats amb el valor esperat i la variància de la funció de densitat; aquests valors han de ser similars. Recordem que, per a l'exemple, el valor esperat és 1.25 i la variància és 0.2708. La mitjana i la variància de les dades simulades en aquest exemple són:

```
mean(Sim.data)
```

```
[1] 1.251668
```

```
var(Sim.data)
```

```
[1] 0.2707881
```

5.3. Distribucions de probabilitat contínues més comunes

Com ja s'ha esmentat en la sessió anterior, hi ha funcions de probabilitat tant en VAD com en VAC que són típiques i àmpliament utilitzades en estadística. En aquesta sessió, s'estudia dues de les funcions de distribució o models de probabilitat continus més importants en enginyeria. El que es vegi aquí podrà extrapolar-se a la resta de models de probabilitat continus, tenint en compte els paràmetres per a cada cas.

Distribució uniforme

Si una VAC pren qualsevol valor en un interval $[a, b]$ amb la mateixa probabilitat, diem que aquesta variable segueix una distribució uniforme contínua. La seva funció de densitat, la seva mitjana i la seva variància són expressades per:

$$f(x) = \frac{1}{b-a} \qquad \forall x \in [a, b],$$

$$E(X) = \frac{a+b}{2},$$

$$V(X) = \sigma^2 = \frac{(b-a)^2}{12}.$$

Distribució exponencial

La família de distribucions exponencials proporciona models que s'utilitzen molt en la ciència i en l'enginyeria. La VAC, que és igual a la distància en els successius esdeveniments d'un procés de Poisson amb una mitjana d'esdeveniments $\lambda > 0$ per interval la unitat, segueix una distribució exponencial amb paràmetre λ. La seva funció de densitat, la seva mitjana i la seva variància són expressades per:

$$f(x) = \lambda e^{-\lambda x} \qquad \forall x \in \mathbb{R}^+,$$

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \sigma^2 = \frac{1}{\lambda^2}.$$

Distribució normal

La distribució de probabilitat contínua més important en tot el camp de l'estadística és la distribució normal. La seva representació gràfica, denominada *corba normal*, és la corba amb forma de campana, la qual descriu aproximadament el fenomen que es presenta en la naturalesa, en la indústria i en la recerca. Les mesures físiques en àrees com ara els experiments meteorològics o els estudis pluvials i les mesures de parts manufacturades solen quedar més ben explicades utilitzant una distribució normal.

Si una VAC X té una distribució normal amb paràmetres μ i σ ($X \hookrightarrow N(\mu, \sigma)$) o $\left(N\left(\mu, \sigma^2\right)\right)$, on $-\infty < \mu < \infty$ i $0 < \sigma$, la seva funció de densitat, la seva mitjana i la seva variància són expressades per:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \forall x \in \mathbb{R},$$

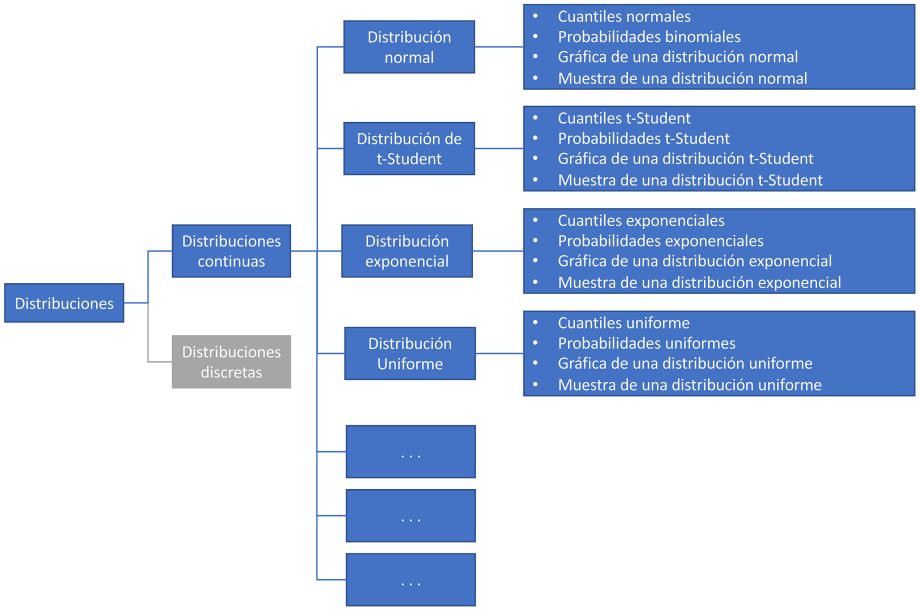
$$E(X) = \mu, \qquad V(X) = \sigma^2.$$

Cada corba de densitat és simètrica respecte a μ i té forma de campana, de manera que el centre de la campana (punt de simetria) és alhora la mitjana i la mediana de la distribució. El valor de σ és la distància des de μ fins als punts d'inflexió de la corba (els punts en què la corba passa de ser còncava cap amunt a ser còncava cap avall). Els valors alts de σ estenen la forma de la corba al llarg de μ, mentre que els valors petits de σ produeixen corbes amb un pic alt en μ.

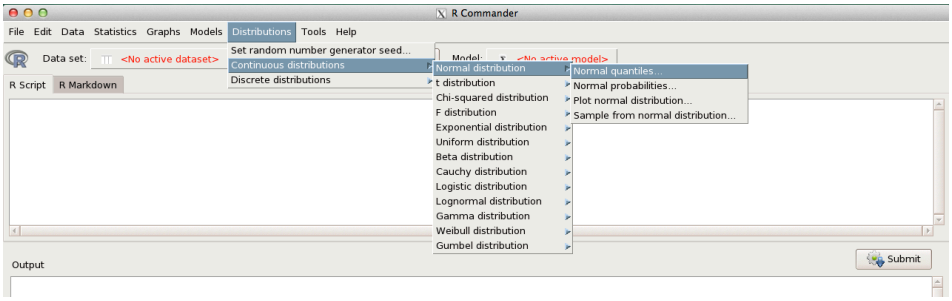
Donada una funció de distribució contínua comunament utilitzada en enginyeria (uniforme, exponencial, normal, t-Student, etc.), amb els seus respectius paràmetres, fent ús de *R* i/o *R-Commander* es poden obtenir molt fàcilment la probabilitat d'un interval, els quantils i els gràfics de distribucions estadístiques estàndard (que poden usar-se, per exemple, com a substitut de les taules estadístiques); a més, es poden generar mostrejos o simulacions d'aquestes distribucions.

Si es disposa de *R-Commander*, a la barra superior hi tenim, en la ruta *Distribuciones > Distribuciones continuas*, les distribucions més utilitzades. L'arbre del submenú complet per a distribucions de probabilitat discretes es mostra a la figura següent. La major part d'opcions del menú ens porten a diferents quadres de diàleg. Les opcions del menú estan inactives (en gris) si no es poden aplicar al context actual.

Si no es disposa de *R-Commander*, o si es prefereix usar *R-Console* o *Rstudio*, o simplement s'està creant un script per a ús posterior, també indiquem quines són les instruccions escaients per a cada cas. D'altra banda, per no ser repetitius, en aquesta guia només s'expliquen totes les opcions concernents a la distribució normal; les altres distribucions tenen les mateixes opcions, amb la diferència que els paràmetres són diferents per a cada cas. Per exemple, la distribució normal té com a paràmetres μ i σ, mentre que la distribució exponencial només té λ.



Donada una VAC X que segueix una distribució normal de paràmetres μ i σ, és a dir, $X \hookrightarrow N(\mu, \sigma^2)$, es pot calcular la funció de densitat i probabilitats entre dos valors de X, crear els gràfics de les funcions de densitat i de distribució, calcular-ne els quantils normals i fer un mostreig o simulació d'experiments normals, tal com es mostra a la figura.

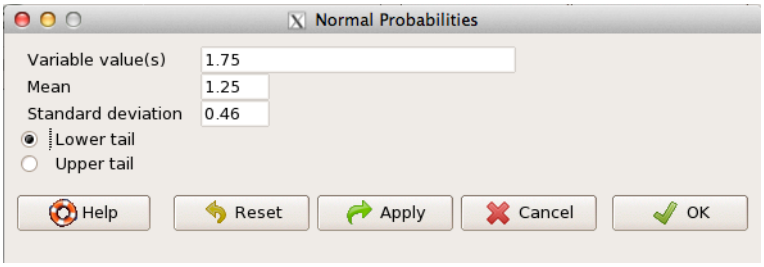


5.3.1. Probabilitats

Calculem la probabilitat que la variable X sigui com a màxim x, $P(X \leq x)$.

Exemple: El temps de reacció en trànsit d'un senyal de fre (llums de fre) pot modelitzar-se com una distribució normal amb una mitjana d'1.25 segons i una desviació típica de 0.46 segons. Quina probabilitat hi ha que el temps de reacció sigui inferior a 1.75 segons?

Si definim la VAC X, que denota el temps de reacció dels llums de fre, aleshores $X \hookrightarrow N(1.25, 0.46^2)$. Per tant, $P(X \leq 1.75)$ s'obté en introduir els paràmetres següents en el quadre de diàleg que es genera en executar *Distribuciones* > *Distribuciones continuas* > *Distribución normal* > *Probabilidades normales*.



A la finestra d'instruccions, es genera el codi següent, amb la respectiva resposta a la finestra de resultats:

```
pnorm(c(1.75), mean=1.25, sd=0.46, lower.tail=TRUE)
```

```
[1] 0.861472
```

Per tant, la probabilitat que el temps sigui inferior a 1.75 segons és $P(X \leq 1.75) = 0.861472$. Noteu que la probabilitat es calcula mitjançant la funció pnorm(x, mean=mu, sd=sigma, lower.tail=TRUE), on x és el valor del qual es vol calcular la probabilitat acumulada, mu és la mitjana, sigma és la desviació típica i lower.tail=TRUE indica que es calcularà la probabilitat de la cua esquerra. D'aquesta manera, si es prefereix usar directament les ordres en *R-Console* o *Rstudio*, es pot executar la instrucció donada.

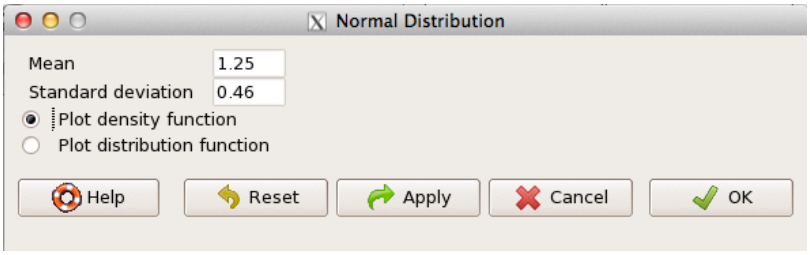
A més, $P(X > x)$ també es pot calcular seleccionant l'opció *Cola derecha* en lloc de *Cola izquierda* o canviant l'opció lower.tail=TRUE. D'aquesta manera, continuant amb l'exemple, $P(X > 1.75) = P(X \geq 1.75) = 1 - P(X \leq 1.75)$.

```
pnorm(c(1.75), mean=1.25, sd=0.46, lower.tail=FALSE)
```

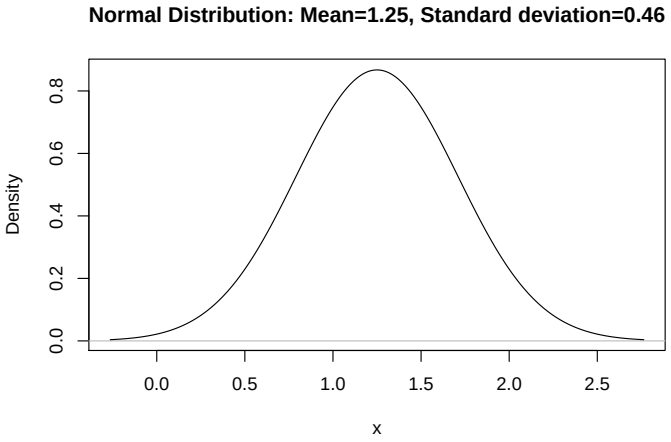
```
[1] 0.138528
```

5.3.2. Gràfic d'una distribució

Aquesta opció permet generar la representació gràfica de la funció de densitat o la funció de distribució de la distribució donada. La funció de densitat de l'exemple es genera de la manera següent:

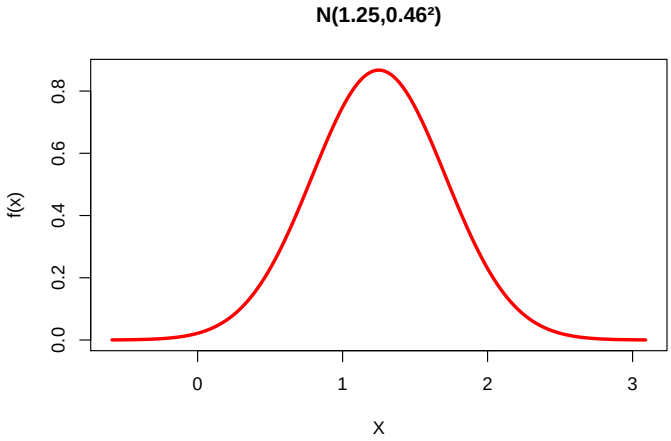



```
local({  .x <- seq(-0.264, 2.764, length.out=1000) plotDistr
(.x, dnorm(.x, mean=1.25, sd=0.46), cdf=FALSE, xlab="x",
ylab="Density", main=paste("Normal Distribution: Mean=1.25, Standard
deviation=0.46")) })
```



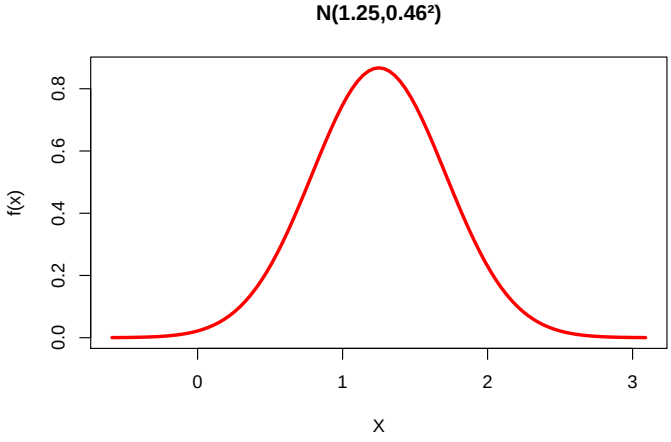
Noteu que els valors de la funció de densitat, $f(x)$, es calculen mitjançant la funció dnorm(x, mean=mu, sd=sigma), on x és el vector de valors de X de què es vol calcular la funció de densitat. Per tant, per fer el gràfic directament des de *R-Console* o *Rstudio* es poden utilitzar les instruccions explicades al principi de la guia de la manera següent:

```
mu = 1.25
sigma = 0.46
x = seq(mu-4*sigma, mu+4*sigma, length=1000) # Resultats possibles
f = dnorm(x, mean=mu, sd=sigma) # Funció de densitat
plot(x, f, type="l", col="red", lwd=3, main="N(1.25,0.46²)",
xlab="X", ylab="f(x)")
```



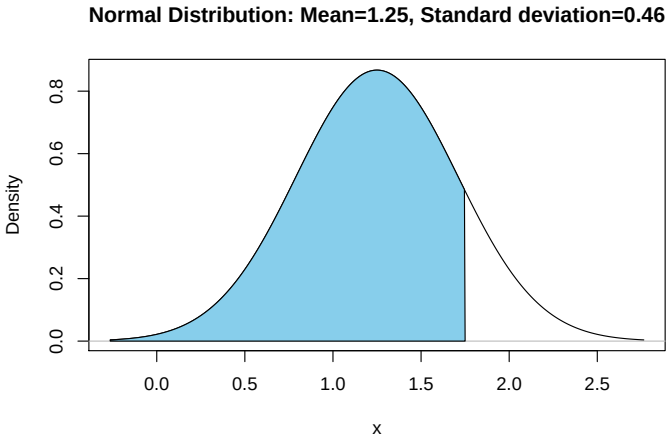
O, utilitzant la funció curve():

```
mu = 1.25
sigma = 0.46
curve(dnorm(x, mean=mu, sd=sigma), xlim = c(mu-4*sigma, mu+4*sigma),
col="red", lwd=3, main="N(1.25,0.46²)", xlab="X", ylab="f(x)")
```

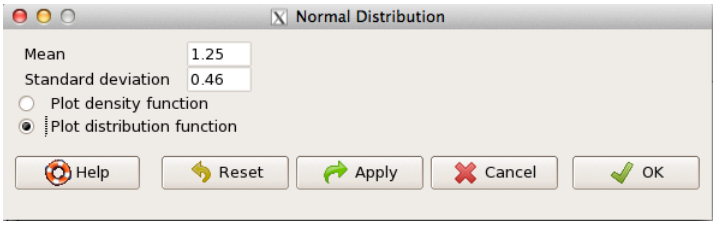


Independentment de com s'hagi generat el gràfic de la funció de densitat, és possible incloure o destacar alguna probabilitat dins el gràfic (actiu). Continuant amb l'exemple de quina probabilitat hi ha que el temps de reacció sigui inferior a 1.75 segons, es pot representar agregant el codi següent:

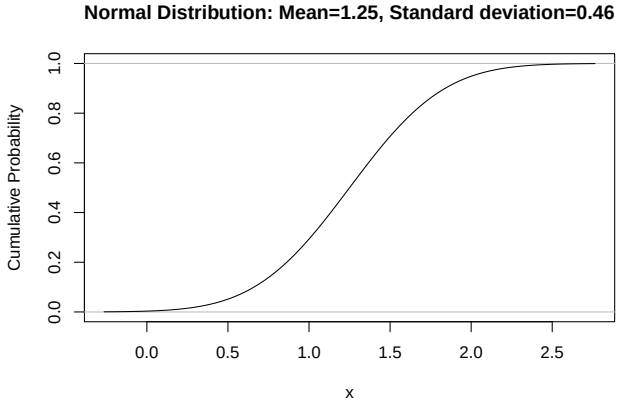
```
cord.x=c(-0.264, seq(-0.264,1.75,0.01),1.75)
#Vector de vèrtexs en x per al polígon
cord.y=c(0, dnorm(seq(-0.264,1.75,0.01),1.25,0.46),0)
#Vector de vèrtexs en y
polygon(cord.x, cord.y, col='skyblue')
```



D'altra banda, el gràfic de la funció de distribució es pot generar introduint els paràmetres següents en el quadre de diàleg:

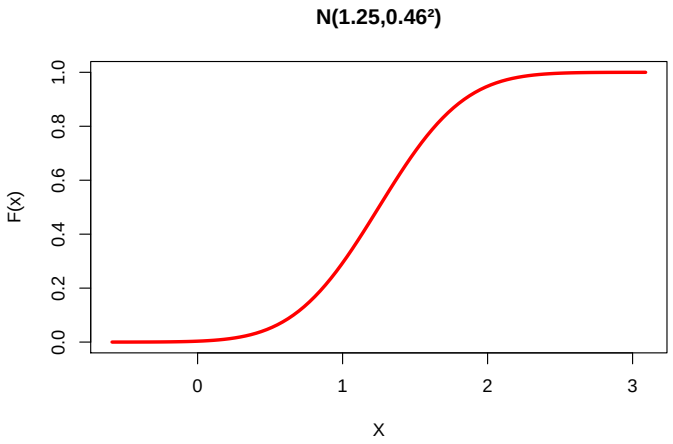



```
local({  .x <- seq(-0.264, 2.764, length.out=1000)
plotDistr(.x, pnorm(.x, mean=1.25, sd=0.46),
cdf=TRUE, xlab="x", ylab=Çumulative Probability",
main=paste("Normal Distribution: Mean=1.25, Standard deviation=0.46"))
})
```



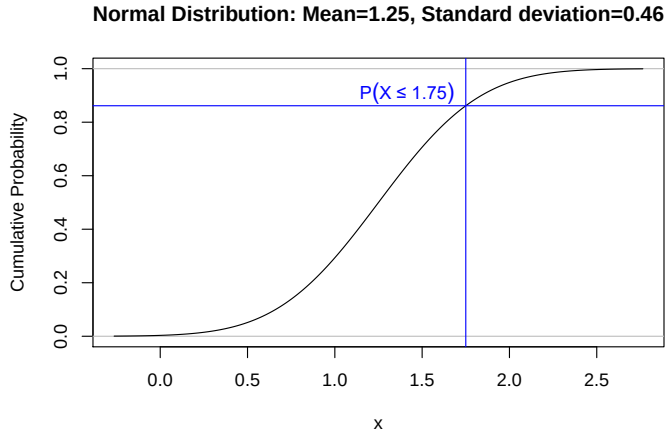
Igualment, sense necessitat d'usar *R-Commander*, mitjançant la funció curve() es pot generar el gràfic de la distribució amb l'opció de personalitzar-la completament:

```
mu = 1.25; sigma = 0.46
curve(pnorm(x, mean=mu, sd=sigma),
xlim = c(mu-4*sigma, mu+4*sigma), col="red", lwd=3,
main="N(1.25,0.46²)", xlab="X", ylab="F(x)")
```



Anàlogament, es pot representar la probabilitat que el temps de reacció sigui inferior a 1.75 segons:

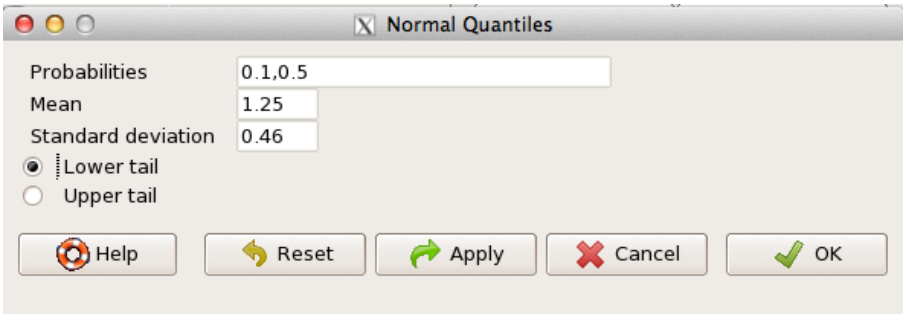
```
x = 1.75
F_x = pnorm(c(1.75), mean=1.25, sd=0.46, lower.tail=TRUE)
abline(v=x, col="blue");
abline(h=F_x, col="blue")
text(x, F_x,expression(P(X<=1.75)), pos=2, col="blue")
```



5.3.3. Quantils

Aquesta funció calcula la "funció inversa" de la funció de distribució donada. El p-èsim quantil es defineix com el valor x tal que $F(x) = P(X \leq x) = p$. Aquest es calcula seleccionant l'opció *Cola izquierda*. També es pot calcular el valor de x tal que $P(X > x) \geq p$ seleccionant l'opció *Cola derecha*. En l'exemple donat, es determinen el desè quantil, la mediana i el valor de x tal que $P(X > x) = 0.15$.

El desè quantil i la mediana indiquen el valor de x tal que $P(X \leq x) = 0.1$ i $P(X \leq x) = 0.5$, respectivament. Per tant, si seleccionem la primera opció del submenú *Distribución normal*, apareix la finestra següent:




```
qnorm(c(0.1,0.5), mean=1.25, sd=0.46, lower.tail=TRUE)
```

```
[1] 0.6604863 1.2500000
```

Aleshores, el desè quantil és 0.66, ja que $P(X \leq 0.66) = 0.1$ i, com caldria esperar, la mediana és igual a la mitjana, ja que $P(X \leq 1.25) = 0.5$. Per calcular el valor de x tal que $P(X > x) = 0.15$, el paràmetre *Probabilidades* s'ha de canviar a 0.15 i seleccionar *Cola derecha*, i es generen les instruccions següents i el seu resultat:

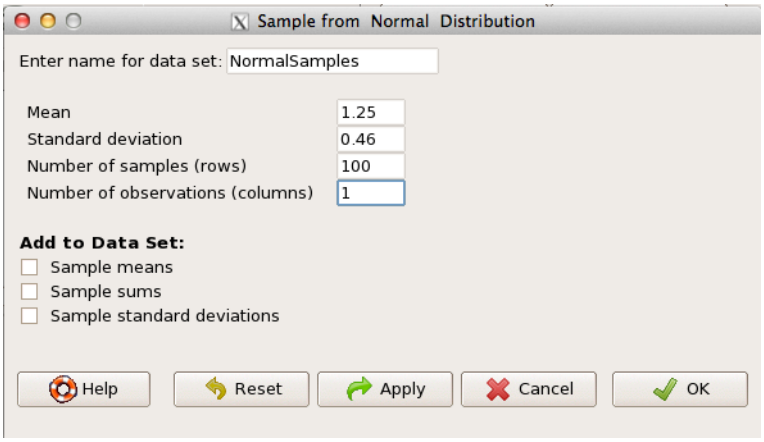
```
qnorm(c(0.15), mean=1.25, sd=0.46, lower.tail=FALSE)
```

```
[1] 1.726759
```

que indica que $P(X > 1.727) = 0.15$.

5.3.4. Mostreig

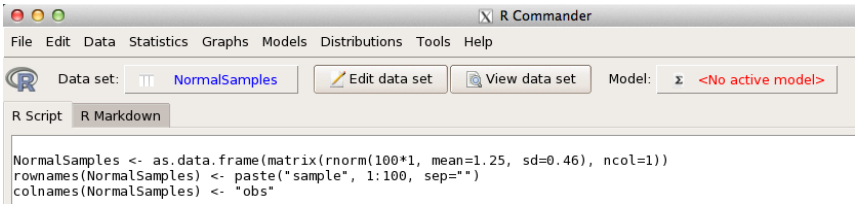
El mostreig simula escenaris aleatoris per a la distribució donada. Amb l'exemple previ, si volem simular una observació de 100 mostres, en què cada mostra consisteixi en una càrrega dinàmica sobre el pont, s'han d'introduir els paràmetres següents al quadre de diàleg:



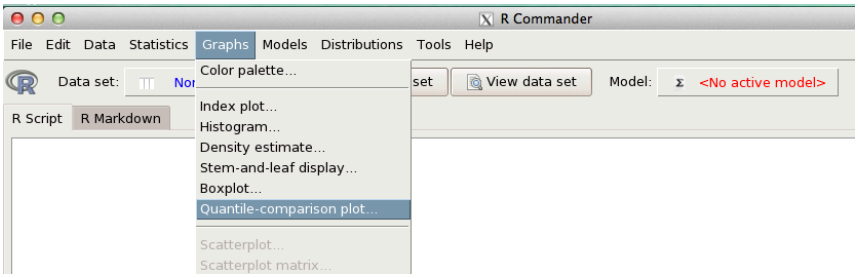
També es poden afegir algunes mesures de les dades simulades, com la mitjana, la sumatòria i la desviació típica:

```
NormalSamples <- as.data.frame(matrix(rnorm(100*1, mean=1.25,
sd=0.46), ncol=1))
rownames(NormalSamples) <- paste("sample", 1:100, sep="")
colnames(NormalSamples) <- "obs"
```

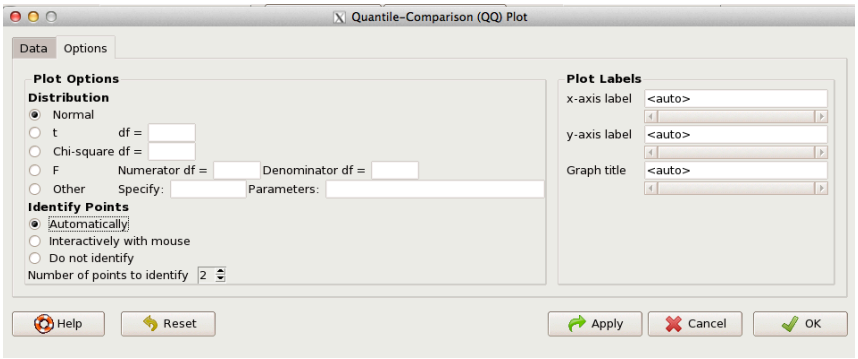
Una vegada hem simulat les dades, aquestes ja apareixen en el conjunt de dades de *R-Commander*. En conseqüència, el nom del conjunt de dades apareix en el botó que hi ha al costat de l'extrem esquerre a la finestra principal. Les dades simulades es desen en el vector denominat NormalSamples i es poden veure fent clic a *Visualizar conjunto de datos* o *View data set*.



El conjunt de dades simulades es pot mostrar a través de l'índex de gràfics, histogrames, etc. (veg. la figura següent). Un mètode important per comparar si les dades simulades segueixen una distribució normal és utilitzar el gràfic de comparació de quantils (QQ en anglès), un mètode gràfic per comparar dues distribucions de probabilitat mostrant-ne els quantils, l'un enfront de l'altre. Si les dues distribucions comparades són similars, els punts en el gràfic QQ formen una línia $y = x$, aproximadament. Si les distribucions són linealment dependents, els punts en el gràfic QQ formen una línia, però no necessàriament $y = x$.

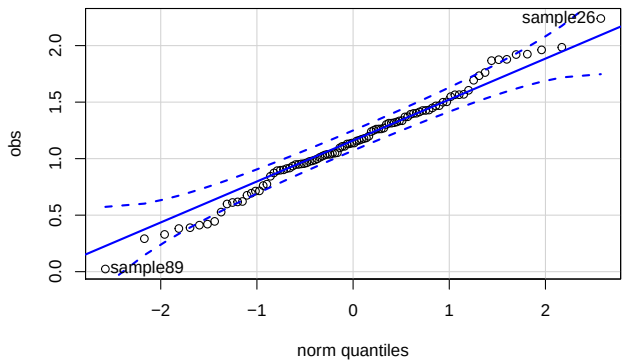


Per comparar les dades simulades amb una distribució normal, hem de seleccionar a *Opciones* el paràmetre *Normal*, com veiem a continuació:



Quan acceptem, es generen les ordres i la sortida següents:

```
with(NormalSamples, qqPlot(obs, dist="norm", id=list(method="y", n=2,
labels=rownames(NormalSamples))))
```




```
sample89 sample26
      89       26
```

Tips & Tricks!

En cas de no disposar o fer ús de *R-Commander*, es poden introduir directament en *R-Console* o *Rstudio* les funcions següents:

- dnorm() calcula la funció de distribució d'una distribució normal, és a dir, $f(x)$ per a tots els valors possibles de x.
- pnorm() calcula la probabilitat acumulada d'una distribució normal, és a dir, $F(x) = P(X \leq x)$ o $1 - F(x) = P(X > x)$ per a un valor de x donat.
- qnorm() calcula el p-èsim quantil d'una distribució normal, és a dir, el valor de x tal que $F(x) = P(X \leq x) \geq p$ per a un valor de p donat.
- rnorm() efectua un mostreig o simulació d'un nombre determinat d'experiments d'una distribució normal.

Noteu que la primera lletra indica: funció de distribució (d), probabilitat (p), quantil (q) o mostra aleatòria (r). Les altres lletres indiquen la distribució: normal (normal), exponencial (exp), t-Student (t), etc. D'aquesta manera, si el que es vol és calcular un quantil d'una distribució de t-Student, la funció corresponent és qt(), per a la distribució exponencial és qexp(), etc.

Una altra manera de representar gràficament una corba contínua definida entre dos valors és mitjançant la funció corbi(expr,from=a,to=b), on expr és el nom d'una funció o una expressió escrita com a funció de x, per exemple pnorm(x,mean,sd).

5.4. Exercicis

1. En una ciutat, s'estima que la temperatura màxima el mes de juny segueix una distribució normal, amb una mitjana de 23^{o}C i una desviació típica de 5^{o}C.

   a) Calculeu la probabilitat que un dia qualsevol la temperatura màxima se situï entre 21^{o}C i 27^{o}C.

   b) Representeu gràficament la funció de distribució i la probabilitat anteriors.

   c) Simuleu una mostra de 200 observacions i compareu-les amb la funció de distribució.

2. Se suposa que els resultats d'un examen segueixen una distribució normal amb una mitjana de 78 i una variància de 36.

   a) Quina probabilitat hi ha que una persona que es presenta a l'examen obtingui una qualificació superior a 72?

b) Representeu gràficament la funció de distribució i la probabilitat anteriors.

c) Simuleu una mostra de 10 observacions i compareu-les amb la funció de distribució.

3. Considerem X la variable aleatòria la funció de densitat de la qual és expressada per:

$$f(x) = \begin{cases} \dfrac{4}{\pi(1+x^2)} & 0 \leq x \leq 1 \\ 0 & \text{resta.} \end{cases}$$

a) Calculeu la probabilitat que X estigui entre 0.4 i 0.6.

b) Representeu gràficament la funció de distribució i la probabilitat anteriors.

c) Simuleu una mostra de 100 observacions i compareu-les amb la funció de distribució.

6
Mostreig i teorema del límit central

6.1. Introducció i objectius

Els resistors que s'utilitzen en la fabricació de productes electrònics estan etiquetats amb una resistència "nominal" i amb un percentatge de tolerància. Per exemple, es preveu que una resistència de 330 ohms (Ω) amb una tolerància del 5 % tindrà una resistència real R d'entre 313.5 Ω i 346.5 Ω distribuïda uniformement.

Si es consideren cinc resistències d'aquest tipus, seleccionades a l'atzar de la població de tots els resistors amb aquestes especificacions, que estan connectats en sèrie, la resistència total R_T del sistema és expressada per $R_T = R_1 + R_2 + ... + R_5$, on R_i són els valors de les resistències, valors que són aleatoris, independents i distribuïts amb una uniformitat idèntica. Per tant, la resistència del sistema R_T també és una variable aleatòria que té associat un valor esperat $E(R_T)$, una variància $V(R_T)$ i una funció de densitat $f(R_T)$. Però, quins són? Com es calculen? Està R_T també distribuïda uniformement? Què succeiria si en lloc de 5 tinguéssim un nombre prou gran de resistències connectades en sèrie? Què canviaria si el valor real de les resistències R no estigués distribuït uniformement sinó normalment?

En aquesta sessió, començarem formalitzant el propòsit de la inferència estadística, concretament quin comportament tenen la suma, la mitjana i la variància d'una mostra d'una variable aleatòria; en altres paraules, el mostreig aleatori i l'aplicació de la teoria de distribució de les mostres. D'aquesta manera, en les sessions següents es discutirà el problema de l'estimació dels paràmetres de la població i de les proves de contrast d'hipòtesi a partir d'una mostra donada.

En finalitzar la sessió, l'estudiant ha de ser capaç de:

- Entendre el comportament de la suma i/o mitjana d'una mostra aleatòria.
- Comprendre el teorema del límit central i les seves aplicacions en l'estimació de paràmetres.

- Simular una mostra aleatòria d'una població donada i representar els resultats fent ús de *R*.
- Comprovar per mitjà de simulacions els teoremes de mostreig i del límit central.

6.2. Mostreig

Un dels objectius més importants de l'estadística és la inferència estadística. Fer inferència sobre alguna cosa significa treure'n conclusions a partir del raonament i l'evidència. Així doncs, la inferència estadística es pot definir com el conjunt de teories, mètodes i pràctiques per formular judicis sobre els paràmetres d'una població a partir de les seves relacions estadístiques, basades en una mostra representativa d'aquesta població. En altres paraules, la inferència estadística utilitza una mostra per conèixer una cosa relacionada amb una població molt més gran. Com que la inferència es basa en les mostres, és convenient estudiar primer quin comportament té la mostra i quina relació té amb la població.

6.2.1. Mostra aleatòria

Donada una població amb variable aleatòria X (discreta o contínua), una *mostra aleatòria* és un conjunt de valors o dades aleatoris, independents i distribuïts idènticament, obtinguts a partir de la variable aleatòria X i que es distribueixen igual que aquesta.

Per exemple, $X_1, X_2, X_3, ..., X_n$ són mostres aleatòries d'una variable aleatòria X que es distribueix normalment amb una mitjana de 100 i una desviació típica de 15, si $X_1, X_2, X_3, ..., X_n$ són independents i cadascuna té una distribució normal amb una mitjana de 100 i una desviació típica de 15. Similarment, seran mostres aleatòries d'una distribució exponencial amb $\lambda = 12$ si són independents i cadascuna és exponencial amb el mateix valor de λ.

Considerem una població amb una variable aleatòria X la funció de densitat de la qual és $f_X(x)$, el seu valor esperat és $E(X)$ o μ_X i la seva variància és $V(X)$ o σ_X^2. Sigui $x_1 = \{x_{1,1}, x_{1,2}, ..., x_{1,n}\}$ una mostra aleatòria de mida n d'aquesta variable X, on $x_{1,1}$ és el valor de la variable del primer individu o objecte seleccionat, $x_{1,2}$ és el valor de la mateixa variable per al segon individu o objecte, etc. D'aquesta mostra, se'n pot visualitzar la taula de freqüències (o histograma); també se'n poden calcular alguns estadístics, com ara la suma dels elements t_1, la mitjana mostral $\bar{x}_1$ i la variància mostral s_1^2, entre d'altres, on:

$$t_1 = \sum_{i=1}^{n} x_{1,i}, \qquad \bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1,i} = \frac{t_1}{n}, \qquad S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{1,i} - \bar{x}_1\right)^2.$$

Exemple 1: En la introducció, s'esmenta que la resistència real d'uns resistors etiquetats com de 330 Ω es pot considerar una variable aleatòria R que es distribueix uniformement amb una mitjana de 330 Ω entre 313.5 Ω i 346.5 Ω. Per tant, $\mu_R = 330$ i $\sigma_R^2 = (346.5 - 313.5)^2/12 = 90.75$.

Exemple 2: La quantitat de temps que un pacient que se sotmet a un procediment especial passa en un determinat centre de cirurgia ambulatòria és una variable aleatòria W que es distribueix normalment amb un valor mitjà de 4.5 h i una desviació típica d'1.4 h. Per tant: $\mu_W = 4.5$ i $\sigma_W^2 = 1.4^2 = 1.96$.

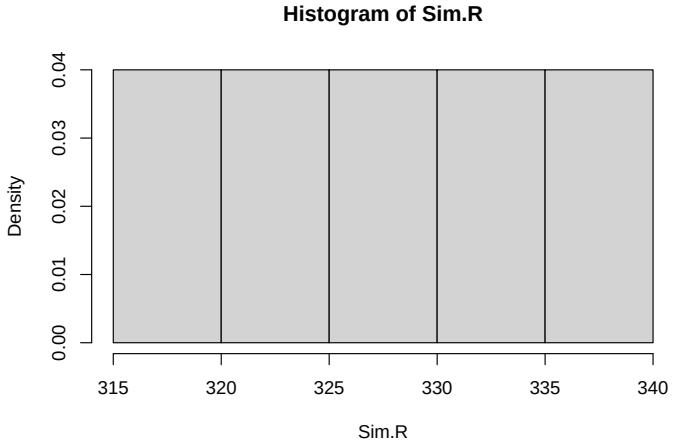
Recordem de la sessió anterior que, per simular esdeveniments o generar mostres d'una població amb una funció de densitat donada, en *R* s'utilitza la funció sample(). Si bé aquesta distribució és de les més utilitzades en enginyeria, hi ha funcions específiques com ara rnorm(), rbinom(), runif(), entre d'altres.

A continuació, generarem una mostra (simularem esdeveniments) per a cadascun dels exemples fixant primer la llavor per reproduir els resultats. Addicionalment, visualitzarem l'histograma i calcularem la suma dels elements de la mostra t, la mitjana mostral $\bar{x}$ i la variància mostral s^2 de cada mostra. Per a l'exemple 1, se simularà la selecció aleatòria de 5 resistències.

```
n.R = 5 # Mida de la mostra
set.seed(10) # Fixació de la llavor de l'aleatorietat
Sim.R = runif(n.R, min=313.5, max=346.5);
Sim.R # Mostra en l'exemple 1
```

```
[1] 330.2468 323.6234 327.5880 336.3724 316.3095
```

```
hist(Sim.R,prob=T) # Histograma
```



```
sum.R = sum(Sim.R); sum.R # Suma de les observacions de la mostra
```

```
[1] 1634.14
```

```
mean.R = mean(Sim.R); mean.R # Mitjana de la mostra
```

```
[1] 326.828
```

```
var.R = var(Sim.R); var.R # Variància de la mostra
[1] 56.06735
```

Per a l'exemple 2, se simula el temps de 25 pacients.

```
n.W = 25 # Mida de la mostra
set.seed(10) # Fixació de la llavor de l'aleatorietat
Sim.W = rnorm(n.W, mean=4.5, sd=1.4); Sim.W # Mostra en l'exemple 2
```

```
[1] 4.526245 4.242046 2.580137 3.661165 4.912363 5.045712
    2.808693 3.990854
[9] 2.222658 4.140930 6.042491 5.558094 4.166473 5.882423
    5.537946 4.625086
[17] 3.163079 4.226789 5.795730 5.176170 3.665165 1.440598
     3.555188 1.533314
[25] 2.728723
```

```
hist(Sim.W,prob=T) # Histograma
```


```
sum.W = sum(Sim.W); sum.W # Suma de les observacions de la mostra
```

```
[1] 101.2281
```

```
mean.W = mean(Sim.W); mean.W # Variància de la mostra
```

```
[1] 4.049123
```

```
var.W = var(Sim.W); var.W
```

```
[1] 1.740741
```

Heu notat alguna relació entre les mitjanes mostrals, les variàncies mostrals i els histogrames amb les mitjanes poblacionals, les variàncies poblacionals i les funcions de densitat de la població? Què passa si s'augmenta la mida de la mostra?

Si fem una altra mostra $x_2 = \{x_{2,1}, x_{2,2}, ..., x_{2,n}\}$, aquesta tindrà també un histograma, t_2, $\bar{t_2}$, s_2^2, etc. A causa de l'aleatorietat del mostreig, aquests histogrames no han de ser idèntics, com tampoc les sumes, ni les mitjanes, ni les variàncies mostrals. De fet, $(t_1, t_2, ...)$, $(\bar{x}_1, \bar{x}_2, ...)$ i $(s_1^2, s_2^2, ...)$ es poden considerar valors de les noves variables aleatòries T, $\bar{X}$ i S^2, respectivament.

Com que T, $\bar{X}$ i S^2 són variables aleatòries contínues, han de tenir:

- Funció de densitat: $f_T(x)$, $f_{\bar{X}}(x)$, $f_{S^2}(x)$
- Valor esperat: $E(T)$, $E(\bar{X})$, $E(S^2)$
- Variància: $V(T)$, $V(\bar{X})$, $V(S^2)$

Però, quins són? i quina relació tenen amb la distribució i els paràmetres de la població? Per resoldre aquests interrogants, estudiarem la distribució de la suma mostral (T), de la mitjana mostral $\bar{X}$ i de la variància mostral S^2. Per fer-ho, se simularà no solament una mostra (com fins ara), sinó moltes mostres (N=300), i s'emmagatzemaran en un data.frame on cada fila representi una mostra, i les columnes, les observacions.

```
#Exemple 1
N = 300 # Nombre de mostres
n.R = 5 # Mida de la mostra
min.R = 313.5 # Paràmetres de la població
max.R = 346.5
set.seed(10) # Fixació de la llavor de l'aleatorietat
samples = runif(N*n.R, min=min.R, max=max.R)
# Simulació de N * nr mostres
samples.R = as.data.frame(matrix(samples, ncol=n.R))
# Organització en un data.frame

#Exemple 2
N = 300 # Nombre de mostres
n.W = 25 # Mida de la mostra
mean.W = 4.5 # Paràmetres de la població
sd.W = 1.4
set.seed(10) # Fixació de la llavor de l'aleatorietat
samples = rnorm(N*n.W, mean=mean.W, sd=sd.W)
# Simulació de N * nr mostres
samples.W = as.data.frame(matrix(samples, ncol=n.W))
# Organització en un data.frame
```

6.2.2. Distribució de la suma mostral

Sigui $X_1, X_2, ..., X_n$ la mostra aleatòria de mida n d'una població X la funció de densitat de la qual és $f_X(x)$, el valor esperat és $E(X) = \mu_X$ i la variància és $V(X) = \sigma_X^2$. De la suma dels elements de la mostra $T = X_1 + X_2 +, ..., +X_n$ se'n pot deduir que:

- El seu valor esperat és expressat per:

$$\begin{aligned} E(T) &= E(X_1 + X_2 +, ..., +X_n) = E(X_1) + E(X_2) +, ..., +E(X_n) \\ &= E(X) + E(X) +, ..., +E(X) = n\mu_X. \end{aligned}$$

- La seva variància és expressada per:

$$\begin{aligned} V(T) &= V(X_1 + X_2 +, ..., +X_n) = V(X_1) + V(X_2) +, ..., +V(X_n) \\ &= V(X) + V(X) +, ..., +V(X) = n\sigma_X^2. \end{aligned}$$

- Si les X_i estan distribuïdes normalment, aleshores T també està distribuïda normalment, és a dir:

$$X \hookrightarrow N(\mu_X, \sigma_X^2) \qquad \Rightarrow \qquad T \hookrightarrow N(n\mu_X, n\sigma_X^2).$$

Tornant a l'exemple 1, la resistència total de la connexió en sèrie de 5 resistències seleccionades de manera aleatòria $R_T = R_1 + R_2 + ... + R_5$ té com a valor esperat $E(R_T) = n\mu_R = 5 \times 330 = 1650\ \Omega$ i la seva variància és $V(R_T) = n\sigma_R^2 = 5 \times 90.75 = 453.75\ \Omega^2$. Finalment, com que R no està distribuïda normalment, no podem assegurar quina distribució té R_T.

Per a cada mostra feta prèviament, es calcula la suma de les 5 observacions i s'emmagatzemen en el vector sum.samples.R (suma mostral); d'aquest vector se'n calcula la mitjana i la variància, i se'n visualitza l'histograma.

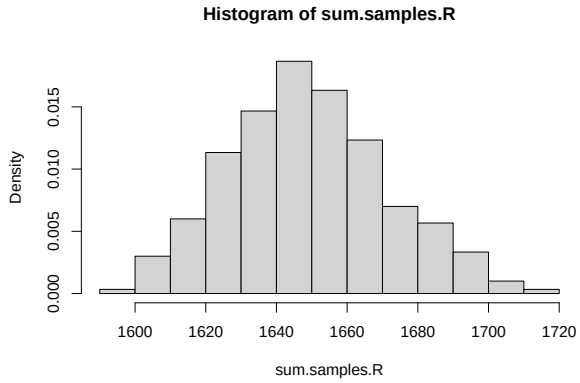
```
sum.samples.R = apply(samples.R,1,sum)
# Calcula la suma de cada mostra (fila)
mean(sum.samples.R) # Mitjana de la suma mostral
```

```
[1] 1649.235
```

```
var(sum.samples.R) # Variància de la suma mostral
```

```
[1] 499.0941
```

```
hist(sum.samples.R,prob=T) # Histograma de la suma mostral
```



D'altra banda, el temps total que 25 pacients romanen en un determinat centre de cirurgia, $W_T = W_1 + W_2 + ... + W_{25}$, té com a valor esperat $E(W_T) = n\mu_W = 25 \times 4.5 = 112.5\ s$ i la seva variància és $V(W_T) = n\sigma_W^2 = 25 \times 1.96 = 49\ s^2$. Finalment, com que W està distribuïda normalment, W_T també es distribueix normalment.

La suma de les 25 observacions s'emmagatzema en el vector sum.samples.W (suma mostral), i d'aquest vector se'n calculen la mitjana i la variància. Finalment, es visualitza l'histograma i la funció de densitat d'una normal amb una mitjana 112.5 ($n\mu_W$) i una variància 49 ($n\sigma_W^2$) o una desviació típica de 7 ($\sqrt{n}\sigma_W$).

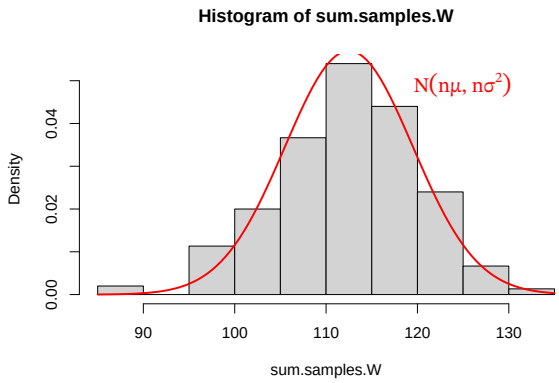
```
sum.samples.W = apply(samples.W,1,sum)
# Calcula la suma de cada mostra (fila)
mean(sum.samples.W) # Mitjana de la suma mostral
```

```
[1] 112.3905
```

```
var(sum.samples.W) # Variància de la suma mostral
```

```
[1] 59.88336
```

```
hist(sum.samples.W,prob=T) # Histograma de la suma mostral
curve(dnorm(x,mean=n.W*mean.W,sd=sqrt(n.W)*sd.W),
add=T, lwd=2, col="red")
text(125,0.049,expression(N(n*mu,n*sigma^2)),col="red",cex=1.3)
```



6.2.3. Distribució de la mitjana mostral

Continuant amb la mostra $X_1, X_2, ..., X_n$ d'una població X, i tenint en compte els resultats de l'apartat anterior, de la mitjana mostral

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{T}{n},$$

se'n pot deduir que:

- El seu valor esperat és expressat per:

$$E(\bar{X}) = E\left(\frac{T}{n}\right) = \frac{1}{n}E(T) = \frac{1}{n}n\mu_X = \mu_X.$$

- La seva variància és expressada per:

$$V(\bar{X}) = V\left(\frac{T}{n}\right) = \frac{1}{n^2}V(T) = \frac{1}{n^2}n\sigma_X^2 = \frac{\sigma_X^2}{n}.$$

- Si les X_i estan distribuïdes normalment, aleshores $\bar{X}$ també està distribuïda normalment, és a dir:

$$X \hookrightarrow N(\mu_X, \sigma_X^2) \qquad \Rightarrow \qquad \bar{X} \hookrightarrow N\left(\mu_X, \frac{\sigma_X^2}{n}\right).$$

D'aquesta manera, la resistència mitjana d'una mostra de 5 resistències seleccionades de manera aleatòria $\bar{R}$ en l'exemple 1 té com a valor esperat $E(\bar{R}) = \mu_R = 330\ \Omega$ i variància $V(\bar{R}) = \sigma_R^2/n = 90.75/5 = 18.15\ \Omega^2$. Finalment, com que R no està distribuïda normalment, no podem assegurar quina distribució té $\bar{R}$.

Com en l'apartat anterior, per a cada mostra es calcula la mitjana de les 5 observacions i aquestes mitjanes s'emmagatzemen en el vector mean.samples.R (mitjana mostral); d'aquest vector se'n calculen la mitjana i la variància, i se'n visualitza l'histograma.

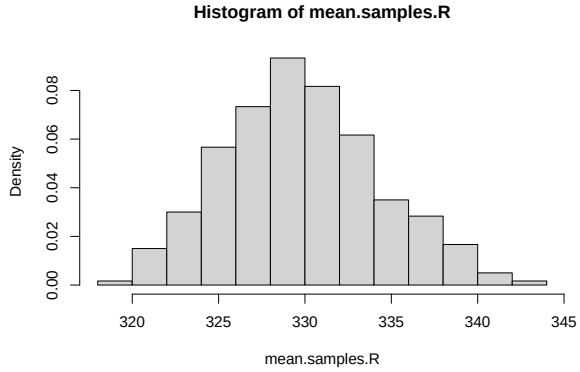
```
mean.samples.R = apply(samples.R,1,mean)
# Calcula la mitjana de cada mostra (fila)
mean(mean.samples.R) # Mitjana de la mitjana mostral
```

```
[1] 329.847
```

```
var(mean.samples.R) # Variància de la mitjana mostral
```

```
[1] 19.96376
```

```
hist(mean.samples.R,prob=T) # Histograma de la mitjana mostral
```



D'altra banda, el temps mitjà en què 25 pacients romanen en un determinat centre de cirurgia $\bar{W}$ té com a valor esperat $E(\bar{W}) = \mu_W = 4.5s$ i la seva variància és $V(\bar{W}) = \frac{\sigma_W^2}{n} = \frac{1.96}{25} = 0.0784s^2$. Finalment, com que W està distribuïda normalment, $\bar{W}$ també es distribueix normalment.

La mitjana de les 25 observacions s'emmagatzema en el vector mean.samples.W (mitjana mostral); d'aquest vector se'n calculen la mitjana i la variància. Finalment, es visualitza l'histograma i la funció de densitat d'una normal amb una mitjana de 4.5 (μ_W) i una variància de 0.0784 (σ_W^2/n) o una desviació típica de 0.28 ($\sigma_W/\sqrt{n}$).

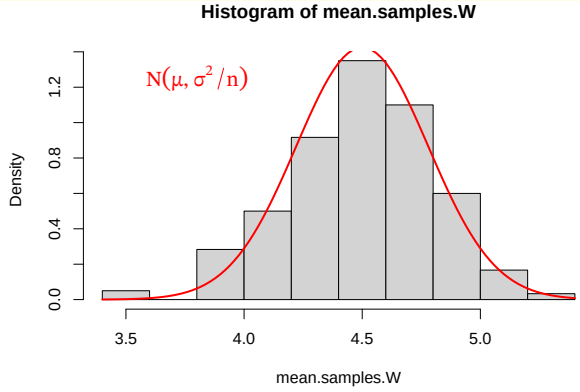
```
mean.samples.W = apply(samples.W,1,mean)
# Calcula la mitjana de cada mostra (fila)
mean(mean.samples.W) # Mitjana de la mitjana mostral
```

```
[1] 4.495621
```

```
var(mean.samples.W) # Variància de la mitjana mostral
```

```
[1] 0.09581337
```

```
hist(mean.samples.W,prob=T) # Histograma de la mitjana mostral
curve(dnorm(x,mean=mean.W,sd=sd.W/sqrt(n.W)), add=T, lwd=2, col="red")
text(3.8,1.25,expression(N(mu,sigma^2/n)),col="red",cex=1.3)
```



6.2.4. Distribució de la variància mostral

Similarment, considerant la mostra $X_1, X_2, ..., X_n$ d'una població X, i tenint en compte els resultats dels apartats anteriors, de la variància mostral

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

se'n pot deduir que:

– El seu valor esperat és expressat per:

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right) = \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X} + \bar{X}^2\right)\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2\right) - \sum_{i=1}^{n}\left(2X_i\bar{X}\right) + \sum_{i=1}^{n}\left(\bar{X}^2\right)\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2\right) - 2\bar{X}\sum_{i=1}^{n}X_i + n\bar{X}^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2\right) - 2n\bar{X}^2 + n\bar{X}^2\right) \\
&= \frac{1}{n-1}E\left(\sum_{i=1}^{n}\left(X_i^2\right) - n\bar{X}^2\right) \\
&= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n}\left(X_i^2\right)\right) - E\left(n\bar{X}^2\right)\right].
\end{aligned}$$

Tenint en compte que $E(X^2) = \sigma_X^2 + \mu_X^2$ i $E(\bar{X}^2) = \frac{\sigma_X^2}{n} + \mu_X^2$, tenim:

$$E(S^2) = \frac{1}{n-1}\left[n\sigma_X^2 + n\mu_X^2 - n\mu_X^2 - \sigma_X^2\right] = \sigma_X^2.$$

– Quan la mida de la mostra tendeix a infinit, la seva variància tendeix a zero:

$$\lim_{n\to\infty} V(S^2) = 0.$$

– Si les X_i estan distribuïdes normalment, aleshores $(n-1)\frac{S^2}{\sigma_X^2}$ està distribuïda segons la *funció khi quadrat* (χ^2) amb $(n-1)$ graus de llibertat, és a dir:

$$X \hookrightarrow N(\mu_X, \sigma_X^2) \qquad \Rightarrow \qquad (n-1)\frac{S^2}{\sigma_X^2} \hookrightarrow \chi^2_{n-1}.$$

Aquesta distribució, també denominada *distribució de Pearson* o *khi quadrada*, és una distribució de probabilitat contínua amb un paràmetre k que representa els graus de llibertat de la variable aleatòria, i la seva funció de densitat és:

$$f(x;k) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} & x > 0, \\ 0 & x \leq 0. \end{cases}$$

on Γ és la funció gamma.

Tenint en compte l'exemple 1, la variància de les resistències d'una mostra de 5 resistències seleccionades de manera aleatòria S^2 té com a valor esperat $E(S^2) = \sigma_R^2 = 90.75\Omega^2$ i, com que R no està distribuïda normalment, no podem assegurar quina distribució té S^2/σ_R^2.

Per a cada mostra simulada, es calcula la variància de les 5 observacions i aquestes s'emmagatzemen en el vector var.samples.R (variància mostral); d'aquest vector se'n calculen la mitjana i la variància, i se'n visualitza l'histograma.

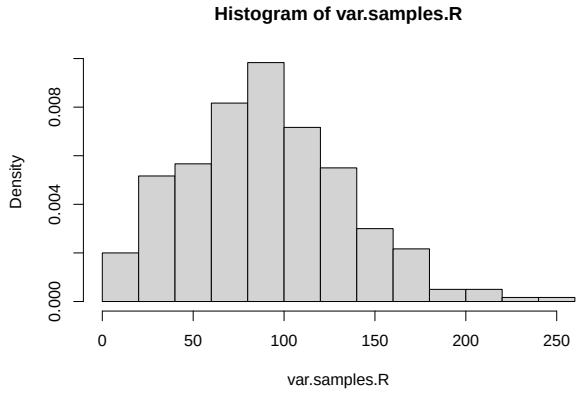
```
var.samples.R = apply(samples.R,1,var)
# Calcula la variància de cada mostra (fila)
mean(var.samples.R) # Mitjana de la variància mostral
```

```
[1] 90.39892
```

```
var(var.samples.R) # Variància de la variància mostral
```

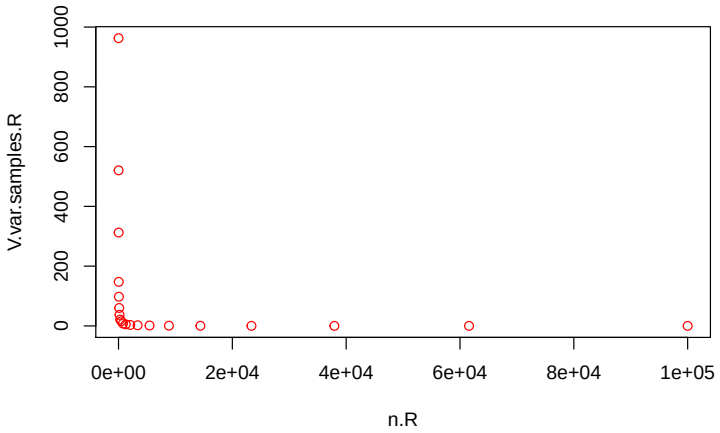
```
[1] 2014.426
```

```
hist(var.samples.R,prob=T) # Histograma de la variància mostral
```



Si es desitja comprovar què succeeix amb la variància quan la grandària de la mostra tendeix a infinit, es repeteix tot el procediment anterior per a valors de n.R entre 10 i 10^5. En el vector V.var.samples.R, s'emmagatzema la variància de la variància mostral per a cada mida de mostra emprada. Finalment, se'n visualitza la tendència en funció de la grandària mostral.

```
n.R = round(10^(seq(1,5,length=20))) # Diferents mides de mostra
V.var.samples.R = rep(0,20) # Inicialització del vector amb zeros
for (i in 1:20) # Inici de les repeticions{
    samples = runif(N*n.R[i], min=min.R, max=max.R)
    samples.R = as.data.frame(matrix(samples, ncol=n.R[i]))
    var.samples.R = apply(samples.R,1,var)
# Calcula la variància de cada mostra (fila)
V.var.samples.R[i] = var(var.samples.R) }
plot(n.R,V.var.samples.R,type="p",col="red") # Gràfic de la tendència
```



Respecte a la variància del temps emprat pels 25 pacients S^2 de l'exemple 2, tenim com a valor esperat $E(S^2) = \sigma_W^2 = 1.96s^2$.

Per a cada mostra simulada, es calcula la variància de les 25 observacions i aquestes s'emmagatzemen en el vector var.samples.W (variància mostral); d'aquest vector se'n calculen la mitjana i la variància, i se'n visualitza l'histograma.

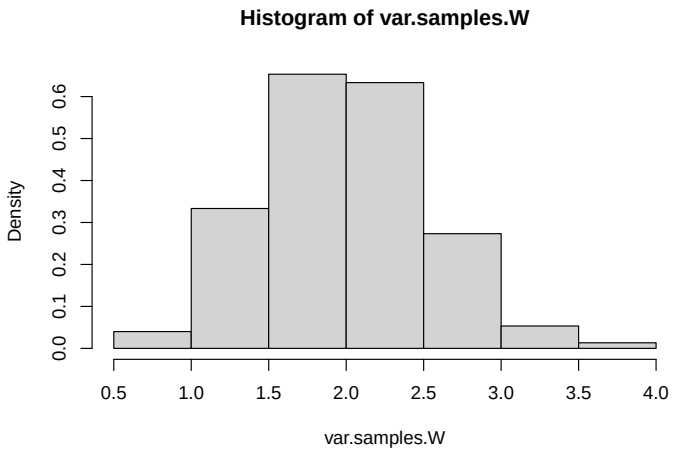
```
var.samples.W = apply(samples.W,1,var)
# Calcula la variància de cada mostra (fila)
mean(var.samples.W) # Mitjana de la variància mostral
```

```
[1] 1.99669
```

```
var(var.samples.W) # Variància de la variància mostral
```

```
[1] 0.2928742
```

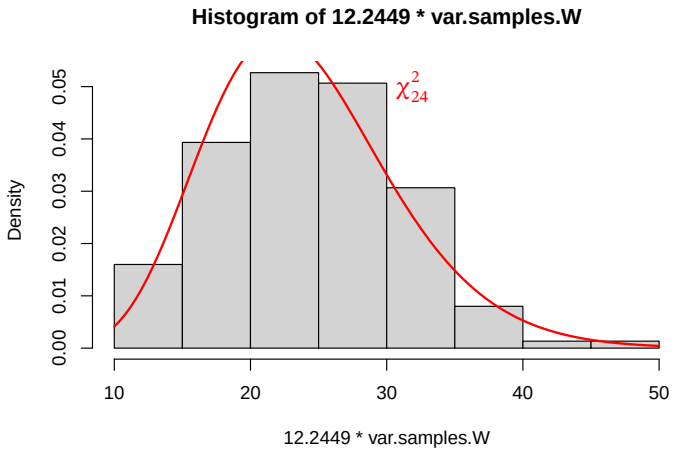
```
hist(var.samples.W,prob=T) # Histograma de la variància mostral
```



Com que W està distribuïda normalment, $(n-1)\dfrac{S^2}{\sigma_W^2}$ està distribuïda segons la funció khi quadrat (χ^2) amb 24 graus de llibertat, és a dir, $12.2449S^2 \hookrightarrow \chi^2_{24}$.

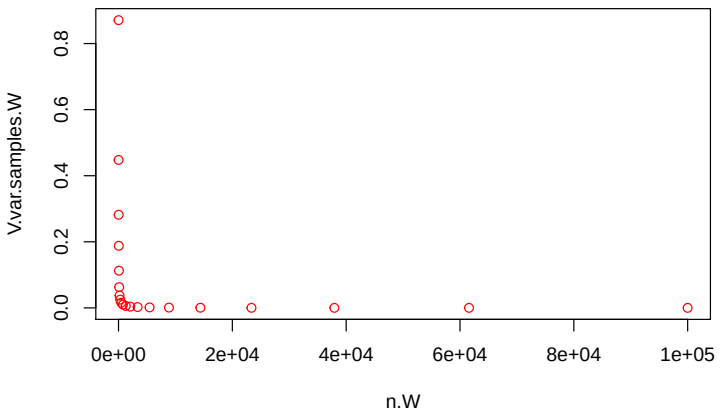
Per corroborar-ho, es representa l'histograma de la variància de la variància mostral i la funció de densitat de la funció χ^2 amb 24 graus de llibertat.

```
# Histograma de la variància mostral dividit per la poblacional
i multiplicat per n-1
hist(12.2449*var.samples.W, prob=T)
curve(dchisq(x, df=(n.W-1)), add=T, lwd=2, col="red")
text(32,0.05, expression(chi[24]^2), col="red", cex=1.3)
```



Finalment, per comprovar què succeeix amb la variància quan la mida de la mostra tendeix a infinit, es repeteix tot el procediment anterior per a valors de n.W entre 10 i 10^5. En el vector V.var.samples.W, s'emmagatzema la variància de la variància mostral per a cada mida de mostra emprada. Finalment, se'n visualitza la tendència en funció de la mida mostral.

```
n.W = round(10^(seq(1,5,length=20))) # Diferents mides de mostra
V.var.samples.W = rep(0,20) # Inicialització del vector amb zeros
for (i in 1:20) # Inici de les repeticions

    samples = rnorm(N*n.W[i], mean=mean.W, sd=sd.W)
    samples.W = as.data.frame(matrix(samples, ncol=n.W[i]))
    var.samples.W = apply(samples.W,1,var)
   # Calcula la variància de cada mostra (fila)
    V.var.samples.W[i] =

var(var.samples.W) }plot(n.W,V.var.samples.W,type="p",col="red") # Gràfic
de la tendència
```



Tips & Tricks!

- matrix(data, nrow=, ncol=) converteix les dades emmagatzemades en *data* en una matriu amb el nombre de files especificat en nrow i el nombre de columnes en ncol.
- as.data.frame() verifica si un objecte és un data.frame, o el converteix si és possible.
- apply() aplica una funció a tots els elements d'un objecte en la direcció especificada: 1 per a les files o 2 per a les columnes.
- dchisq() calcula la funció de densitat de la distribució khi quadrat (χ^2).

6.3. Teorema del límit central

En la secció anterior, s'ha estudiat el comportament probabilístic que té la suma dels elements d'una mostra, la mitjana i la seva variància. En resum, estadístics mostrals com la mitjana i la variància de les variables aleatòries. La mitjana, la variància i la suma mostral es relacionen directament amb els paràmetres de la població. No obstant això, la distribució de probabilitat d'aquestes variables aleatòries és desconeguda, a excepció del cas en què la població tingui una distribució normal.

Encara que, en aplicacions reals en enginyeria, molts processos o sistemes amb incertesa (o aleatorietat) generin variables que segueixen una distribució de probabilitat normal, altres punts no la segueixen o simplement no sabem amb certesa com es distribueixen.

D'altra banda, en estadística, molts mètodes es basen en el fet que la població es distribueix normalment, com per exemple en inferència. Per tant, si tenim un cas en què no es pot garantir la normalitat de la variable, com apliquem el mètode? Per superar aquest obstacle, recorrem a un dels resultats més notables de la teoria estadística, que és el teorema central del límit o teorema del límit central. Es considera el teorema fonamental de l'estadística, i per això porta en el seu nom la paraula "central". I s'enuncia a continuació.

Sigui $X_1, X_2, X_3, ..., X_n$ un conjunt de n variables aleatòries, independents i distribuïdes idènticament (amb la mateixa funció de densitat) amb una mitjana μ i una variància σ^2 finita. Sigui la seva suma $T_n = X_1 + X_2 + X_3 + ... + X_n$, aleshores:

$$\lim_{n\to\infty} P\left(\frac{T_n - n\mu}{\sigma\sqrt{n}} \leq z\right) = \Phi(z),$$

on $\Phi(z)$ és la funció de distribució $N(0,1)$.

Si les n variables aleatòries provenen de la mostra d'una població, la seva suma tendeix a seguir de manera asimptòtica una distribució normal, amb mitjana $n\mu$ i variància $n\sigma^2$ sempre que n sigui prou gran. D'altra banda, la mitjana mostral és expressada per:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{T}{n},$$

que segueix una distribució normal, amb una mitjana $\mu_{\bar{X}} = \mu$ i una variància $\sigma^2_{\bar{X}} = \frac{\sigma}{n}$.

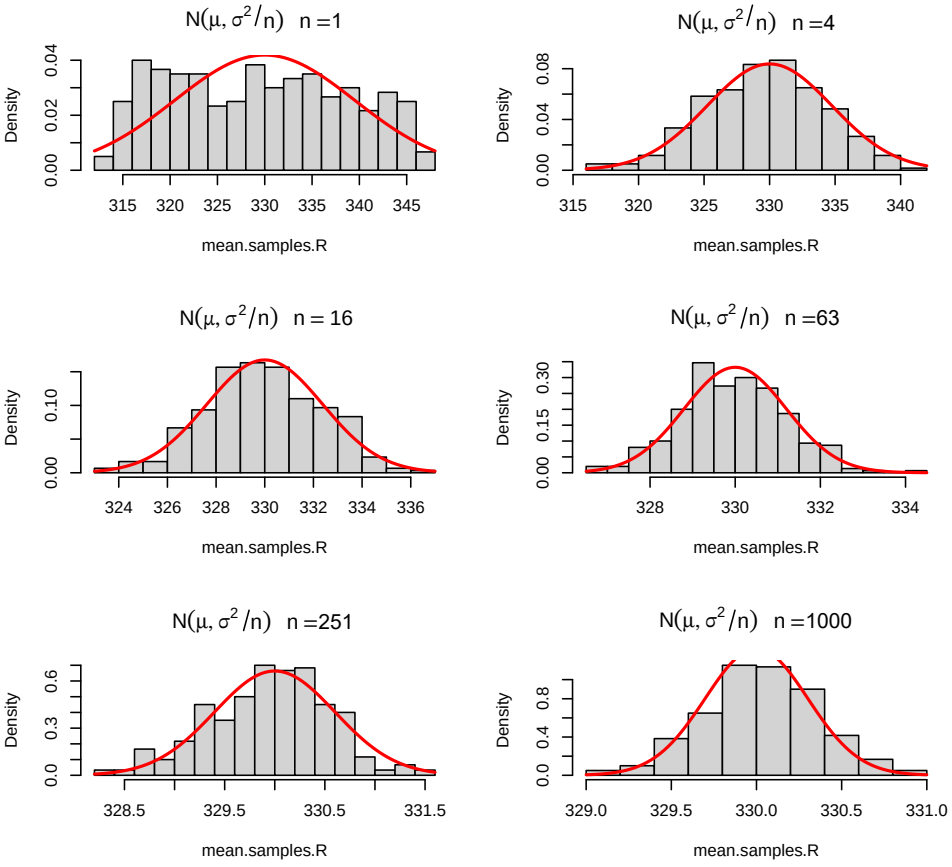
Intuïtivament, això ens indica que, en prendre una mostra aleatòria de mida prou gran d'una població amb qualsevol distribució probabilística, la distribució de la mitjana d'aquesta mostra i d'altres estadístics, com ara la proporció o la mediana, pot aproximar-se mitjançant una distribució normal, la mitjana de la qual (valor esperat)

precisament coincideix amb el valor del paràmetre poblacional. Per tant, se supera l'obstacle de no tenir una població amb distribució normal.

Suposem que es vol analitzar el consum d'energia elèctrica en una ciutat. És evident que a cada llar el consum d'electricitat es produeix de manera aleatòria. Així, per a cada llar i, es pot assignar una variable aleatòria X_i, que descriu el consum particular, amb la seva distribució de probabilitat i els paràmetres respectius (mitjana i variància). A cada llar es pot consumir energia elèctrica de manera molt diferent i, per tant, amb distribució de probabilitat diferent. No obstant això, gràcies al teorema del límit central, es pot assegurar que la quantitat total d'energia elèctrica consumida en aquesta ciutat (la suma total de nombroses llars) s'aproxima a una distribució normal.

Per corroborar el teorema del límit central, utilitzem l'exemple 1, en què la població (resistència real d'uns resistors) no es distribueix normalment, però sabem que té una mitjana de $\mu_R = 330\ \Omega$ i una variància de $\sigma_R^2 = 90.75\ \Omega^2$. Es repeteix el procediment anterior simulant diverses mostres amb diferents mides (n.R entre 1 i 1000) i es compara cadascun dels histogrames resultants amb una distribució normal de mitjana μ_R i variància σ_R^2/n.

```
N = 300 # Nombre de mostres
min.R = 313.5 # Paràmetres de la població
max.R = 346.5
mean.R = (min.R + max.R)/2
sd.R = sqrt((max.R - min.R)^2/12)
n.R = round(10^(seq(0,3,length=6))) # Diferents mides de mostra
M.mean.samples.R = rep(0,6) # Inicialització del vector amb zeros
par(mfrow=c(3,2)) # Es divideix la figura en 6
for (i in 1:6) # Inici de les repeticions{
    samples = runif(N*n.R[i], min=min.R, max=max.R)
    samples.R = as.data.frame(matrix(samples, ncol=n.R[i]))
    mean.samples.R = apply(samples.R,1,mean)
   # Calcula la mitjana de cada mostra (fila)
    title = bquote(N(mu,sigma^2/n)    n==.(n.R[i]))
   # Cadena de caràcters per al títol
   hist(mean.samples.R, prob=T, breaks=12, main=title)
   # Histograma de la mitjana mostral
   curve(dnorm(x,mean=mean.R,sd=sd.R/sqrt(n.R[i])),
   add=T, lwd=2, col="red") }
```



6.4. Exercicis

1. Es fabrica un cert tipus de fil amb una resistència a la tracció mitjana de 78,3 kg i una desviació típica de 5,6 kg.

   a) Calculeu la probabilitat que, si se selecciona aleatòriament un fil, aquest tingui una resistència a la tracció menor que 79 kg.

   b) Si se selecciona una mostra de 5 fils, calculeu la probabilitat que la mitjana de la resistència d'aquesta mostra sigui menor que 79 kg.

   c) Si la mostra és de 50 fils, calculeu la probabilitat que la mitjana de la resistència d'aquesta mostra sigui menor que 79 kg.

2. En una població, hi ha una taxa d'infecció d'una determinada malaltia d'1:100000 a l'any. Es considera X com la VAD que representa el nombre d'habitants infectats en

un nucli urbà de 3 milions d'habitants; per tant, X segueix una distribució binomial amb paràmetres $n = 3 * 10^6$ i $p = 1 \times 10^{-}5$.

a) Simuleu una mostra de 10 habitants; calculeu-ne la mitjana, la variància i l'histograma de resultats. Es pot aproximar a una distribució normal? Si la resposta és afirmativa, quins són els paràmetres de la distribució?

b) Repetiu el numeral anterior amb una mostra de 200 habitants.

3. Els pesos dels homes adults d'una determinada població es distribueixen normalment, amb una mitjana de 80 kg i una desviació estàndard de 15 kg.

a) Trobeu la probabilitat que un home seleccionat a l'atzar pesi més de 85 kg. Creeu el gràfic de la funció de distribució i representeu la probabilitat calculada.

b) Simuleu una mostra de 25 observacions; calculeu-ne la mitjana, la variància i l'histograma dels resultats. Es pot aproximar a una distribució normal? Si la resposta és afirmativa, quins són els paràmetres de la distribució?

c) Un elevador en un gimnàs per a homes té un rètol que diu que el pes màxim permès és de 2125 kg. Si 25 homes seleccionats a l'atzar entren a l'elevador, quina és la probabilitat que superin el pes màxim permès? Creeu el gràfic de la funció de distribució de la suma i representeu la probabilitat calculada.

7
Estimació

7.1. Introducció i objectius

Suposem que volem analitzar la mitjana i la variància de l'altura de tots els estudiants d'una universitat. Per conèixer aquests paràmetres, hauríem de prendre l'altura de tots i cadascun dels estudiants i després calcular-ne la mitjana i la variància. Però, si això és impossible pel cost que representa, o simplement perquè la població (i, per tant, el seu espai mostral) és infinit (com en gairebé totes les aplicacions en enginyeria), només podríem fer inferències sobre aquests paràmetres a partir d'una mostra donada.

La inferència és un dels principals objectius de l'estadística i consisteix a aconseguir informació sobre paràmetres desconeguts d'una població a partir d'un conjunt de dades obtingudes d'una mostra aleatòria. Hi ha dues maneres de fer aquesta inferència: per estimació i per contrast d'hipòtesi.

L'estimació d'aquests paràmetres es pot fer de manera puntual, és a dir, se suggereix un valor d'aquest paràmetre, per exemple: $\hat{\mu}_X = 176$ cm i $\hat{\sigma}_X = 10$ cm, on el símbol "barret" ($\hat{}$) indica que no és un valor real, sinó una estimació del paràmetre de la VA X. Com que aquesta estimació és altament dependent de la mostra, tindrà un error que no seria fàcil d'interpretar. Per tant, en la major part dels casos es prefereix suggerir no un valor, sinó un interval que contingui el valor real del paràmetre, per exemple: $\hat{\mu}_X \in [174, 178]$ cm i $\hat{\sigma}_X \in [9, 11]$ cm. Com que aquest interval depèn també de la mostra (per cada mostra que s'utilitzi, tindrem un interval diferent), s'ha de garantir que una proporció significativa dels intervals calculats contingui el valor real del paràmetre; aquesta proporció es denomina *nivell de confiança de l'interval* i, per simplificar, l'interval es denomina *interval de confiança del* $100(1-\alpha)\%$.

En aquesta sessió, s'analitzen i implementen els mètodes clàssics per al càlcul de l'interval de confiança dels paràmetres més comuns: mitjana i variància. Per tant, en finalitzar, l'estudiant ha de ser capaç de:

- Comprendre el concepte d'estimació: puntual i per intervals.
- Fer en **R** l'estimació per intervals de la mitjana i la variància d'una població a partir d'una mostra.
- Comprendre la influència de la mida de la mostra en l'estimació per intervals.
- Comprendre el concepte del nivell de confiança d'un interval.

7.2. Estimació de la mitjana d'una població

Per il·lustrar el procediment per a l'estimació de la mitjana poblacional, s'utilitza un conjunt de dades d'una mostra de 237 estudiants d'estadística en una universitat australiana. Aquest conjunt de dades, anomenat *survey*, pertany al paquet MASS, que està inclòs en la instal·lació bàsica de **R**, però s'ha de carregar amb anterioritat de la manera següent:

```
library(MASS) # Carrega la biblioteca MASS
head(survey) # Visualitza les primeres observacions
```

```
     Sex Wr.Hnd NW.Hnd W.Hnd    Fold Pulse    Clap Exer Smoke
1 Female   18.5   18.0 Right  R on L    92    Left Some Never
2   Male   19.5   20.5  Left  R on L   104    Left None Regul
3   Male   18.0   13.3 Right  L on R    87 Neither None Occas
4   Male   18.8   18.9 Right  R on L    NA Neither None Never
5   Male   20.0   20.0 Right Neither    35   Right Some Never
6 Female   18.0   17.7 Right  L on R    64   Right Some Never
  Height      M.I       Age
1 173.00    Metric    18.250
2 177.80  Imperial    17.583
3     NA      <NA>    16.917
4 160.00    Metric    20.333
5 165.00    Metric    23.667
6 172.72  Imperial    21.000
```

Aquest conjunt de dades conté, entre altra informació, el sexe de cada estudiant seleccionat (Sex), la seva freqüència cardíaca (Pulse), la seva altura (Height) i la seva edat (Age). En el transcurs d'aquesta sessió, es faran les estimacions dels paràmetres de la variable altura, encara que el procediment és extensible a qualsevol de les variables quantitatives/numèriques del conjunt de dades. Com que alguns dels estudiants enquestats no han contestat totes les preguntes, hi ha algunes observacions amb valors no disponibles, raó per la qual hem de filtrar-los usant la funció na.omit(), convertir-los a un vector numèric amb la funció as.numeric() i desar-los per a l'ús posterior en l'objecte height.

```
height = as.numeric(na.omit(survey$Height))
# Vector de dades sense valors disponibles (NA)
```

7.2.1. Estimació puntual de la mitjana

L'objectiu d'un estimador puntual és obtenir un valor numèric únic que s'aproximi a algun paràmetre (únic i desconegut) de la població a partir d'un estadístic de la mostra (aleatòria però coneguda). D'aquesta manera, l'estadístic $\hat{\theta}$ es denomina estimador puntual del paràmetre θ.

Recordant les propietats de la mitjana mostral de la sessió anterior $E(\bar{X}) = \mu_X$, el valor que esperem que s'obtingui en calcular la mitjana d'una mostra és igual al valor de la mitjana de la població. Per tant, la mitjana de la mostra és un bon estimador de la mitjana poblacional ($\hat{\mu}_X = \bar{X}$). A més, aquest estimador posseeix propietats essencials com ara no esbiaixament, eficiència, convergència i robustesa.

```
est_mu = mean(height); est_mu
```

```
[1] 172.3809
```

Si volem anar una mica més enllà d'una estimació puntual (un sol valor plausible) de la mitjana poblacional, necessitem una manera de quantificar-ne la precisió. És a dir, necessitem definir un interval entorn d'aquesta estimació puntual amb un nivell de confiança donat, és a dir, el nivell de seguretat que tenim que l'interval inclogui el paràmetre. El mètode, encara que és general, presenta petites diferències en funció de la distribució de la població, del coneixement o desconeixement de la variància de la població i de la mida de la mostra.

7.2.2. Interval de confiança de la mitjana d'una població amb distribució normal i variància coneguda

Com que sabem que la població es distribueix normalment, d'acord amb les propietats de la mitjana mostral estudiades en la sessió anterior, podem afirmar que la mitjana mostral també es distribueix normalment i la seva variància és igual a la variància de la població dividida per n:

$$\bar{X} \hookrightarrow N\left(\mu_X, \frac{\sigma_X^2}{n}\right),$$

on μ_X és el paràmetre desconegut i, per tant, per estimar.

Per a qualsevol mostra aleatòria, els punts extrems de l'interval estimat per a la mitjana de la població amb un nivell de confiança de $(1-\alpha)\,\%$ són expressats per:

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

on $z_{\alpha/2}$ denota el $100\left(1-\frac{\alpha}{2}\right)$ percentil de la distribució normal estàndard, $\frac{\sigma}{\sqrt{n}}$ és la desviació típica de la mitjana mostral (aquí la denominem error estàndard) i $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ és el marge d'error.

Assumint que l'altura de tots els estudiants de l'exemple té una distribució normal i que la desviació típica és coneguda, $\sigma = 9.48$ cm (desviació típica poblacional), l'interval d'estimació per a la mitjana de l'altura dels estudiants amb un 95% de confiança es calcula com segueix:

```
sigma = 9.48 # Sigma coneguda
alpha = 0.05 # Atès que el nivell de confiança (1-alpha) és 0.95,
#alpha = 0.05
n = length(height) # Mida de la mostra
SE = sigma/sqrt(n); SE # Error estàndard
```

```
[1] 0.6557453
```

Atès que l'interval està centrat en $\bar{x}$, el $100(1-\alpha)\% = 95\%$ de confiança implica el 97.5 $(1-\alpha/2 = 0.975)$ percentil de la distribució normal en la cua superior. Per tant, $z_{\alpha/2}$ és expressat per qnorm(0.975) (qnorm(1-alpha/2)). Ho multipliquem per l'error estàndard SE i obtenim el marge d'error E.

```
E = qnorm(1-alpha/2)*SE; E # Marge d'error
```

```
[1] 1.285237
```

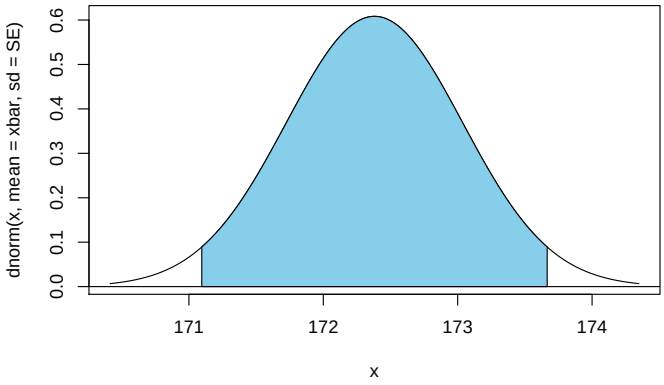
A la mitjana mostral, hi sumem i en restem aquest valor per obtenir els extrems de l'interval.

```
xbar = mean(height) # Mitjana mostral
IC = xbar + c(-E,E); IC # Interval estimat
```

```
[1] 171.0956 173.6661
```

Finalment, representem gràficament l'interval.

```
# Gràfic de la funció de densitat de la mitjana mostral
curve(dnorm(x,mean=xbar,sd=SE),from=xbar-3*SE, to=xbar+3*SE)

# Gràfic de la regió de l'interval
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dnorm(seq(IC[1],IC[2],length=100),mean=xbar,sd=SE),0)
polygon(cord.x,cord.y,col="skyblue")
abline(h=0)
```



Conclusió: Assumint la desviació típica poblacional σ com a 9.48, el marge d'error per a la mitjana de l'altura dels estudiants amb un 95% de confiança és 1.2852 centímetres, i l'interval per a la mitjana poblacional se situa entre 171.10 i 173.67 centímetres.

Fins aquest moment, hem usat la fórmula general per al càlcul de l'interval de confiança de la mitjana; no obstant això, podem aplicar la funció z.test del paquet TeachingDemos. No és un paquet que s'instal·la per defecte en **R**, sinó que s'ha d'instal·lar i carregar prèviament:

```
install.packages("TeachingDemos")
# Instal·lació del paquet (es fa solament una vegada)
```

```
library(TeachingDemos)
# Carregar el paquet (s'ha de carregar sempre que s'utilitzi)
IC = z.test(height, sd=sigma); IC
# Càlcul de l'interval de confiança per a la mitjana
```

```
 One Sample z-test

data:  height
z = 262.88, n = 209.00000, Std. Dev. = 9.48000,
Std. Dev. of the sample
mean = 0.65575, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
171.0956 173.6661
sample estimates:
mean of height
172.3809
```

Si volem canviar el nivell de confiança de l'interval, podem afegir-lo mitjançant l'opció conf.level=:

```
IC = z.test(height, sd=sigma, conf.level=0.9); IC
```

```
One Sample z-test

data:  height
z = 262.88, n = 209.00000,
Std. Dev. = 9.48000, Std. Dev. of the sample
mean = 0.65575, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
171.3023 173.4595
sample estimates:
mean of height
172.3809
```

D'acord amb els intervals calculats, es verifica que, com més gran és el nivell de confiança, més ampli és l'interval.

El resultat està emmagatzemat en una variable de tipus llista que conté, entre altres components, els següents:

- statistic: valor de l'estadístic $z = \dfrac{\bar{x}}{\sigma/\sqrt{n}}$.
- data.name: cadena de caràcters amb el nom de la variable.
- data.parameter: vector amb els graus de llibertat de l'estadístic $n - 1$, la desviació estàndard de la població σ i la mitjana mostral s.
- conf.int: vector amb els extrems de l'interval.
- estimates: mitjana mostral $\bar{x}$ (estimació puntual de la mitjana poblacional $\hat{\mu} = \bar{x}$).

D'aquesta manera, per visualitzar solament l'interval, n'hi ha prou d'executar:

```
IC$conf.int
```

```
[1] 171.3023 173.4595
attr(,"conf.level")
[1] 0.9
```

7.2.3. Interval de confiança de la mitjana d'una població amb distribució normal i variància desconeguda

En el cas anterior, si es calcula el mateix interval de confiança per a la mitjana a partir d'una altra mostra de la mateixa mida, només pot canviar el centre o punt intermedi

de l'interval ($\bar{x}$), no la seva amplitud ($2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$), ja que la variància de la població σ és coneguda i invariant.

Si aquesta variància és desconeguda (és el cas més factible en aplicacions reals), aleshores s'ha d'estimar. Un bon estimador puntual de σ és la variància de la mostra corregida $\hat{\sigma} = S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$. Aquest estimador posseeix les característiques ideals: no esbiaixament, eficiència, convergència i robustesa (coherència). Atès que aquesta estimació depèn directament de la mostra, en calcular el mateix interval de confiança per a la mitjana a partir d'una altra mostra de la mateixa mida, no solament pot canviar el centre de l'interval ($\bar{x}$), sinó també la seva amplitud ($2z_{\alpha/2}\frac{S}{\sqrt{n}}$). Per tant, per calcular l'interval de confiança no s'assumeix que $\bar{x}$ està distribuïda normalment, sinó que segueix una distribució molt similar a la normal, que es denomina *distribució t-Student*, amb $n-1$ graus de llibertat, on n és la grandària de la mostra.

La distribució de *t*-Student és similar a la distribució normal estàndard $N(0,1)$: totes dues són simètriques, amb una mitjana de zero i forma de campana; la principal diferència està en la variància. La variància de la distribució *t*-Student amb $n-1$ graus de llibertat és $V(T) = \frac{n-1}{n-2}$, és a dir, més gran que 1, que és la variància de la normal. Si observem el valor de la variància de la distribució de *t*-Student quan els graus de llibertat tendeixen a infinit, tenim:

$$\lim_{n\to\infty}\frac{n-1}{n-2} = 1,$$

és a dir, quan els graus de llibertat tendeixen a infinit, la distribució *t*-Student tendeix a la distribució normal estàndard.

Tenint en compte això, per a qualsevol mostra aleatòria, els punts extrems de l'interval estimat per a la mitjana de la població amb un nivell de confiança de $(1-\alpha)\,\%$ són expressats per:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},$$

on $t_{\alpha/2,n-1}$ denota el $100\left(1-\frac{\alpha}{2}\right)$ percentil de la distribució de *t*-Student amb $n-1$ graus de llibertat, S és la desviació típica de la mostra, $\frac{S}{\sqrt{n}}$ és la desviació típica de la mitjana mostral (error estàndard) i $t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}$ és el marge d'error.

En l'exemple donat, assumint distribució normal i variància poblacional desconeguda, l'interval de confiança al 95% utilitzant únicament els 10 primers elements de la mostra es calcula de la manera següent:

```
alpha = 0.05 # Atès que el nivell de confiança (1-alpha) és 0.95,
alpha = 0.05
height.10 = height[1:10]

# Se selecciona una mostra de 10 observacions
n = length(height.10) # Mida de la mostra
S = sd(height.10) # Desviació típica de la mostra
SE = S/sqrt(n); SE # Error estàndard
```

```
[1] 2.874441
```

Igual que en el cas anterior, el $100(1-\alpha)\% = 95\%$ de confiança implica el 97.5 ($1-\alpha/2 = 0.975$) percentil de la distribució t-Student amb 9 graus de llibertat ($n-1$) en la cua superior. Per tant, $t_{\alpha/2,n-1}$ és expressat per qt(0.975,df=9) (qt(1-alpha/2,n-1)). Ho multipliquem per l'error estàndard SE i obtenim el marge d'error E.

```
E = qt(1-alpha/2,n-1)*SE; E # Marge d'error
```

```
[1] 6.502436
```

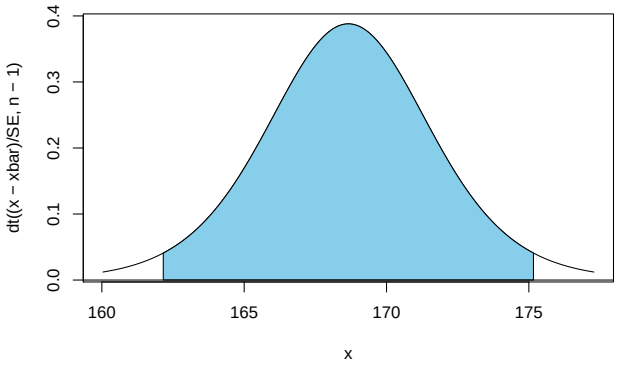
A la mitjana mostral hi sumem i en restem aquest valor per obtenir els extrems de l'interval.

```
xbar = mean(height.10) # Mitjana mostral
IC = xbar + c(-E,E); IC # Interval estimat
```

```
[1] 162.1576 175.1624
```

Finalment, representem gràficament l'interval.

```
# Gràfic de la funció de densitat de la mitjana mostral, distribució
de t-Student centrada en xbar
curve(dt((x-xbar)/SE,n-1),from=xbar-3*SE, to=xbar+3*SE)

# Gràfic de la regió de l'interval
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dt(seq((IC[1]-xbar)/SE,(IC[2]-xbar)/SE,length=100),n-1),0)
polygon(cord.x,cord.y,col="skyblue")
abline(h=0)
```



Conclusió: Desconeixent la desviació típica poblacional σ i tenint en compte les 10 primeres observacions, el marge d'error per a la mitjana de l'altura dels estudiants amb un 95% de confiança és 6.502 centímetres, i l'interval per a la mitjana poblacional es troba entre 162.16 i 175.16 centímetres. Si canviem la mostra, quins resultats s'obtenen? Calculeu diversos intervals utilitzant diferents mostres de grandària més petita que 30.

En **R** també es pot trobar la funció que calcula aquest interval, t.test, del paquet stats, que normalment està integrat en la instal·lació bàsica de **R**.

```
IC = t.test(height.10); IC # Càlcul de l'interval de confiança per a la mitjana
```

```
 One Sample t-test

data:  height.10
t = 58.676, df = 9, p-value = 6.111e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
162.1576 175.1624
sample estimates:
mean of x
168.66
```

De la mateixa manera que amb z.test, es pot canviar el nivell de confiança i l'objecte IC és una llista que conté els mateixos components.

7.2.4. Interval de confiança de la mitjana d'una població amb distribució desconeguda

Si recordem el teorema del límit central de la sessió anterior, en prendre una mostra aleatòria de mida prou elevada d'una població amb qualsevol distribució probabilística, la distribució de la mitjana d'aquesta mostra es pot aproximar mitjançant una

distribució normal que té un valor esperat que coincideix amb el valor de la mitjana poblacional.

Per tant, si tenim una mostra prou gran d'una població amb distribució desconeguda però la seva variància és coneguda, l'interval de confiança al $(1-\alpha)\%$ es calcula com a la secció 2.2:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

D'altra banda, si el valor de la variància poblacional és desconegut, aleshores l'interval es calcula com a la secció 2.3:

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}.$$

Però, com que la distribució de t-Student és similar a la normal estàndard per a alts valors de graus de llibertat, és a dir, per a una mostra prou gran (típicament $n > 30$), aleshores l'interval es pot aproximar de la manera següent:

$$\bar{x} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}.$$

Continuant amb l'exemple donat, tenint la mostra de 209 observacions i considerant que la distribució i la variància d'altura dels estudiants d'una determinada universitat és desconeguda, l'interval de confiança al 95% de l'altura mitjana d'aquests estudiants es calcula de la manera següent:

```
alpha = 0.05 # Atès que el nivell de confiança (1-alpha) és 0.95,
alpha = 0.05
n = length(height) # Mida de la mostra
S = sd(height) # Desviació típica de la mostra
SE = S/sqrt(n); SE # Error estàndard
```

```
[1] 0.6811677
```

```
E = qnorm(1-alpha/2)*SE; E # Marge d'error
```

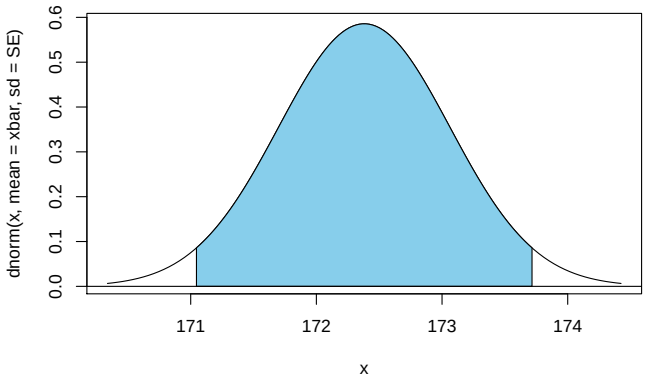
```
[1] 1.335064
```

```
xbar = mean(height) # Mitjana mostral
IC = xbar + c(-E,E); IC # Interval estimat
```

```
[1] 171.0458 173.7159
```

```
# Gràfic de la regió de l'interval
curve(dnorm(x, mean=xbar, sd=SE),from=xbar-3*SE, to=xbar+3*SE)
```

```
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dnorm(seq(IC[1],IC[2],length=100), mean=xbar, sd=SE),0)
polygon(cord.x,cord.y,col="skyblue")
abline(h=0)
```



O, utilitzant la funció z.test:

```
IC = z.test(height, sd=sd(height)); IC
```

```
 One Sample z-test

data:  height
z = 253.07, n = 209.00000, Std. Dev. = 9.84753, Std. Dev. of the sample
mean = 0.68117, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
171.0458 173.7159
sample estimates:
mean of height
172.3809
```

Tenint en compte que la distribució *t*-Student és similar a la normal estàndard per a una mostra gran, aleshores una aproximació pot ser:

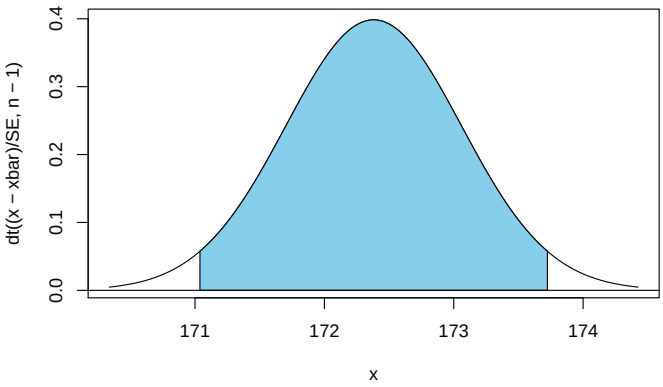
```
E = qt(1-alpha/2,n-1)*SE; E # Marge d'error
```

```
[1] 1.342878
```

```
IC = xbar + c(-E,E); IC # Interval estimat
```

```
[1] 171.0380 173.7237
```

```
# Gràfic de la regió de l'interval
curve(dt((x-xbar)/SE,n-1),from=xbar-3*SE, to=xbar+3*SE)
cord.x=c(IC[1],seq(IC[1],IC[2],length=100),IC[2])
cord.y=c(0,dt(seq((IC[1]-xbar)/SE,(IC[2]-xbar)/SE,length=100),n-1),0)
polygon(cord.x,cord.y,col="skyblue")  abline(h=0)
```



O, utilitzant la funció t.test:

```
IC = t.test(height); IC
```

```
 One Sample t-test

data:  height
t = 253.07, df = 208, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
171.0380 173.7237
sample estimates:
mean of x
172.3809
```

Conclusió: Desconeixent la distribució de la població, però amb una mostra gran ($n = 209$), el marge d'error per a la mitjana de l'altura dels estudiants amb un 95% de confiança és d'1.34 centímetres, i l'interval per a la mitjana poblacional se situa entre 171.05 i 173.72 centímetres. Noteu la petita diferència que hi ha en l'interval calculat fent ús de la normal i fent ús la distribució de t-Student. Si la mostra és petita, per exemple 10, quina és la diferència en utilitzar la distribució normal? Recordeu que aquesta aproximació només és possible si la mostra és gran.

7.2.5. Mida de la mostra

S'ha pogut observar la influència directa que té la mida de la mostra n en la longitud de l'interval de confiança resultant. La mida de la mostra necessària per complir els requisits de l'interval de confiança al $(1 - \alpha)$ % de la mitjana poblacional, és a dir, per a un marge d'error E, és expressada per:

$$n = \frac{\left(z_{\alpha/2}\right)^2 \sigma^2}{E^2}.$$

En l'exemple, sabent que la desviació estàndard poblacional σ de l'altura dels estudiants és 9.48, per trobar la mida necessària de la mostra a fi de tenir un marge d'error d'1.2 centímetres amb un 95% de confiança, es pot calcular com segueix:

```
z.alpha2 = qnorm(.975)
sigma = 9.48 E = 1.2 n = z.alpha2^2*sigma^2/ E^2
```

Conclusió: Basant-nos en el fet que coneixem la desviació estàndard poblacional, la mostra ha de tenir una mida mínima de 240 observacions per a un marge d'error d'1.2 centímetres amb un 95% de confiança. Si la variància de la població és desconeguda, es pot usar l'equació anterior?

7.2.6. Què representa el nivell de confiança?

S'ha definit en tot moment que l'estimador de l'interval té un nivell de confiança de $(1-\alpha)\,\%$. D'altra banda, es pot apreciar que, si es coneix la variància poblacional, la longitud de l'interval $\left(2z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$ només depèn de la mida de la mostra n. Per tant, si creem una altra mostra de la mateixa mida, la longitud de l'interval es manté, però la seva posició no, ja que depèn de $\bar{x}$, i aquest nou interval pot incloure o no el valor real de la mitjana de la població. Tenint en compte tot això, es pot definir que el nivell de confiança de l'interval és la proporció d'intervals estimats que incloguin el valor real del paràmetre. En altres paraules, l'interval estimat és un valor de l'interval aleatori (variable aleatòria) que té $100\alpha\%$ de probabilitat que no inclogui el valor real del paràmetre.

Per verificar-ho, suposem que l'altura mitjana de la població de l'exemple és 172 i la seva desviació estàndard és 9.48. Se simulen 100 mostres de 20 observacions cadascuna. Calculem per a cada mostra un interval de confiança del 95% ($\alpha = 0.05$) i el representem gràficament. Finalment, destaquem aquells intervals estimats que no contenen el valor real de la mitjana de la població.

```
mu = 172 ; sigma = 9.48 # Paràmetres de la població
alpha = 0.05 # Nivell de confiança
N = 100 # Nombre de mostres
n = 20 # Mida de la mostra

# Simulació de les mostres
set.seed(12) # Fixació de la llavor de l'aleatorietat
sim = rnorm(N*n, mean=mu, sd=sigma) # Simulació de N * n mostres
samples = as.data.frame(matrix(sim, ncol=n))

# Organització en un data.frame
```

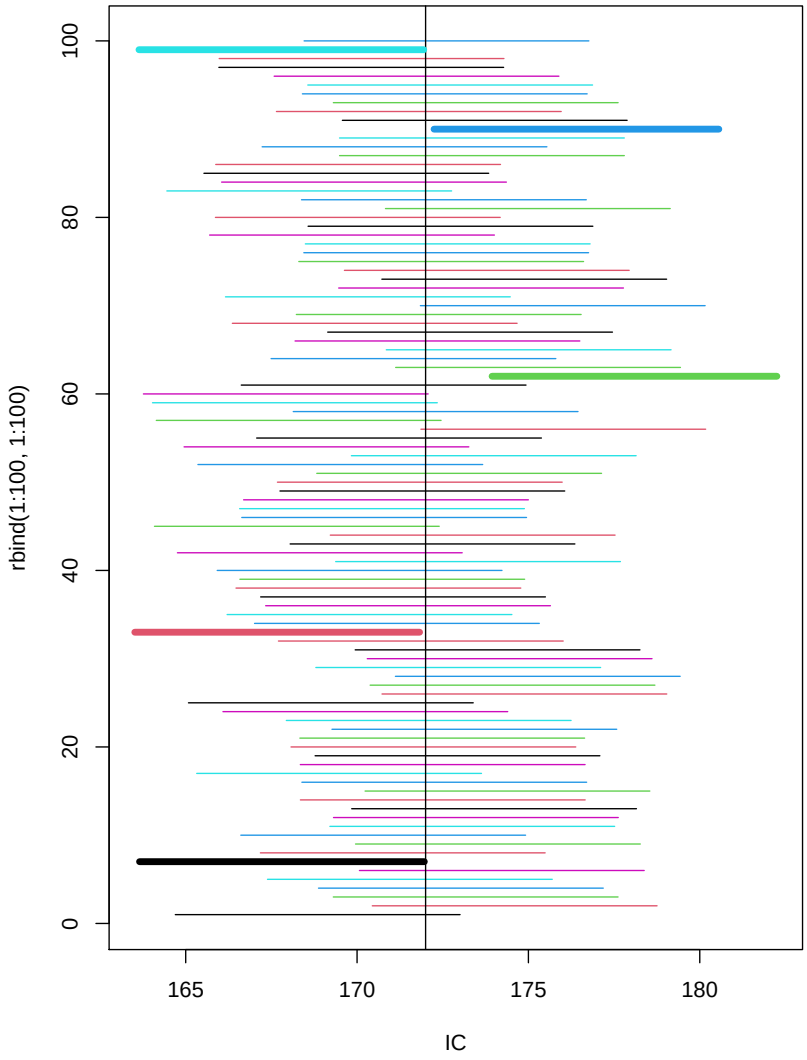
```
# Càlcul dels intervals de confiança
mean.samples = apply(samples,1,mean)

# Mitjana de cada
mostra (fila)
sd.samples = apply(samples,1,sd)

# Desviació típica de cada mostra (fila)
E = qnorm(1-alpha/2)*sigma/sqrt(n) # Marge d'error
IC = rbind(mean.samples -E,mean.samples+E) # Intervals de confiança
IC[,1:7] # Visualitza els 7 primers intervals
```

```
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
[1,] 164.6960 170.4440 169.3063 168.8746 167.3864 170.0693 163.6530
[2,] 173.0054 178.7534 177.6157 177.1840 175.6958 178.3787 171.9624
```

```
# Representació gràfica
matplot(IC,rbind(1:100,1:100),type="l",lty=1)
# Línia horitzontal per interval
abline(v=mu) # Línia vertical que representa el valor poblacional
real
out=which(!(IC[1,]<mu  mu<IC[2,]))
# Detecció d'intervals que NO contenen mu
Error = length(out);
Error # Quants intervals NO contenen el valor real
```

```
[1] 5
```

```
matplot(IC[,out],rbind(out,out),type="l",lty=1,add=T,lwd=5)
# Destaca els intervals
```



7.3. Interval de confiança per a la variància d'una població amb distribució normal

En la sessió anterior, s'ha analitzat i verificat el comportament probabilístic de la variància d'una mostra i la seva relació amb la variància de la població quan aquesta es distribueix normalment. S'ha arribat a la conclusió que la variància de la mostra és una variable aleatòria contínua, ja que depèn de la mostra seleccionada. Per tant, té un valor esperat, una variància i la relació $(n-1)\frac{S^2}{\sigma_X^2}$ està distribuïda d'acord amb la funció de khi quadrat (χ^2) amb $(n-1)$ graus de llibertat.

Tenint en compte que aquesta distribució no és simètrica al voltant de cap punt, un $(1-\alpha)$% d'interval de confiança per a la variància de la població normal significa que s'han de buscar els valors $\chi^2_{1-\frac{\alpha}{2},n-1}$ i $\chi^2_{\frac{\alpha}{2},n-1}$, de manera que:

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha,$$

on $\chi^2_{1-\frac{\alpha}{2},n-1}$ i $\chi^2_{\frac{\alpha}{2},n-1}$ denoten els $100(\alpha/2)$ i $100(1-\alpha/2)$ percentils de la distribució de khi quadrat (χ^2) amb $(n-1)$ graus de llibertat.

D'aquesta manera, l'interval de confiança per a la variància és expressat per:

$$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right).$$

Assumint que l'altura de tots els estudiants de l'exemple té una distribució normal, l'interval de confiança al 95% per a la variància de l'altura dels estudiants es calcula com segueix:

```
alpha = 0.05
n = length(height) # Grandària de la mostra
Ssq = var(height) # Variància de la mostra
chi.alpha.lower = qchisq(alpha/2, df=n-1, lower.tail = TRUE)
#100(alpha/2) percentil
chi.alpha.upper = qchisq(alpha/2, df=n-1, lower.tail = FALSE)
#100(1-alpha/2) percentil
IC = (n-1)*Ssq*c(1/chi.alpha.upper,1/chi.alpha.lower);
IC # Interval de confiança
```

```
[1]  80.73552 118.68446
```

Com que no hi ha una distribució de probabilitat concreta per a la variància mostral (només per a la relació entre les dues variàncies), no té sentit fer un gràfic de l'interval. Per al càlcul de l'interval, també es pot usar la funció sigma.test del paquet TeachingDemos.

```
library(TeachingDemos) # Carregar el paquet
IC = sigma.test(height); IC
# IC per a la variància d'una població normal
```

```
One sample Chi-squared test for variance
```

```
data:  height
X-squared = 20171, df = 208, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
80.73552 118.68446
sample estimates:
var of height
96.9738
```

Tips & Tricks!

- na.omit() retorna l'objecte eliminant les observacions amb valors que manquen.
- as.numeric() converteix objectes de qualsevol tipus a un objecte de tipus numèric.
- z.test(data, sd=) calcula l'interval de confiança per a la mitjana de la població (a partir de la mostra data) quan la variància de la població és coneguda o la mostra és gran. Pertany a la biblioteca TeachingDemos, que s'ha de carregar prèviament.
- t.test(data) calcula l'interval de confiança per a la mitjana de la població (a partir de la mostra data) quan la població es distribueix normalment amb variància desconeguda.
- sigma.test(data) calcula l'interval de confiança per a la variància d'una població que es distribueix normalment (a partir de la mostra data). Pertany a la biblioteca TeachingDemos, que s'ha de carregar prèviament.

7.4. Exercicis

1. Un fabricant de vehicles sap que el consum de gasolina dels seus vehicles es distribueix normalment. Se selecciona una mostra aleatòria simple de cotxes, se n'observa el consum cada cent quilòmetres i s'obtenen les observacions següents: 19.2, 19.4, 18.4, 18.6, 20.5, 20.8. Trobeu l'interval de confiança per al consum mitjà de gasolina de tots els vehicles d'aquest fabricant, amb un nivell de confiança del 99% i representeu-lo gràficament.

2. Un supervisor de control de qualitat en una planta enllaunadora sap que la quantitat exacta en cada llauna varia, perquè hi ha uns certs factors impossibles de controlar que afecten la quantitat d'ompliment. L'ompliment mitjà per llauna és important, però igualment important és la variació de la quantitat d'ompliment. Si

la variància és gran, algunes llaunes contindran molt poc contingut, i unes altres, massa. A fi d'estimar la variació de l'ompliment en l'enllaunadora, el supervisor tria a l'atzar 9 llaunes i pesa el contingut de cada una, i obté el pesatge següent (en unces): 7.96, 7.90, 7.98, 8.01, 7.97, 8.03, 8.02, 8.04 i 8.02. Sabent que el pes es distribueix normalment, estableix un interval de confiança del 95% per a la variància poblacional.

3. Les dades següents són les puntuacions obtingudes per a 45 persones d'una escala de depressió (puntuació més gran significa depressió més gran):

   2; 5 ; 6; 8 ; 8; 9; 9; 10; 11; 11; 11; 13; 13; 14; 14; 14; 14; 14; 14; 15; 15; 16; 16; 16; 16; 16; 16; 16; 16; 17; 17; 17; 18; 18; 18; 19; 19; 19; 19; 19; 19; 19; 19; 20; 20

   a) Trobeu l'interval de confiança per a la demanda mitjana diària amb un nivell de confiança del 80%.

   b) Trobeu l'interval de confiança per a la demanda mitjana diària amb un nivell de confiança del 95%, sabent que la desviació típica és de 4.5.

8
Contrast d'hipòtesi

8.1. Introducció i objectius

En la sessió anterior, s'ha explicat que la inferència estadística consisteix a obtenir informació sobre paràmetres desconeguts d'una població a partir d'un conjunt de dades obtingudes d'una mostra aleatòria. Aquesta informació no solament s'obté a través de l'estimació, sinó que molts dels problemes en enginyeria consisteixen en la formulació de procediments de decisió. S'exposa una informació, i el procediment basat en una mostra ens conduirà a rebutjar o acceptar aquesta hipòtesi.

Per exemple, un fabricant d'aigua embotellada afirma que cada ampolla conté 50 cl d'aigua. Com que aquesta informació figura a l'etiqueta, assumim que és certa, però, ho és?

Per respondre aquesta pregunta, podem aplicar el procediment del contrast estadístic d'hipòtesi, o simplement contrast d'hipòtesi. Es proposa una hipòtesi inicial, que és una conjectura sobre una població (mai sobre la mostra), i es contrasta amb les observacions de la mostra, és a dir, es decideix si la propietat de la població que ha estat proposada (hipòtesi) és compatible amb el que s'ha observat en la mostra d'aquesta població.

Aquest mètode està molt relacionat amb els intervals de confiança, però amb un enfocament diferent. En lloc d'estimar el paràmetre desconegut, suposem un valor i analíticament rebutgem o no la hipòtesi.

En aquesta sessió, s'analitza i implementa el procediment per al contrast d'hipòtesi de la mitjana d'una població. Per tant, en finalitzar, l'estudiant ha de ser capaç de:

- Comprendre el concepte de contrast d'hipòtesi.
- Elaborar en *R* el contrast de la mitjana d'una població a partir d'una mostra.
- Comprendre la influència de la formulació correcta de les hipòtesis.
- Comprendre el concepte del nivell de significança d'un interval.

8.2. Plantejament general del problema de contrast

En aquest apartat, es descriu el procediment general per a l'elaboració d'un contrast d'hipòtesi de la mitjana d'una població. Encara que sigui general, aquest plantejament també es pot estendre al contrast d'altres paràmetres com la proporció, la variància, etc. La diferència està principalment en l'elecció de l'estadístic de contrast i la seva distribució respectiva. Per il·lustrar-ho, al principi prendrem com a exemple el que s'ha presentat a la introducció, en què un fabricant d'aigua embotellada afirma que cada ampolla conté 50 cl d'aigua. Per definir els casos particulars, es definiran tres exemples addicionals.

8.2.1. Formular les hipòtesis

Com a punt de partida, tenim l'afirmació del fabricant que la mitjana de la població és de 50 cl. Aquesta afirmació l'anomenem *hipòtesi inicial* o *hipòtesi nul·la* i es denota típicament per H_0. Per contrastar-la, necessitem una *hipòtesi alternativa* que contradigui la nul·la, això és, que siguin mútuament excloents, que denominarem H_1.

El resultat del contrast ens portarà a decidir si les observacions (mostra) són congruents amb la hipòtesi inicial. Si no ho són, això significa que la mostra refuta la hipòtesi inicial i, per tant, s'accepta la hipòtesi alternativa. En canvi, si les observacions són congruents, no es pot demostrar que aquesta sigui certa, perquè pot existir una altra mostra que la contradigui. *En una caixa de pomes, una poma en bon estat no demostra que totes les pomes de la caixa estiguin bé, però una poma podrida sí que demostra que no totes les pomes de la caixa estan bé.*

Tenint en compte això, podem concloure que la hipòtesi nul·la es pot rebutjar, però *mai* no es pot acceptar. Per tant, la tria de la hipòtesi alternativa (que sí que s'accepta, en rebutjar la nul·la) depèn del que vulguem demostrar, i és per això que se sol anomenar *hipòtesi de recerca*.

Per exemple, si representem el consumidor d'aigua embotellada i volem *demostrar que el fabricant menteix*, suposem que la quantitat d'aigua mitjana és igual a 50 cl ($H_0 : \mu = 50$ cl) i la contrastem amb la hipòtesi que aquesta quantitat és diferent de 50 cl ($H_1 : \mu \neq 50$ cl). D'aquesta manera, el resultat del test pot ser:

- Rebutjar H_0, la qual cosa implica que acceptem que la quantitat mitjana d'aigua és *diferent de 50 cl*, que és precisament el que desitgem.

- No rebutjar H_0, la qual cosa implica que no acceptem que la quantitat d'aigua és igual ni diferent de 50 cl. Podríem dir que, amb la mostra aportada, la prova queda inconclusa.

D'altra banda, com a representants del consumidor però alhora volent demostrar que el fabricant *no solament menteix, sinó que a més la quantitat és menor*, suposem que

la quantitat d'aigua mitjana és igual a 50 cl ($H_0 : \mu = 50$ cl) i la contrastem amb la hipòtesi que aquesta quantitat és menor que 50 cl ($H_1 : \mu < 50$ cl). D'aquesta manera, el resultat del test pot ser:

- Rebutjar H_0, la qual cosa implica que acceptem que la quantitat mitjana d'aigua és *menor que 50 cl*, que és just el que desitgem.

- No rebutjar H_0, la qual cosa implica que no acceptem cap hipòtesi i la prova queda inconclusa.

Finalment, si representem el fabricant, volem demostrar que *la quantitat és fins i tot més gran que l'estipulada a l'etiqueta*. Per tant, suposem que la quantitat d'aigua mitjana és igual a 50 cl ($H_0 : \mu = 50$ cl) i la contrastem amb la hipòtesi que aquesta quantitat és més gran que 50 cl ($H_1 : \mu > 50$ cl). D'aquesta manera, el resultat del test pot ser:

- Rebutjar H_0, la qual cosa implica que acceptem que la quantitat mitjana d'aigua és *més gran que 50 cl*, que és just el que desitgem.

- No rebutjar H_0, la qual cosa implica que no acceptem cap hipòtesi i no demostrem que el fabricant menteix.

8.2.2. Especificar el nivell de significança α

Considerant el que s'ha descrit fins ara, es presenta l'existència d'un problema de rebuig/no rebuig de la hipòtesi nul·la. Això implica una possibilitat de fracassar en la decisió presa, és a dir, rebutjar H_0 quan aquesta és certa (error de tipus I) o no rebutjar-la quan realment és falsa (error de tipus II).

Com que no sabem el valor real del paràmetre μ, l'única manera de quantificar aquest error és per mitjà de probabilitats. Per tant, es defineix l'error de tipus I com la probabilitat de rebutjar una hipòtesi que en realitat és certa:

$$\alpha = P(\text{Rebutjar } H_0 | H_0 \text{ és certa}),$$

i l'error de tipus II com la probabilitat de no rebutjar una hipòtesi que en realitat és falsa:

$$\beta = P(\text{No rebutjar } H_0 | H_0 \text{ és falsa}).$$

α és el *nivell de significança del contrast*. Com que és la probabilitat de cometre un error, se li assigna un valor petit, típicament 0.05 o 0.01. D'altra banda, $1-\beta$ (probabilitat de rebutjar una hipòtesi falsa) es denomina *potència del contrast*.

8.2.3. Seleccionar el tipus de contrast

Depenent de la formulació de les hipòtesis, podem tenir tres tipus de contrastos:

Contrast bilateral (dues cues)

És el cas en què les hipòtesis nul·la i alternativa per a la prova de la mitjana poblacional es formulen de la manera següent:

$$\begin{aligned} H_0 &: \quad \mu = \mu_0, \\ H_1 &: \quad \mu \neq \mu_0. \end{aligned}$$

on μ_0 és el valor suposat de la mitjana poblacional μ. Com que l'objectiu final és decidir si la mitjana poblacional és *diferent* del valor suposat, n'hi ha prou que la mostra tingui una mitjana estadísticament diferent de μ_0, és a dir, que sigui més petita o més gran (bilateral).

Contrast unilateral inferior (cua esquerra)

És el cas en què les hipòtesis nul·la i alternativa per a la prova de la mitjana poblacional es formulen de la manera següent:

$$\begin{aligned} H_0 &: \quad \mu = \mu_0, \\ H_1 &: \quad \mu < \mu_0. \end{aligned}$$

Ara, com que l'objectiu final és decidir si la mitjana poblacional és *més petita* que el valor suposat, n'hi ha prou que la mostra tingui una mitjana estadísticament menor a μ_0 (lateral inferior).

Contrast unilateral superior (cua dreta)

És el cas en què les hipòtesis nul·la i alternativa per a la prova de la mitjana poblacional es formulen de la manera següent:

$$\begin{aligned} H_0 &: \quad \mu = \mu_0, \\ H_1 &: \quad \mu > \mu_0. \end{aligned}$$

De la mateixa manera, com que l'objectiu final és decidir si la mitjana poblacional és *més gran* que el valor suposat, n'hi ha prou que la mostra tingui una mitjana estadísticament més gran que μ_0 (lateral superior).

8.2.4. Determinar l'estadístic de contrast

En general, tot nombre que, obtingut a partir de les observacions d'una mostra, serveixi per prendre la decisió sobre rebutjar o no H_0, s'anomena *estadístic de contrast*. Com que aquesta sessió s'enfoca al contrast de la mitjana poblacional, l'estadístic de contrast és la mitjana de la mostra. Aquesta mitjana mostral s'ha de tipificar per poder fer la comparació que permeti prendre la decisió. Per tipificar una variable, recordem que se n'ha de restar la mitjana $E(\bar{X})$ i dividir per la seva desviació típica $\sigma_{\bar{X}}$, és a dir:

$$\bar{X}_{tipificada} = \frac{\bar{X} - E(\bar{X})}{\sigma_{\bar{X}}}.$$

Com ja s'ha analitzat en la sessió anterior, les propietats de la distribució de la mitjana mostral, $E(\bar{X}) = \mu$, on μ és la mitjana de la població, no les coneixem, però hem assumit en la hipòtesi inicial que és μ_0. En l'exemple de l'embotelladora d'aigua $\mu_0 = 50$. També s'ha comprovat que $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, on n és la grandària de la mostra i σ és la desviació típica de la població, que pot ser coneguda o no. En el cas que sigui desconeguda, s'ha d'estimar, o sigui, $\hat{\sigma} = S$.

Finalment, la distribució de la mitjana mostral (normal o t-Student) depèn de si la població està normalment distribuïda o no i de si es coneix la variància poblacional o no, i a més a més depèn de la mida de la mostra. Tenint en compte tot això, en el contrast d'hipòtesi de la mitjana d'una població podem tenir tres estadístics de contrast.

Població amb distribució normal i variància coneguda

Com que la població està distribuïda normalment, la mitjana mostral també ho està i, per tant:

$$\bar{X} \hookrightarrow N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \Rightarrow \qquad z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

on z_{obs} és l'estadístic de contrast.

Població amb distribució normal, variància desconeguda i mostra petita

Com que la població està distribuïda normalment, la mitjana mostral també ho està. Però, pel fet de no conèixer la variància poblacional σ, aquesta s'estima amb la variància de la mostra S. Com que aquesta estimació depèn de la mostra i de la seva mida n, la mitjana mostral no segueix exactament una distribució normal, sinó una distribució similar anomenada de t-Student, amb $n - 1$ graus de llibertat, per la qual cosa:

$$\bar{X} \hookrightarrow t_{n-1}\left(\mu, \frac{S^2}{n}\right) \qquad \Rightarrow \qquad t_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}},$$

on t_{obs} és l'estadístic de contrast.

Població amb qualsevol distribució, variància desconeguda i mostra gran

D'acord amb el teorema del límit central, independentment de la distribució de la població, pel fet de tenir una mostra gran, la distribució de la mitjana mostral es pot aproximar a una de normal. D'altra banda, el fet de no conèixer σ determina que la distribució de la mitjana mostral sigui la de t-Student amb $n-1$ graus de llibertat. No obstant això, com que la distribució de t-Student tendeix a una de normal quan els graus de llibertat tendeixen a infinit, atès que la mostra és gran, la distribució de la mitjana mostral es pot aproximar a una de normal. Per tant:

$$\bar{X} \hookrightarrow N\left(\mu, \frac{S^2}{n}\right) \qquad \Rightarrow \qquad z_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}},$$

on z_{obs} és l'estadístic de contrast.

Com s'ha vist fins ara, ens podem trobar davant nou casos possibles depenent del tipus de contrast (bilateral, lateral superior o lateral inferior) i de l'estadístic de contrast (z_{obs} a partir d'una població normal i variància coneguda, t_{obs} a partir d'una població normal, variància desconeguda i mostra petita, o z_{obs} a partir de qualsevol distribució, variància coneguda i mostra prou gran). Per continuar amb la sessió, ens basarem en tres exemples diferents, però que serveixen per entendre els nou casos possibles.

Exemple 1: Contrast bilateral d'una població normal i variància coneguda Una companyia petroliera afirma que, en un jaciment de petroli, l'àrea mitjana dels porus de la roca és de 7000 píxels. Per demostrar que el fabricant s'equivoca, es pren una mostra de 12 nuclis tallats en 4 seccions transversals i a cada secció transversal s'hi mesura l'àrea dels porus en píxels. Les dades s'emmagatzemen en l'estructura de dades data.frame anomenada *rock* de la biblioteca datasets, que està instal·lada per defecte en *R*. Suposant que aquesta àrea està distribuïda normalment amb una desviació típica de 1500 píxels, farem un contrast d'hipòtesi amb un nivell de significança de 0.05.

Les hipòtesis nul·la i alternativa del test de la mitjana poblacional són:

$$\begin{aligned} H_0 &: \quad \mu = 7000, \\ H_1 &: \quad \mu \neq 7000. \end{aligned}$$

Per tant, és un contrast bilateral i l'estadístic de contrast és:

$$z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 7000}{1500/\sqrt{48}}.$$

```
# S'inicialitza l'script
data(rock) # Carrega de dades
```

```
area = rock$area # Defineix el vector amb les dades a utilitzar
sigma.area = 1500 # Desviació típica de la població
mu.area = 7000 # Àrea mitjana que s'assumeix que és certa
alpha.area = 0.05 # Nivell de significança
```

Exemple 2: Contrast unilateral inferior d'una població normal, variància desconeguda i mostra petita El científic Edgar Anderson va recollir dades per quantificar la variació morfològica de la flor *Iris* de tres espècies relacionades: *setosa*, *virginica* i *versicolor*. Va mesurar quatre trets de cada mostra: la llargada i amplada del sèpal i del pètal, en centímetres. Mitjançant el contrast d'hipòtesi amb un nivell de significança de 0.1, demostreu que la longitud mitjana del sèpal de la varietat *setosa* és més gran que 1.5 cm, suposant que aquesta longitud es distribueix normalment. Utilitzeu les 10 primeres observacions de l'estructura de dades anomenada *Iris* de la biblioteca datasets.

Les hipòtesis nul·la i alternativa del test de la mitjana poblacional són:

$$\begin{aligned} H_0 &: \quad \mu = 1.5, \\ H_1 &: \quad \mu < 1.5. \end{aligned}$$

Per tant, és un contrast unilateral inferior i l'estadístic de contrast és:

$$t_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{\bar{x} - 1.5}{S/\sqrt{10}}.$$

```
# S'inicialitza l'script
data(iris) # Carrega les dades
sepal = iris$Petal.Length[iris$Species=="setosa"]
sepal = sepal[1:10] # Defineix el vector amb les dades a utilitzar
mu.sepal = 1.5 # Longitud mitjana que s'assumeix que és certa
alpha.sepal = 0.1 # Nivell de significança
```

Exemple 3: Contrast unilateral superior d'una població amb qualsevol distribució, variància desconeguda i mostra prou gran L'Ajuntament de Nova York afirma que la concentració d'ozó en l'aire de la ciutat és menor que 37 parts per bilió. Un grup ambientalista vol refutar aquesta afirmació mitjançant un contrast amb un nivell de significança de 0.01 i pren 116 mesures de la concentració d'ozó. Les dades s'emmagatzemen en l'estructura de dades anomenada *airquality* de la biblioteca datasets.

Les hipòtesis nul·la i alternativa del test de la mitjana poblacional són:

$$\begin{aligned} H_0 &: \quad \mu = 37, \\ H_1 &: \quad \mu > 37. \end{aligned}$$

Per tant, és un contrast unilateral superior i l'estadístic de contrast és:

$$z_{obs} = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{\bar{x} - 37}{S/\sqrt{116}}.$$

```
# S'inicialitza l'script
data("airquality") # Carrega les dadess
ozono = as.numeric(na.omit(airquality[,1])) # Defineix un vector amb
les dades a utilitzar
mu.ozono = 37 # Concentració mitjana que s'assumeix que és certa
alpha.ozono = 0.01 # Nivell de significança
```

8.2.5. Definir el criteri de decisió

La pregunta plantejada fins ara és si la mitjana de la mostra $\bar{x}$ té un valor igual, més gran o més petit (segons el tipus de contrast) que el valor que s'ha suposat per a la mitjana de la població. Aquesta és una comparació estadística o probabilística, la qual cosa implica que, tenint en compte el nivell de significança, la distribució de probabilitat i el tipus de contrast, es defineixen l'interval o els límits en què es considera si aquestes mitjanes són iguals o quina és més gran o més petita. Aquest criteri es denomina *criteri de decisió basat en el valor crític*, en què es determinen les zones de rebuig o de fallada al rebuig de la hipòtesi plantejada.

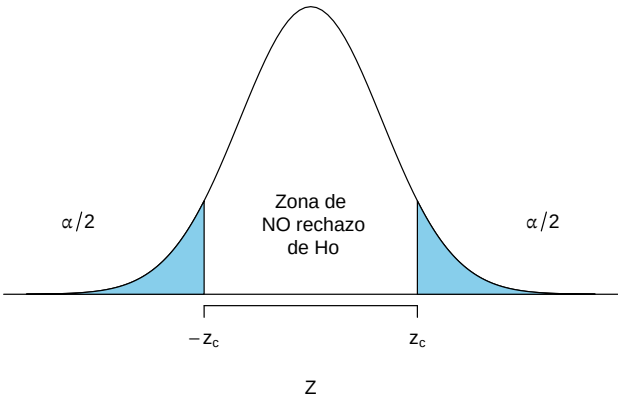
Una altra manera de prendre la decisió de la comparació és calcular la probabilitat que, si prenem una altra mostra, el seu estadístic corresponent sigui més gran o més petit, més gran o més petit (segons el tipus de contrast) que l'estadístic observat. Aquesta probabilitat es denomina *p*-valor i es compara amb el nivell de significança del contrast (α). És el *criteri de decisió basat en el p-valor*.

Com s'ha dit anteriorment, la definició de les zones de rebuig / no rebuig i el càlcul del *p*-valor depenen de la distribució de l'estadístic, entre altres factors. Per al contrast de la mitjana poblacional, l'estadístic de contrast pot tenir una distribució normal o una de *t*-Student amb $n-1$ graus de llibertat, depenent de la distribució de la població, de si es coneix o no la variància de la població i de la mida de la mostra. En qualsevol cas, la forma de la distribució és similar: la campana gaussiana. Per tant, per definir aquestes zones ens basem en el fet que l'estadístic de contrast és z_{obs}, però el resultat és igualment vàlid si l'estadístic de contrast és t_{obs}.

D'altra banda, s'ha esmentat que les zones i el *p*-valor depenen també del tipus de contrast. Així doncs, a continuació s'explica com es defineixen aquests criteris segons cada cas, utilitzant els exemples descrits anteriorment.

Criteri de decisió basat en el valor crític

Contrast bilateral (dues cues) Com que la hipòtesi alternativa és $H_1 : \mu \neq \mu_0$, i tenint en compte el nivell de significança α, els límits de l'interval de la zona de rebuig de H_0 i, per tant, de la d'acceptació de H_1, estan definits pel valor crític $\pm z_c$ tal que:



$$P(-z_c < Z < z_c) = 1 - \alpha.$$

És a dir, es rebutja H_0 quan l'estadístic observat és més petit que el valor crític negatiu ($z_{obs} < -z_c$) o més gran que el valor crític positiu ($z_{obs} > z_c$).

En l'exemple 1, tenim un contrast bilateral d'una població normal amb variància coneguda i l'estadístic de contrast és z_{obs}, ja que la mitjana mostral està distribuïda normalment. Per tant, el valor crític i la zona de no rebuig són:

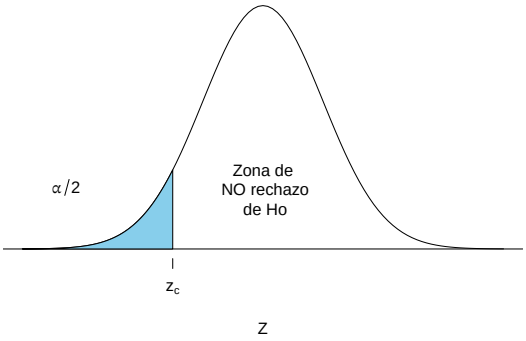
```
# Interval de no rebuig per a l'àrea mitjana dels porus de la roca
zc.area = qnorm(alpha.area/2, lower.tail=FALSE)
No_rec = c(-zc.area,zc.area); No_rec
```

```
[1] -1.959964  1.959964
```

Per tant, es rebutja H_0 si l'estadístic de la mostra tipificat z_{obs} és més petit que -1.96 o més gran que 1.96.

Contrast unilateral inferior (cua esquerra) Com que la hipòtesi alternativa és $H_1 : \mu < \mu_0$, i tenint en compte el nivell de significança α, els límits de l'interval de la zona de rebuig de H_0 i, per tant, de la d'acceptació de H_1, estan definits pel valor crític z_c tal que:

$$P(z_c < Z) = 1 - \alpha.$$



És a dir, es rebutja H_0 quan l'estadístic observat és menor que el valor crític ($z_{obs} < z_c$).

En l'exemple 2, tenim un contrast unilateral inferior d'una població normal amb variància desconeguda i l'estadístic de contrast és t_{obs}, ja que la mitjana mostral segueix una distribució de t-Student amb 9 graus de llibertat (la mida de la mostra és 10). Per tant, el valor crític i la zona de no rebuig són:

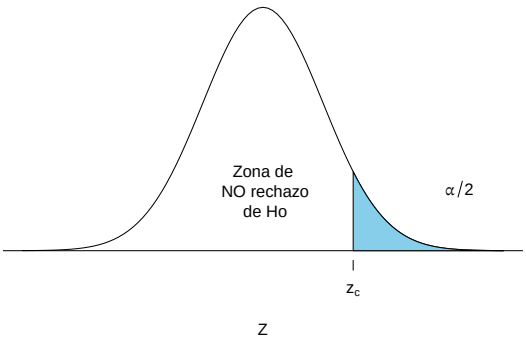
```
# Interval de no rebuig per a la longitud mitjana del sèpal
de la varietat 'setosa'
n.sepal = length(sepal) # Grandària de la mostra
tc.sepal = qt(alpha.sepal, df=n.sepal-1, lower.tail=TRUE)
No_rec = c(tc.sepal, Ïnf"); No_rec
```

```
[1] "-1.38302873839663" "Inf"
```

Per tant, es rebutja H_0 si l'estadístic de la mostra tipificat t_{obs} és menor que -1.383.

Contrast unilateral superior (cua dreta) Com que la hipòtesi alternativa és H_1 : $\mu > \mu_0$, i tenint en compte el nivell de significança α, els límits de l'interval de la zona de rebuig de H_0 i, per tant, de la d'acceptació de H_1, estan definits per un valor crític z_c tal que:

$$P(Z < z_c) = 1 - \alpha.$$



És a dir, es rebutja H_0 quan l'estadístic observat és més gran que el valor crític ($z_{obs} > z_c$).

En l'exemple 3, tenim un contrast unilateral superior d'una població amb distribució i variància desconeguda però amb una mostra prou gran. L'estadístic de contrast és z_{obs}, ja que, segons el teorema del límit central, la distribució de la mitjana mostral s'aproxima a una normal. Per tant, el valor crític i la zona de no rebuig són:

```
# Interval de no rebuig per a la concentració mitjana d'ozó
zc.ozono = qnorm(alpha.ozono, lower.tail=FALSE)
No_rec = c(-Inf", zc.ozono); No_rec
```

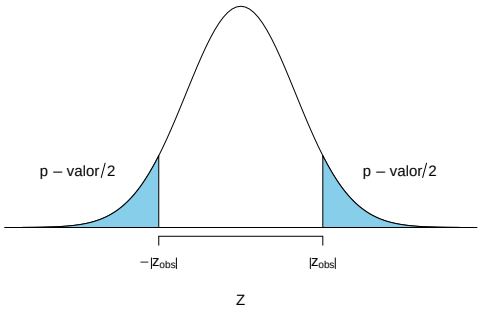
```
[1] "-Inf"     "2.32634787404084"
```

Per tant, es rebutja H_0 si l'estadístic de la mostra tipificat z_{obs} és més gran que 2.326.

Criteri de decisió basat en el p-valor

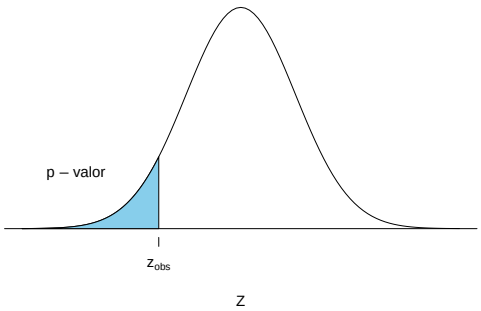
Contrast bilateral (dues cues) Com que la hipòtesi alternativa és $H_1 : \mu \neq \mu_0$, el p-valor de la mostra observada és expressat per:

$$p - valor = P(-|z_{obs}| < Z < |z_{obs}|) = 2P(Z > |z_{obs}|) = 2P(Z < -|z_{obs}|).$$



Contrast unilateral inferior (cua esquerra) Com que la hipòtesi alternativa és $H_1 : \mu < \mu_0$, el p-valor de la mostra observada és expressat per:

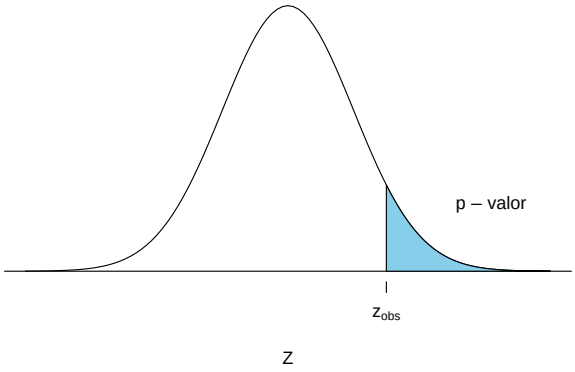
$$p - valor = P(Z < z_{obs}).$$



Contrast unilateral superior (cua dreta) Com que la hipòtesi alternativa és H_1 : $\mu > \mu_0$, el p-valor de la mostra observada és expressat per:

$$p - valor = P(Z > z_{obs}).$$

En qualsevol dels casos, es rebutja H_0 quan el p-valor és menor que el nivell de significança ($p - value < \alpha$).



8.2.6. Calcular l'estadístic observat (de la mostra) i el seu p-valor

En l'exemple 1, l'estadístic de contrast és $z_{obs} = \dfrac{\bar{x} - 7000}{1500/\sqrt{48}}$, ja que la mitjana mostral està distribuïda normalment. Tenint en compte la mostra donada, l'estadístic observat i el seu p-valor són expressats per:

```
# Estadístic i p-valor per a l'exemple 1
n.area = length(area) # Mida de la mostra
xbar.area = mean(area) # Mitjana de la mostra
zob.area = (xbar.area-mu.area)/(sigma.area/sqrt(n.area));
zob.area # Estadístic observat
```

```
[1] 0.8670839
```

```
pvalue.area = 2*pnorm(zob.area,lower.tail=FALSE);
pvalue.area # p-valor
```

```
[1] 0.3858961
```

En l'exemple 2, l'estadístic de contrast és $t_{obs} = \dfrac{\bar{x} - 1.5}{S/\sqrt{10}}$, ja que la mitjana mostral està distribuïda segons la distribució t-Student amb 9 graus de llibertat. Tenint en compte la mostra donada, l'estadístic observat i el seu p-valor són expressats per:

```
# Estadístic i p-valor per a l'exemple 2
xbar.sepal = mean(sepal) # Mitjana de la mostra
S.sepal = sd(sepal) # Desviació típica de la mostra
tob.sepal = (xbar.sepal-mu.sepal)/(S.sepal/sqrt(n.sepal)); tob.sepal
```

```
[1] -1.46385
```

```
pvalue.sepal = pt(tob.sepal,df=n.sepal-1,lower.tail=TRUE); pvalue.sepal
```

```
[1] 0.08863385
```

En l'exemple 3, l'estadístic de contrast és $z_{obs} = \dfrac{\bar{x} - 37}{S/\sqrt{116}}$, ja que, segons el teorema del límit central, la mitjana mostral s'aproxima a una distribució normal. Tenint en compte la mostra donada, l'estadístic observat i el seu p-valor són expressats per:

```
# Estadístic i p-valor per a l'exemple 3
n.ozono = length(ozono) # Mida de la mostra
xbar.ozono = mean(ozono) # Mitjana de la mostra
S.ozono = sd(ozono) # Desviació típica de la mostra
zob.ozono = (xbar.ozono-mu.ozono)/(S.ozono/sqrt(n.ozono)); zob.ozono
```

```
[1] 1.674686
```

```
pvalue.ozono = pnorm(zob.ozono, lower.tail=FALSE); pvalue.ozono
```

```
 [1] 0.04699788
```

L'estadístic observat i el p-valor del contrast d'hipòtesi per a la mitjana poblacional també es poden calcular per mitjà de les funcions z.test() de la biblioteca TeachingDemos o t.test() de la biblioteca stats que hem estudiat en la sessió anterior per al càlcul dels intervals de confiança. Per calcular l'interval, només s'introdueixen les dades de la mostra i la desviació típica de la població (només per a z.test()) i en el cas que es vulgui canviar el nivell de significança que hi ha per defecte. Ara, per efectuar el contrast, també s'ha d'estipular el valor de la mitjana per a la hipòtesi nul·la (mu=) i el tipus de contrast o la hipòtesi alternativa (alternative=), les opcions de les quals són: two.sided, less, greater. D'aquesta manera:

Exemple 1:

```
library(TeachingDemos)
z.test(area, sd=sigma.area, mu=mu.area) #Contrast per a l'exemple 1
```

```
One Sample z-test
data:  area
z = 0.86708, n = 48.00, Std. Dev. = 1500.00, Std. Dev. of the sample
```

```
mean = 216.51, p-value = 0.3859
alternative hypothesis: true mean is not equal to 7000
95 percent confidence interval:
6763.385 7612.074
sample estimates:
mean of area
7187.729
```

Entre altres coses, la funció ens retorna $z_{obs} = 0.86708$ i p-valor$= 0.3859$, que són iguals als calculats prèviament.

Ejemplo 2:

```
t.test(sepal, mu=mu.sepal, alternative="less", conf.level=0.9)
# Contrast per a l'exemple 2
```

```
One Sample t-test

data:  sepal
t = -1.4639, df = 9, p-value = 0.08863
alternative hypothesis: true mean is less than 1.5
90 percent confidence interval:
-Inf 1.497239
sample estimates:
mean of x
1.45
```

Observem també que l'estadístic observat i el p-valor coincideixen amb els calculats anteriorment, $t_{obs} = -1.4639$, p-valor$= 0.08863$.

Exemple 3:

```
t.test(ozono, mu=mu.ozono, alternative="greater", conf.level=0.99)
```

```
One Sample t-test

data:  ozono
t = 1.6747, df = 115, p-value = 0.04836
alternative hypothesis: true mean is greater than 37
99 percent confidence interval:
34.9034     Inf
sample estimates:
mean of x
42.12931
```

```
z.test(ozono, sd=S.ozono, mu=mu.ozono, alternative="greater",
conf.level=0.99)
```

```
One Sample z-test

data:  ozono
z = 1.6747, n = 116.0000,
Std. Dev. = 32.9879,
Std.Dev. of the sample
mean = 3.0628,
p-value = 0.047
alternative hypothesis: true mean is greater than 37
99 percent confidence interval:
35.00406      Inf
sample estimates:
mean of ozono
42.12931
```

Com que la mostra és gran, es pot observar la bona aproximació entre els estadístics observats i el seu p-valor ($z_{obs} = 1.6747$, p-valor$= 0.047$).

8.2.7. Rebutjar o no la hipòtesi inicial (resultat del contrast)

El resultat del contrast consisteix a rebutjar o no la hipòtesi inicial d'acord amb el criteri de decisió adoptat i l'estadístic calculat a partir de la mostra.

En l'exemple 1, tenim que $\alpha = 0.05$ i s'ha calculat $z_c = 1.96$, $z_{obs} = 0.867$ i p-valor$=$ 0.386. Si tenim en compte el criteri del valor crític, no es pot rebutjar H_0, ja que $-z_c < z_{obs} < z_c$, és a dir, z_{obs} se situa en la zona de no rebuig. D'altra banda, si considerem el criteri del p-valor, com a p-valor$> \alpha$, el resultat és el mateix: *no rebutjar* H_0.

En l'exemple 2, tenim que $\alpha = 0.1$ i s'ha calculat $t_c = -1.38$, $t_{obs} = -1.464$ i p-valor$=$ 0.089. Si tenim en compte el criteri del valor crític, es rebutja H_0, ja que $t_{obs} < t_c$, és a dir, t_{obs} se situa en la zona de rebuig. D'altra banda, si considerem el criteri del p-valor, com a p-valor$< \alpha$, el resultat és el mateix, *rebutjar* H_0. Feu el contrast prenent una altra mostra també de 10 observacions. S'obté el mateix resultat?

Finalment, en l'exemple 3, tenim que $\alpha = 0.01$ i s'ha calculat $z_c = 2.326$, $z_{obs} = 1.675$ i p-valor$= 0.047$. Si tenim en compte el criteri del valor crític, no es rebutja H_0, ja que $z_{obs} < z_c$, és a dir, z_{obs} se situa en la zona de no rebuig. D'altra banda, si considerem el criteri del p-valor, com a p-valor$> \alpha$, el resultat, com és d'esperar, és el mateix: *no rebutjar* H_0.

8.2.8. Conclusió

Per acabar, és altament recomanable fer una declaració per interpretar la decisió en el context del problema original. Si el resultat del contrast és rebutjar H_0, la conclusió és que hi ha prova suficient per rebutjar que la mitjana és igual a μ_0 (més gran que o més petita que μ_0, segons el cas). Per tant, hi ha prova suficient per defensar que la mitjana és diferent de μ_0 (més petita que o més gran que μ_0, segons el cas).

Si, per contra, el resultat és no rebutjar H_0, es conclou que no hi ha prova suficient per rebutjar que la mitjana sigui igual a μ_0 (més gran que o més petita que μ_0, segons el cas), però tampoc per defensar que la mitjana és diferent (més petita que o més gran que μ_0, segons el cas) de μ_0. En aquest supòsit, atès que no es pot afirmar res, què creieu que s'hauria de fer en una situació real?

Segons els resultats obtinguts en els exemples, es conclou el següent: per a l'exemple 1, no hi ha prova suficient per rebutjar que la mitjana de l'àrea dels porus de la roca del jaciment sigui igual a 7000 píxels, però tampoc n'hi ha per defensar que sigui diferent de 7000 píxels.

D'altra banda, en l'exemple 2 es conclou que hi ha prova suficient per rebutjar que la mitjana de la longitud del sèpal de la varietat *setosa* és igual a 1.5 cm o més gran; per tant, hi ha prova suficient per defensar que la mitjana de la longitud és menor que 1.5 cm.

Finalment, en l'exemple 3, la conclusió és que no hi ha prova suficient per rebutjar que la mitjana de la concentració d'ozó sigui igual o més petita que 37, però tampoc n'hi ha per defensar que sigui més gran que 37.

Tips & Tricks!

- En la formulació de les hipòtesis s'ha de tenir en compte:
  - Fallar en el rebuig de H_0 no significa que H_0 s'accepti com a cert.
  - Si es vol reafirmar alguna cosa, cal posar-ho en H_1.
  - Si es vol desmentir, cal posar-ho en H_0.
- z.test(data, sd=, alternative=) efectua el contrast d'hipòtesi per a la mitjana de la població (a partir de la mostra data) quan la variància de la població és coneguda o la mostra és gran. S'ha d'especificar el tipus de contrast en el paràmetre alternative = i les opcions són: two.sided, less o greater. Pertany a la biblioteca TeachingDemos, que s'ha de carregar prèviament.
- t.test(data, alternative=) efectua el contrast d'hipòtesi per a la mitjana de la població (a partir de la mostra data) quan la població es distribueix normalment amb variància desconeguda. Igualment, s'ha d'especificar el tipus de contrast.

8.3. Exercicis

1. S'especifica que un cert tipus de ferro ha de contenir 0.85 g de silici per cada 100 g de ferro (0.85%). S'ha determinat el contingut de silici (distribuït normalment) de cadascuna de les 25 mostres de ferro seleccionades a l'atzar, i s'ha trobat que la mitjana és de 0.888 i que la desviació típica és de 0.1807. És admissible que el contingut de silici és diferent del desitjat amb un nivell de significació de 0.05?

2. D'una mostra de 50 lents utilitzades en ulleres, s'obté un gruix mitjà de 3.05 mm i una desviació típica de la mostra de 0.34 mm. El gruix mitjà cert desitjat d'aquestes lents és de 3.20 mm. ¿Suggereixen les dades amb força que el gruix mitjà real de tals lents és una mica diferent del que es desitja? Demostreu-ho utilitzant un nivell de significança de 0.05.

3. Un fabricant afirma que la resistència al trencament mitjà de les seves bandes de goma és de 3 kg. Un investigador sospita que l'afirmació del fabricador és massa alta. Una mostra de 15 bandes de goma produeix una resistència mitjana al trencament de 2.6 kg i una desviació estàndard d'1.1 kg. Sabent que la resistència al trencament es distribueix normalment, en fer una prova de contrast amb un nivell de significança de 0.01, ¿decidim que l'afirmació del fabricant és massa alta? I si la prova es fa amb un nivell de significança de 0.1, què decidim?

4. Un enginyer mesura la duresa Brinell de 25 peces de ferro dúctil recuites subcríticament. Les dades resultants són:

   170 167 174 179 179 156 163 156 187 156 183 179 174 179 170 156 187 179 183 174 187 167 159 170 179

   L'enginyer planteja la hipòtesi que la duresa Brinell mitjana de totes aquestes peces de ferro dúctil és més gran que 170. Per tant, li interessa provar les hipòtesis amb un nivell de significança de 0.05. A quina conclusió arriba?