

# ESTADÍSTICA AMB R i RStudio

---

- **INSTAL·LAR R**

<https://cran.r-project.org/>

- **INSTAL·LAR R-STUDIO**

<https://rstudio.com/products/rstudio/> Download RStudio Desktop

- **CREAR UN SCRIPT D'R**

- **Des de RStudio:** *File -> New file -> R script*

- **GUARDAR L'SCRIPT**

*File -> Save as...*

- **ESPECIFICAR DIRECTORI DE TREBALL**

- **Directament des de RStudio:** *Session -> Set working directory -> choose directory*
- **També es pot fer amb codi:**  
*setwd( )* # Exemple: *setwd("C:/Bioestadistica")* Atenció: les barres són cap a la dreta /

- **ENTRAR LES DADES COM A VECTOR**

*c( )* # Exemple:

```
tractament<-c(0, 0, 0, 1, 1)
bmi<-c(21.52, 22.73, 21.89, 20.17, 24.13)
diagnostic<-c("malalt", "sa", "sa", "malalt", "sa")
```

- **CREAR UNA TAULA DE DADES (data frame):** taula de dades on la primera fila conté el nom de les columnes o variables i cada columna és del mateix tipus (numèrica o caràcter)

*data.frame( )* # Exemple: *Dades <-data.frame(diagnostic, tractament, bmi)*  
# Amb aquesta instrucció hem creat un data frame que conté 3 variables o columnes

- **IMPORTAR UN FITXER D'EXCEL**

**Des de RStudio:** *File -> Import Data Set -> From Excel*

- **IMPORTAR LES DADES DES D'UN FITXER DE TEXT O CSV**

*read.table( )* # Exemple: *Dades <-read.table("dades.txt",header=TRUE,sep=" ",dec=".")*

*read.csv( )* # Exemple: *Dades <- read.csv("dades.csv", header=T, sep=" ", dec=".");*

*# header=T* vol dir que a la primera fila del fitxer hi ha el nom de les columnes o variables  
*# sep="\t"* vol dir que els valors de cada fila estan separats per tabulacions  
*# sep=" "* vol dir que els valors de cada fila estan separats per un espai  
*# sep=","* vol dir que els valors de cada fila estan separats per comes  
*# dec="."* vol dir que el símbol decimal és un punt

- **MIRAR LES PRIMERES FILES D'UNA TAULA**

*head( )* # Exemple: *head(Dades)*

- **NOMBRE DE FILES D'UNA TAULA**

*nrow( )* # Exemple: *nrow(Dades)*

- **NOMBRE DE COLUMNES D'UNA TAULA**

*ncol( )* # Exemple: *ncol(Dades)*

- **CRIDAR UN ELEMENT D'UNA TAULA**

*taula[fila, columna]*

*Dades[2,3]* # Exemple: L'element de la fila 2 i columna 3 de la taula Dades

- **CRIDAR UNA VARIABLE D'UNA TAULA**

Per cridar una variable que està dins d'una taula cal especificar

*taula\$variable* # Exemple: La variable bmi de la taula Dades: *Dades\$bmi*

També es pot cridar especificant la columna que ocupa dins la taula:

*Dades[,3]* # bmi està a la 3a columna del data frame Dades

- **LONGITUD O NOMBRE D'ELEMENTS D'UNA VARIABLE**

*length( )* # Exemple: *length(Dades\$bmi)*

- **SELECCIONAR SUBCONJUNTS DE DADES**

# Exemple: Volem seleccionar els individus (files) que tenen un bmi>22:

*Dades[Dades\$bmi>22 , ]*

# Podem guardar aquest subconjunt de dades en un nou data frame anomenat, per exemple, "dades.bmi22" amb la següent instrucció:

*dades.bmi22 <- Dades[Dades\$bmi>22 , ]* # cal assignar amb una fletxa a l'esquerra el nom

- **DEFINIR LES VARIABLES CATEGÒRIQUES I ELS NOMS DE LES CATEGORIES**

De vegades entrem les variables categòriques amb números i hem d'indicar què significa cada valor numèric amb la funció *factor( )*:

Per exemple, en la variable *tractament*:

*tractament<-c(0, 0, 0, 1, 1)*

el valor 0 significa "placebo" i el valor 1 significa "tractat". Això ho podem indicar així:

*tractament<-factor(tractament, levels=c(0,1), labels=c("placebo", "tractat"))*

*# levels són les diferents categories*

*# labels és el nom de la categoria*

# ANÀLISI D'UNA VARIABLE

---

<b>Variable Contínua</b>	<ul style="list-style-type: none"><li>○ <b>Resums numèrics (summary statistics)</b> <i>summary( )</i> # Exemple <i>summary(bmi)</i> <i>min( )</i> # mínim <i>max( )</i> # màxim <i>mean( )</i> # mitjana (mean, average) <i>median( )</i> # mediana (median) <i>sd( )</i> # desviació típica (standard deviation) <i>IQR( )</i> # rang interquartíl·lic (interquartile rang) <i>quantile( , )</i> # percentil Ex. percentil 95%: <i>quantile(x, 0.95)</i></li><li>○ <b>Histograma</b> <i>hist( )</i> # Exemple <i>hist(bmi)</i></li><li>○ <b>Diagrama de Caixa (box plot / box-and-whisker plot)</b> <i>boxplot( )</i> # Exemple <i>boxplot(bmi)</i></li></ul>
<b>Variable Categòrica</b>	<ul style="list-style-type: none"><li>○ <b>Taula de freqüències (frequency table)</b> <i>table( )</i> # Taula freq. absolutes <i>prop.table(table( ))</i> # Taula freq. relatives <i>100*prop.table(table( ))</i> # Taula percentatges  # Exemples: <i>table(tractament)</i> <i>prop.table(table(tractament))</i> <i>100*prop.table(table(tractament))</i></li><li>○ <b>Diagrama de barres (bar plot)</b> <i>barplot(table( ))</i> # Exemple <i>barplot(table(tractament))</i></li><li>○ <b>Diagrama de sectors (pie chart)</b> <i>pie(table( ))</i> # Exemple <i>pie(table(tractament))</i></li></ul>

# RELACIÓ ENTRE DUES VARIABLES

	Relació entre dues variables
<b>Contínua &amp; contínua</b>	<ul style="list-style-type: none"><li>○ <b>Coeficient de correlació</b> <code>cor( )</code> # Exemple: <code>x&lt;-c(2, 4, 1, 3, 6, 5)</code> <code>y&lt;-c(3, 5, 2, 2, 6, 3)</code> <code>cor(x,y)</code></li><li>○ <b>Recta de regressió</b> <code>lm(y~x)</code></li><li>○ <b>Diagrama de dispersió i recta de regressió</b> <code>plot(x,y)</code> <code>abline(lm(y~x))</code></li></ul>
<b>Contínua &amp; categòrica</b>	<ul style="list-style-type: none"><li>○ <b>Resums numèrics de la variable contínua per a cada categoria de la variable categòrica</b> <code>tapply(&lt;contínua&gt;, &lt;categòrica&gt;, &lt;funció&gt; )</code> # Exemple: <code>tapply(bmi, tractament, summary)</code> <code>tapply(bmi, tractament, mean)</code></li><li>○ <b>Diagrames de caixes múltiples</b> <code>boxplot(&lt;contínua&gt;~&lt;categòrica&gt; )</code> # Exemple: <code>boxplot(bmi~tractament)</code></li></ul>
<b>Categòrica &amp; categòrica</b>	<ul style="list-style-type: none"><li>○ <b>Taula de contingència</b> <code>table( , )</code> # taula freq. absolutes <code>prop.table( )</code> # proporció total <code>prop.table( ,1)</code> # proporció fila <code>prop.table( ,2)</code> # proporció columna <code>100*prop.table( ,1)</code> # percentatge fila  # Exemples: <code>taula&lt;-table(tractament, diagnostic)</code> <code>prop.table(taula, 1)</code></li><li>○ <b>Diagrames de barres apilats</b> <code>barplot(table( ))</code> # Exemple: <code>barplot(table(tractament, diagnostic))</code></li></ul>

# Proves d'hipòtesis d'igualtat de mitjanes

<i>y variable continua x variable categòrica</i>	<b>Test de normalitat: Shapiro-Wilk</b> H0: les dades y en cada categoria segueixen una distribució normal H1: les dades y en alguna categoria no segueixen una distribució normal  <i>tapply(&lt;contínua&gt;,&lt;categòrica&gt;,function(x) shapiro.test(x))</i>	
	Si p-valor de Shapiro >0.05 <b>Les dades segueixen una distribució normal</b>	Si p-valor de Shapiro <0.05 <b>Les dades NO segueixen una distribució normal</b>
<b>Test per a una mitjana</b> H0: mitjana=valor predeterminat H1: mitjana≠valor predeterminat	Test t per a una mostra  <i>t.test(y, mu=valor)</i>	Test de Wilcoxon per a una mostra  <i>wilcox.test(y, mu=valor)</i>
<b>Test d'igualtat de dues mitjanes</b> H0: mitjana1=mitjana2 H1: mitjana1≠ mitjana2	Test t per a mostres independents <i>(cal fer prèviament el test d'igualtat de variàncies)</i>  <i>t.test(y~x,var.equal=T) # si les variàncies són iguals</i> <i>t.test(y~x,var.equal=F) # si les variàncies són diferents</i>	Test de Wilcoxon per a mostres independents  <i>wilcox.test(y~x)</i>
<b>Test d'igualtat de dues mitjanes amb dades aparellades</b> H0: mitjana1=mitjana2 H1: mitjana1≠ mitjana2	Test t per a dades aparellades  <i>d&lt;-y1-y2</i> <i>t.test(d,mu=0)</i>	Test de Wilcoxon per a dades aparellades  <i>wilcox.test(y1,y2,paired=TRUE)</i>
<b>Test d'igualtat de més de dues mitjanes</b> H0: mitjana1 = mitjana2 = mitjana3 = ... = mitjanak H1: alguna de les mitjanes és diferent	ANOVA d'un factor <i>(cal fer el bptest de variàncies)</i>  <i>summary(aov(y~x))</i>  Post-hoc analysis: <i>TukeyHSD(aov)</i>	Test de Kruskal-Wallis  <i>kruskal.test(y~x)</i>
<b>Test d'igualtat de 2 variàncies</b> H0: variància1= variància2 H1: variància1≠ variància2	Prova F d'igualtat de variàncies  <i>var.test(y~x)</i>	
<b>Test d'igualtat de més de 2 variàncies</b> H0: variàncies iguals H1: algun grup té var dif	bptest d'igualtat de variàncies  <i>bptest(lm(y ~x),studentize = FALSE)</i>	

# Altres proves d'hipòtesis

---

<b>Test d'igualtat de proporcions</b> H0: proporció1= proporció2 H1: proporció1≠ proporció2	Prova d'igualtat de dues proporcions  <i>prop.test(table(x1,x2)) # x1 i x2 són factors amb 2 categories</i>
<b>Test d'independència de dos variables categòriques</b> H0: X i Y són independents H1: X i Y estan relacionades	Prova xi-quadrat d'independència de dos factors  <i>chisq.test(table(x1,x2)) # x1 i x2 són variables categòriques</i>

## Resum models de regressió amb R

---

<b>Regressió lineal</b> Y numèrica contínua X1, X2 variables explicatives	<i>model&lt;-lm(y~x1+x2, data = data)</i> <i>summary(model)</i>  <i>Cal verificar la normalitat dels residus:</i> <i>shapiro.test(residuals(model))</i>
<b>Regressió logística</b> Y binària X1, X2 variables explicatives	<i>model&lt;-glm(y~x1+x2, data = data, family = "binomial")</i> <i>summary(model)</i>
<b>Selecció de variables en regressió</b> (step-wise regression)	<i>step(model)</i>
<b>Diagnòstics en regressió:</b> Plots de residus vs predicció	<i>plot(predict(model), residuals(model))</i> <i>abline(a=0, b=0)</i>

# VARIABLES ALEATÒRIES AMB R

$f(x)$  or  $P(X = x)$        $P(X \leq x)$        $P(X \leq q) = \alpha$

Table 3.2: Built-in-functions for random variables used in this chapter.

Distribution	parameters	density	distribution	quantiles	random sampling
Bin	$n, p$	<code>dbinom(<math>x, n, p</math>)</code>	<code>pbinom(<math>x, n, p</math>)</code>	<code>qbinom(<math>\alpha, n, p</math>)</code>	<code>rbinom(10, <math>n, p</math>)</code>
Normal	$\mu, \sigma$	<code>dnorm(<math>x, \mu, \sigma</math>)</code>	<code>pnorm(<math>x, \mu, \sigma</math>)</code>	<code>qnorm(<math>\alpha, \mu, \sigma</math>)</code>	<code>rnorm(10, <math>\mu, \sigma</math>)</code>
Chi-squared	$m$	<code>dchisq(<math>x, m</math>)</code>	<code>pchisq(<math>x, m</math>)</code>	<code>qchisq(<math>\alpha, m</math>)</code>	<code>rchisq(10, <math>m</math>)</code>
T	$m$	<code>dt(<math>x, m</math>)</code>	<code>pt(<math>x, m</math>)</code>	<code>qt(<math>\alpha, m</math>)</code>	<code>rt(10, <math>m</math>)</code>
F	$m, n$	<code>df(<math>x, m, n</math>)</code>	<code>pf(<math>x, m, n</math>)</code>	<code>qf(<math>\alpha, m, n</math>)</code>	<code>rf(10, <math>m, n</math>)</code>

- Altres distribucions:**

Geomètrica: `dgeom()`

Binomial negativa: `dnbinom()`

Poisson: `dpois()`

Hipergeomètrica: `dhyper()`

Exponencial: `dexp()`

- Exemples Binomial**

$X$  Binomial de paràmetres  $n = 8$  i  $p = 0.35$

$P(X = 4)$ : `dbinom(4, 8, 0.35)`

$P(X \leq 4)$ : `pbinom(4, 8, 0.35)`

Percentil del 95%: `qbinom(0.95, 8, 0.35)`

Mostra aleatòria de 25 valors de  $X$ : `rbinom(25, 8, 0.35)`

- Exemples distribució Normal**

$X$  Normal de paràmetres  $\mu = 10$  i  $\sigma = 3$

$P(X \leq 15)$ : `pnorm(15, 10, 3)`

$P(X > 20)$ : `1-pnorm(20, 10, 3)`

$P(12 \leq X \leq 20)$ : `pnorm(20, 10, 3) - pnorm(12, 10, 3)`

Percentil del 95%: `qnorm(0.95, 10, 3)`

Mostra aleatòria de 25 valors de  $X$ : `rnorm(25, 10, 3)`