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

Variable Contínua	○ **Resums numèrics (summary statistics)** *summary( ) # Exemple summary(bmi)* *min( ) # mínim* *max( ) # màxim* *mean( ) # mitjana (mean, average)* *median( ) # mediana (median)* *sd( ) # desviació típica (standard deviation)* *IQR( ) # rang interquartíl·lic (interquartile rang)* *quantile(,) # percentil Ex. percentil 95%: quantile(x, 0.95)* ○ **Histograma** *hist( ) # Exemple hist(bmi)* ○ **Diagrama de Caixa (box plot / box-and-whisker plot)** *boxplot( ) # Exemple boxplot(bmi)*
Variable Categòrica	○ **Taula de freqüències (frequency table)** *table( ) # Taula freq. absolutes* *prop.table(table( )) # Taula freq. relatives* *100*prop.table(table( )) # Taula percentatges* *# Exemples:* *table(tractament)* *prop.table(table(tractament))* *100*prop.table(table(tractament))* ○ **Diagrama de barres (bar plot)** *barplot(table( )) # Exemple barplot(table(tractament))* ○ **Diagrama de sectors (pie chart)** *pie(table( )) # Exemple pie(table(tractament))*

RELACIÓ ENTRE DUES VARIABLES

	Relació entre dues variables
Contínua & contínua	○ **Coeficient de correlació** ○ *cor( )* *# Exemple:* *x<-c(2, 4, 1, 3, 6, 5)* *y<-c(3, 5, 2, 2, 6, 3)* *cor(x,y)* ○ **Recta de regressió** *lm(y~x)* ○ **Diagrama de dispersió i recta de regressió** *plot(x,y)* *abline(lm(y~x))*
Contínua & categòrica	○ **Resums numèrics de la variable contínua per a cada categoria de la variable categòrica** *tapply(<continua>, <categòrica>, <funció>)* *# Exemple:* *tapply(bmi, tractament, summary)* *tapply(bmi, tractament, mean)* ○ **Diagrames de caixes múltiples** *boxplot(<contínua>~<categòrica>)* *# Exemple:* *boxplot(bmi~tractament)*
Categòrica & categòrica	○ **Taula de contingència** *table(,) # taula freq. absolutes* *prop.table( ) # proporció total* *prop.table(,1) # proporció fila* *prop.table(,2) # proporció columna* *100*prop.table(,1) # percentatge fila* *# Exemples:* *taula<-table(tractament, diagnostic)* *prop.table(taula, 1)* ○ **Diagrames de barres apilats** *barplot(table( ))* *# Exemple:* *barplot(table(tractament, diagnostic))*

Proves d'hipòtesis d'igualtat de mitjanes

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

Altres proves d'hipòtesis

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

Resum models de regressió amb R

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

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	para-meters	density	distribution	quantiles	random sampling
Bin	n, p	`dbinom`(x, n, p)	`pbinom`(x, n, p)	`qbinom`(α, n, p)	`rbinom`$(10, n, p)$
Normal	μ, σ	`dnorm`(x, μ, σ)	`pnorm`(x, μ, σ)	`qnorm` (α, μ, σ)	`rnorm`$(10, \mu, \sigma)$
Chi-squared	m	`dchisq`(x, m)	`pchisq`(x, m)	`qchisq`(α, m)	`rchisq`$(10, m)$
T	m	`dt`(x, m)	`pt`(x, m)	`qt`(α, m)	`rt`$(10, m)$
F	m,n	`df`(x, m, n)	`pf`(x, m, n)	`qf`(α, m, n)	`rf`$(10, m, n)$

- **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)`