TEMA 1: Anàlisi exploratòria de dades

1. POBLACIÓ I MOSTRA

Una **població** és un conjunt d'individus o elements que presenten característiques en comú.

Una **variable (atribut)** és la característica d'interès, que pot prendre diversos valors que canvien segons els individus.

Una **mostra** és un subconjunt de *n* individus de la població per als quals s'observa la variable o variables d'interès. El valor *n* es coneix com la **grandària de la mostra.**

Exemple

Són poblacions:

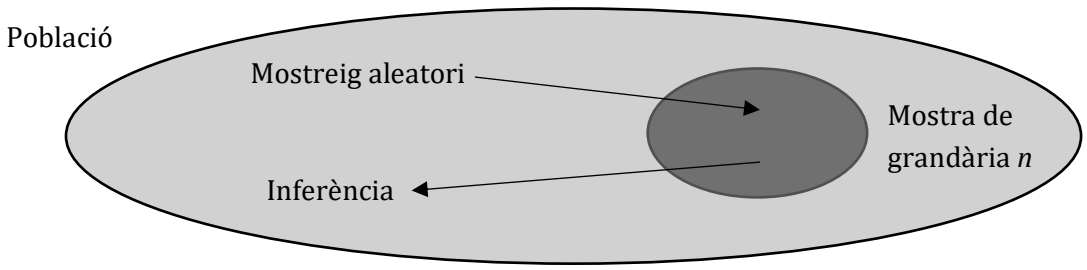
- Totes les plantacions de bedolls a Florida.
- Tots els óssos rentadors del Parc Estatal Muntanya d'Or.
- Totes les persones amb esquizofrènia als Estats Units.

Són mostres:

- *n = 8* plantacions de bedolls crescuts en un hivernacle a Florida.
- *n = 13* óssos rentadors capturats en el càmping de Muntanya d'Or.
- *n = 42* pacients amb esquizofrènia que han contestat a un determinat anunci d'un periòdic estatunidenc.

Una **mostra aleatòria simple** de *n* elements és una mostra en què:

a) Tots els individus tenen la mateixa probabilitat de ser triats.
b) Els membres de la mostra es trien independentment entre si. Això significa que l'oportunitat que un determinat membre de la població siga seleccionat no depèn que un altre haja estat ja seleccionat.



Exemple (disseny d'un experiment)

Projecte d'investigació: estudi de l'efectivitat d'un antitèrmic per a adults amb símptomes de la grip.

Experiment: se seleccionen aleatòriament 50 adults amb símptomes de la grip, se'ls administra l'antitèrmic i s'anota la temperatura (abans i després).

Població: tots els adults amb símptomes de la grip.

Mostra: els 50 adults amb símptomes de la grip seleccionats.

Variables observades: temperatura abans i temperatura després dels 50 individus de la mostra.

Variable d'interès: diferència de temperatura (abans - després) en tots el individus.

Com les nostres mostres es trien aleatòriament, sempre estarà present l'**error de mostreig**. No obstant això, si es mostreja de forma no aleatòria, l'efecte d'aquest error es pot agreujar de manera impredictible introduint un **biaix de mostreig**, que és una tendència sistemàtica que presenten certs elements de la població a ser seleccionats de forma més probable que uns altres. Els dos exemples següents il·lustren el biaix de mostreig.

Exemple (longituds de peixos)

Un biòleg planeja estudiar la distribució de la longitud del cos d'una certa població de peixos. La mostra es recull utilitzant una xarxa de pesca.

Serà llavors més probable que els peixos petits escapen pels forats de la xarxa. Per tant, serà menys probable recollir peixos petits que grans.

Exemple (Sacarosa en arrels de remolatxa)

Un agrònom planeja mostrejar arrels de remolatxa en un camp de cultiu per a mesurar el seu contingut de sacarosa. Suposem que pren tots els espècimens d'una zona petita del camp de cultiu seleccionada aleatòriament.

La mostra no reflectiria les variacions ambientals i d'entorn al llarg de tot el camp de cultiu, per tant tindríem un mostra massa homogènia.

2. TIPUS DE VARIABLES

Una **variable** és una característica d'un individu de la població a la qual podem assignar un nombre o una categoria.

Exemple

Considerem com a població tots el individus que viuen a la Comunitat Valenciana. Les variables associades a cada individu podrien ser:

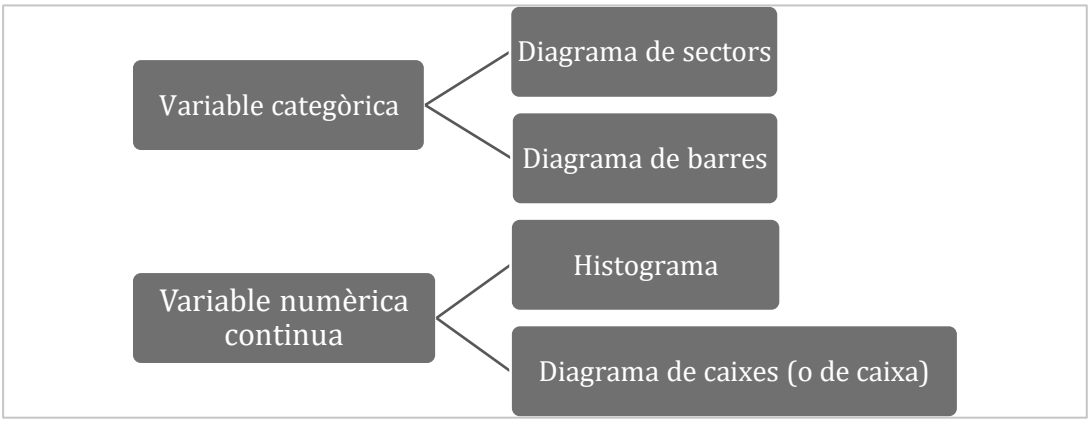
- Grup sanguini: {A, B, AB, O}.
- Nombre de fills: {0, 1, 2, ... , N}.
- Sexe: {M, F}.
- Estatura en centímetres: [50, 300]

Tipus de variables

1. Qualitatives o **categòriques** (els seus valors s'anomenen nivells o categories).
   - Sexe: {M, F}.
   - Grup sanguini: {A, B, AB, O}.
   - Fumador: {Sí, No}.
2. Quantitatives o **numèriques** (els seus valors són numèrics).
   2.1. **Discretes** (té un nombre finit de valors possibles)
   - Nombre d'ous de cranc: {0, 1, 2, ... , N}.
   - Dosis administrades en un tractament: {0, 1, 2, ...}.

   2.2. **Contínues**
   - Pes.
   - Temperatura.
   - Estatura.
   - Pressió arterial.
   - Longitud d'una arrel.

3. DESCRIPCIÓ GRÀFICA DE VARIABLES



En el cas d'una **variable numèrica discreta**:

a) Si la grandària de la mostra és petita, la podem tractar com una variable categòrica i la representarem amb un diagrama de barres.
b) Si la grandària de la mostra és gran, la tractarem com una variable numèrica continua i la representarem amb un histograma o amb un diagrama de caixa.

VARIABLE CATEGÒRICA

Abans d'explicar com es dibuixa un diagrama de sectors i un diagrama de barres hem de presentar una eina fonamental: la **taula de freqüències**. Aquesta eina ens dóna la representació numèrica d'una variable categòrica. Està formada per les freqüències absolutes i les freqüències relatives.

- Siga una mostra de *n* elements, la **freqüència absoluta** d'un valor d'una variable categòrica és el nombre de vegades que s'observa aqueix valor en la mostra. Denotem per n_i la freqüència absoluta de la categoria $i, i = 1, \dots, k$, on k és el nombre total de categories.
- La freqüència relativa d'un valor d'una variable categòrica és la proporció de vegades que s'observa aquest valor en la mostra. Es pot representar en percentatge.
  Denotem per $f_i = n_i/n$ la freqüència relativa de la categoria $i, i = 1, \dots, k$, on k és el nombre total de categories.

R-Commander:

Primer hem d'importar les dades amb les quals volem treballar:

Dades → importar → d'un arxiu Excel

Fem la taula de freqüències

Estadístics → resums → distribució de freqüències

La **moda** és el valor observat amb major freqüència.

El **diagrama de sectors** consisteix en un gràfic circular dividit en sectors. Cada sector ve donat per les freqüències relatives.

R-Commander: (amb les dades ja importades)

Gràfiques → gràfica de sectors

El **diagrama de barres** és la representació de les dades mitjançant barres (horitzontals o verticals). Cada barra correspon a una categoria.

R-Commander: (amb les dades ja importades)

Gràfiques → gràfica de barres

Recordem que les variables quantitatives discretes (nombre d'ous, nombre de fills...) es descriuen com les categòriques si prenen un nombre petit de valors. En aquest cas hem d'usar taules de freqüències per a la representació numèrica i diagrames de barres per a la gràfica. No hem d'utilitzar diagrames de sectors, ja que en aquestes variables l'ordre està present i s'ha de veure en la gràfica.

Exemple (Una mostra de grup sanguini)

Les dades observades d'un experiment han estat ($n = 362$):

0 0 B 0 B 0 A A 0 A B 0 A 0 A 0 0 0 A A A 0 0 B A A A A 0 0 A B A0 A 0 0
A AB B 0 A A A A B A A B 0 A 0 A 0 A A 0 B A B 0 0 A 0 A 0 0 0 A 0 B 0 A
A 0 A 0 B A 0 0 B B 0 B A 0 A 0 B AB A B 0 B 0 B 0 A B A 0 AB 0 A 0 B 0 0
B A 0 B A 0 A A B B AB A 0 A 0 0 B A A A B 0 0A 0 0 B A 0 A 0 0 B 0 0 A
0 A 0 B B A A B 0 0 A A 0 0 A A A 0 0 BA 0 0 B 0 0 0 0 0 0 0 B B AB B 0
A 0 A 0 B A A B B 0 A 0 0 0 B B AA B A A AB 0 B A 0 B A 0 B 0 A B B A A
B 0 0 A A 0 0 0 A A 0 A 0 0B 0 A B B A 0 B 0 0 B 0 0 0 0 0 A A A AB A B
0 B B A B 0 A B 0 A 00 A A 0 B B B B 0 0 B B A 0 0 B B B A 0 0 B 0 B A B
A A B A A 0 0A 0 A A B B 0 0 0 0 B 0 A B AB B A AB AB 0 B B B A A 0 A 0
A 0 0 B 0A A 0 0 A A 0 0 0 0 0 A 0 A 0 A 0 A 0 A 0 0 B A B B A A B 0 B B

Taula de freqüències (posem en blau l'eixida de *R-Commander*)

Categoria	n_i	f_i
0	145	40,06
A	120	33,15
B	87	24,03
AB	10	2,76

```
counts:
gs
 G1: 0  G2: A  G3: B G4: AB
    145    120     87     10

percentages:
gs
 G1: 0  G2: A  G3: B G4: AB
 40.06  33.15  24.03   2.76
```

En aquest cas la **moda** és la primera categoria: grup sanguini 0.

Diagrama de sectors

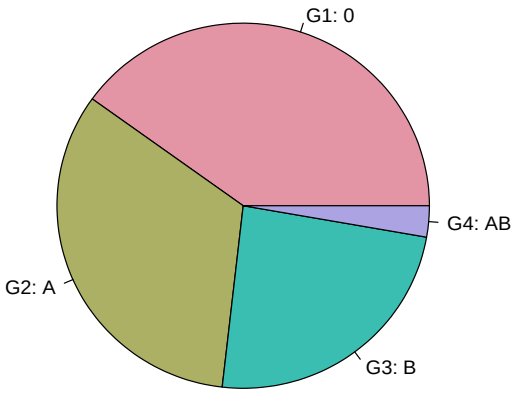
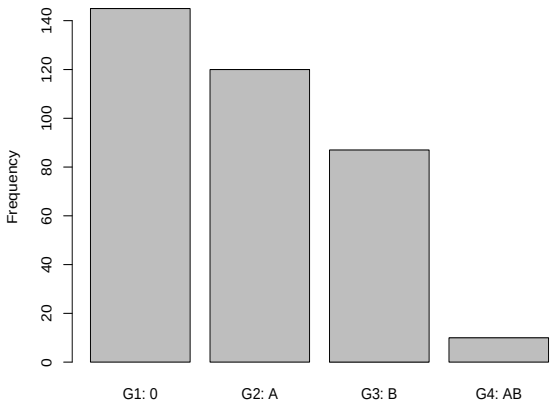


Diagrama de barres



ARIABLES QUANTITATIVES CONTÍNUES

Un histograma és una representació gràfica associada a una variable estadística contínua (o discreta amb molts valors diferents). L'histograma es fa a partir d'una taula de freqüències amb les dades agrupades en classes (o intervals), ja que tenim moltes possibilitats diferents.

R-Commander: (amb les dades ja importades)

Gràfiques → histograma

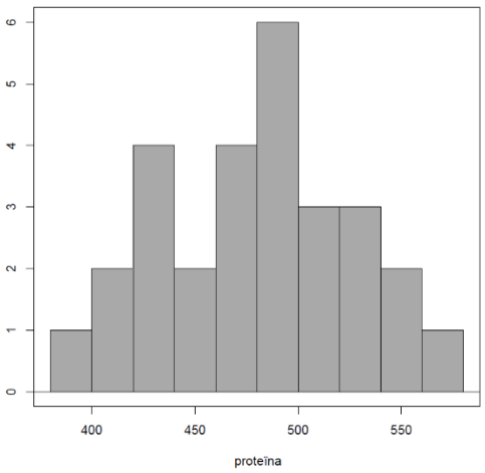
En l'exemple següent veiem clarament com es fa un histograma.

Exemple (proteïna en la llet de vaca)

S'ha analitzat la producció total anual de proteïna (en libres) de 28 vaques Holstein de dos anys d'edat. Les dades següents corresponen a la producció total anual de proteïna del 28 individus analitzats:

425	481	477	434	410	397	438	545	528	496	502
529	500	465	539	408	513	496	477	445	546	471
495	445	565	499	508	426					

El valor mínim és 397 i el màxim 565. Aagrupem les dades en deu intervals amb una amplitud de 20, ja que $(56 - 397)/10$:



Freqüències agrupades

Intervals	*Freqüència*
(380,400]	1
(400,420]	2
(420,440]	4
(440,460]	2
(460,480]	4
(480,500]	6
(500,520]	5
(520,540]	5
(540,560]	2
(560,580]	1

Ens falta explicar el diagrama de caixes com a representació gràfica d'una variable numèrica. Per fer-lo hem de presentar abans alguns conceptes.

4. DESCRIPCIÓ NUMÈRICA D'UNA MOSTRA

Per descriure numèricament una variable categòrica utilitzem les taules de freqüències. En particular, podem donar el valor de la moda, el valor observat amb major freqüència.

En el cas de variables numèriques continues hem de calcular les mesures de tendència central, la localització, la dispersió i la forma. A continuació expliquem aquestes mesures.

MESURES DE TENDÈNCIA CENTRAL

Les mesures de tendència central ens informen sobre el valor al voltant del qual tendeixen a agrupar-se les dades. Com a mesures de tendència central trobem la mitjana, la mediana i la moda.

- **Mitjana**: és la mitjana aritmètica dels valors de la mostra.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n),$$

  on x_i són les observacions (valors de la mostra) i *n* és la grandària de la mostra (nombre total d'observacions)
- **Mediana**: és el valor que ocupa la posició central en la mostra ordenada.
  - Si *n* és imparell: valor en la posició $(n+1)/2$.
  - Si *n* és parell: mitjana dels valors en les posicions $n/2$ i $n/2+1$.
- **Moda**: és el valor més freqüent en la mostra (el que es repeteix més vegades).

Encara que la mitjana i la mediana són mesures de tendència central, la mitjana és la més coneguda. Aquesta mesura utilitza totes les dades. És fàcil de calcular i té importants propietats estadístiques. Però hi ha una gran diferència entre ambdues mesures: la mitjana és molt sensible als valors extrems, mentre que la mediana no. En l'exemple següent veurem com es calculen la mitjana i la mediana i com afecten els valors extrems.

Exemple

Mostra de grandària $n = 8$:

175	164	188	176	167	158	162	182
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

Mitjana.

$$\bar{x} = \frac{1}{8}(175 + 164 + 188 + 176 + 167 + 158 + 162 + 182) = 171{,}5.$$

Mediana. Primer ordenem la mostra:

158	162	164	**167**	**175**	176	182	188

Com n és parell, calculem la mitjana entre els dos valors centrals:

$$Me = \frac{167 + 175}{2} = 171.$$

Exemple de com és sensible la mitjana als valors extrems:

Mostra 1: 158, 162, 164, 167, 175, 176, 182, 188.

Mitjana = 171,5.

Mediana = 171.

Mostra 2: 158, 162, 164, 167, 175, 176, **1820**, 188.

Mitjana = 376,25

Mediana = 171.

MESURES DE LOCALITZACIÓ

Les mesures de localització divideixen la mostra ordenada en grups. Com a mesures de localització trobem els quartils i els percentils (quantils d'ordre α).

- **Quartils**: hi ha tres quartils, per tant, divideixen la mostra en quatre parts.
  - **Q1**: és el valor que deixa per sota el 25% de la mostra.
  - **Q2**: és la mediana. Deixa per sota el 50% de la mostra.
  - **Q3**: és el valor que deixa per sota el 75% de la mostra.
- **Percentil 100α**: valor que deixa per sota el 100α% dels valors de la mostra.

MESURES DE DISPERSIÓ

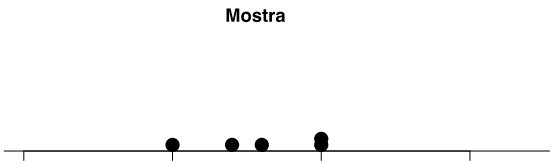
Una bona descripció de les dades hauria de caracteritzar, a més del centre, el grau de dispersió de les dades. Si dues mostres tenen la mateixa mitjana, gairebé totes les observacions de la mostra són iguals o difereixen substancialment? En l'exemple següent observem que la distribució de les observacions pot ser diferent.

Exemple

Mostra a: 70, 72, 73, 75, 75

Mitjana: 73.

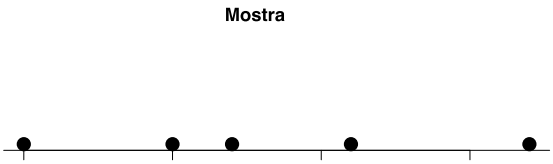
Desviació típica: 2,12.



Mostra b: 65, 70, 72, 76, 82

Mitjana: 73.

Desviació típica: 6,4.



Per tant, les mesures de dispersió indiquen valors que representen la "separació" de les dades entre si o respecte de la mitjana. Com a mesures de dispersió trobem el rang, el rang interquartílic, la variància, la desviació típica i el coeficient de variació.

- **Rang**: és la diferència entre els valors màxim i mínim.
- **Rang interquartílic**: és la diferència entre el primer i el tercer quartil (IQR = Q3 – Q1). És el valor que deixa en mig el 50% de la mostra.
- **Variància:** es calcula així

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- **Desviació típica:** és l'arrel quadrada de la variància

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- **Coeficient de variació:** és la desviació típica expressada com un percentatge de la mitjana

$$CV = \frac{s}{\bar{x}}$$

Ens permet fer comparacions de variabilitat entre diferents mostres, inclús si tenen diferents unitats.

Exemple

Mostra 1

$$\bar{x}_1 = 12\ cm,\ s_1 = 3\ cm \implies CV_1 = \frac{s_1}{\bar{x}_1} = \frac{3}{12} = 0{,}25$$

Mostra 2

$$\bar{x}_2 = 27\ kg, s_2 = 7\ kg \implies CV_2 = \frac{s_2}{\bar{x}_2} = \frac{7}{27} = 0{,}2592$$

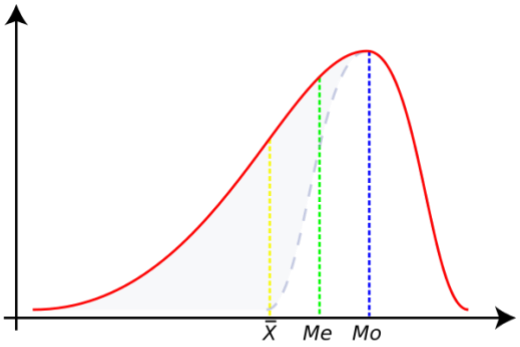
Encara que la mitjana i la desviació típica és molt diferent en ambdues mostres, el coeficient de variació és similar. Per tant, les observacions es distribueixen de manera similar. En el cas de la mostra 2 veiem que té un poc més de dispersió.

MESURES DE FORMA

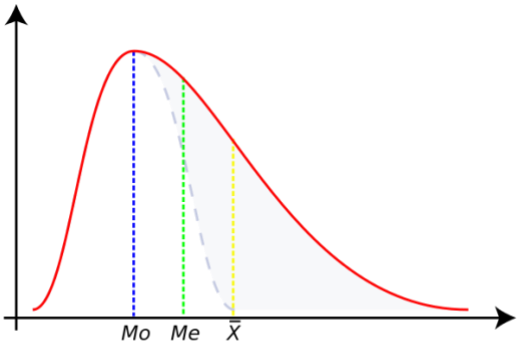
Com a mesures de forma veurem el coeficient d'asimetria i la curtosi.

- **Coeficient d'asimetria**: mesura el grau d'allunyament de la distribució de les dades respecte d'una distribució simètrica.
  - En una distribució simètrica el coeficient d'asimetria val 0.
  - El coeficient d'asimetria és positiu si la cua més llarga és la de la dreta, i negatiu en cas contrari.

Les diferències entre la mitjana i la mediana indiquen asimetria.

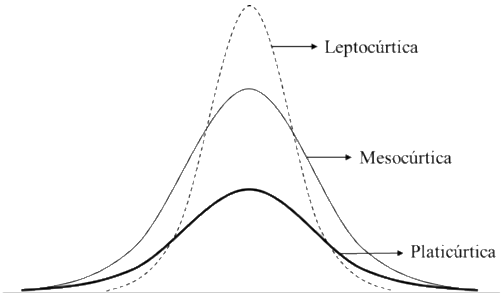


Asimetria negativa



Asimetria positiva

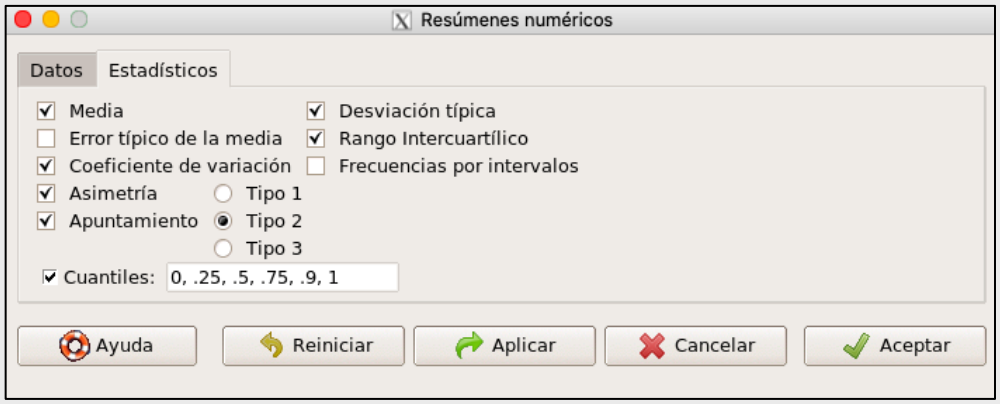
- **Curtosi:** és el grau d'apuntament de la distribució de dades.
  - Mesocúrtica: curtosi aproximadament 0.
  - Leptocúrtica: curtosi positiva.
  - Platicúrtica: curtosi negativa.



Per obtenir descripció numèrica de les dades amb *R-Commander*, fem el següent.

R-Commander: (amb les dades ja importades)

Estadístics → resums → resums numèrics → estadístics *(seleccionem quines mesures volem calcular)*



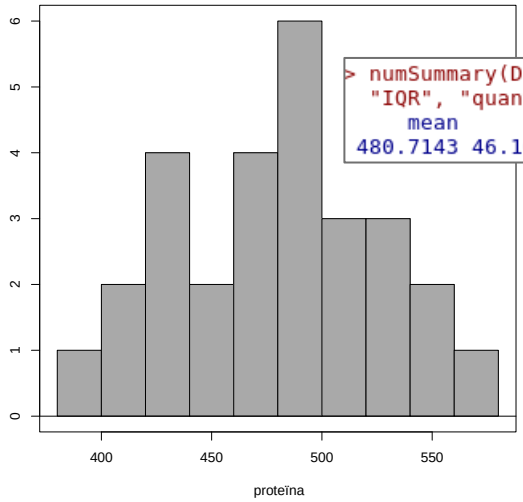
Exemple (proteïna en la llet de vaca)

Mostrem a continuació una eixida de *R-Commander*:

```
> numSummary(Dades[,"Proteïna", drop=FALSE], statistics=c("mean", "sd",
  "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
    mean       sd IQR  0%    25% 50%    75% 100%  n
480.7143 46.10845  66 397 443.25 488 509.25  565 28
```

La mitjana (mean) és 480,7143; la desviació típica (sd) és 46,10845; el rang interquartílic és 66 (Q3 – Q1 = 509,25 – 443,25 = 66), i el coeficient de variació (cv) és 0,09591654. També hem calculat els tres quartils (Q1: 25%, Q2: 50% i Q3: 75%) i el percentil 90%. A més, en la taula veiem el valor mínim (0%), 397, el màxim (100%), 565, i la grandària de la mostra (n), 28.

També podem dibuixar l'histograma de les dades, ja que és una variable numèrica continua, i calcular la simetria i la curtosi:



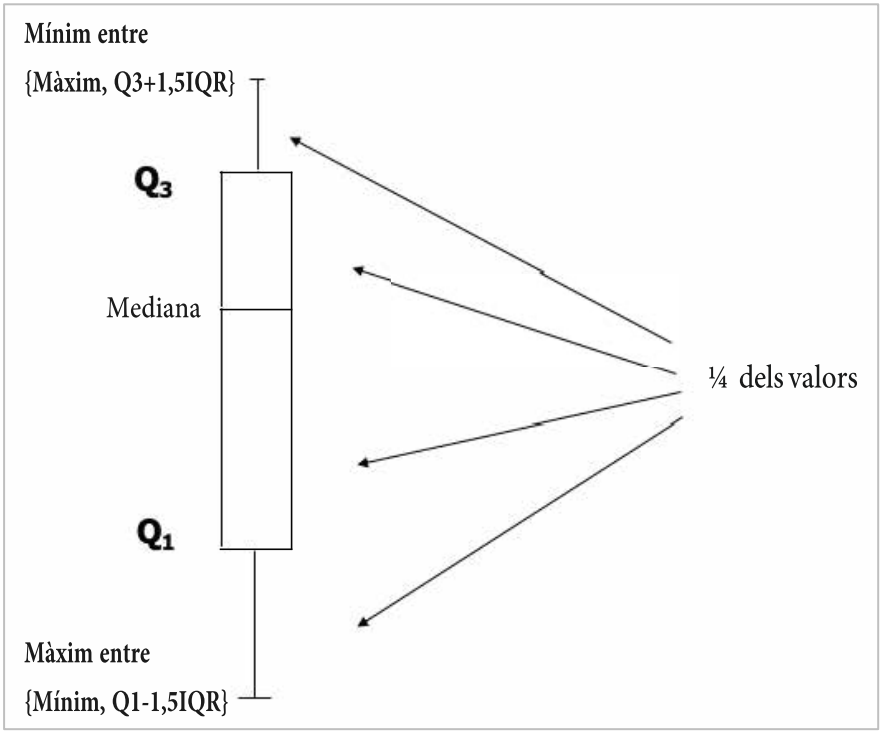

```
> numSummary(Dades[,"Proteïna", drop=FALSE], statistics=c("mean", "sd",
  "IQR", "quantiles", "skewness", "kurtosis"), quantiles=c(0,1), type="2")
    mean       sd IQR   skewness   kurtosis  0% 100%  n
480.7143 46.10845  66 -0.1185642 -0.8871086 397  565 28
```

Skewness → asimetria negativa (cua més llarga a l'esquerra)

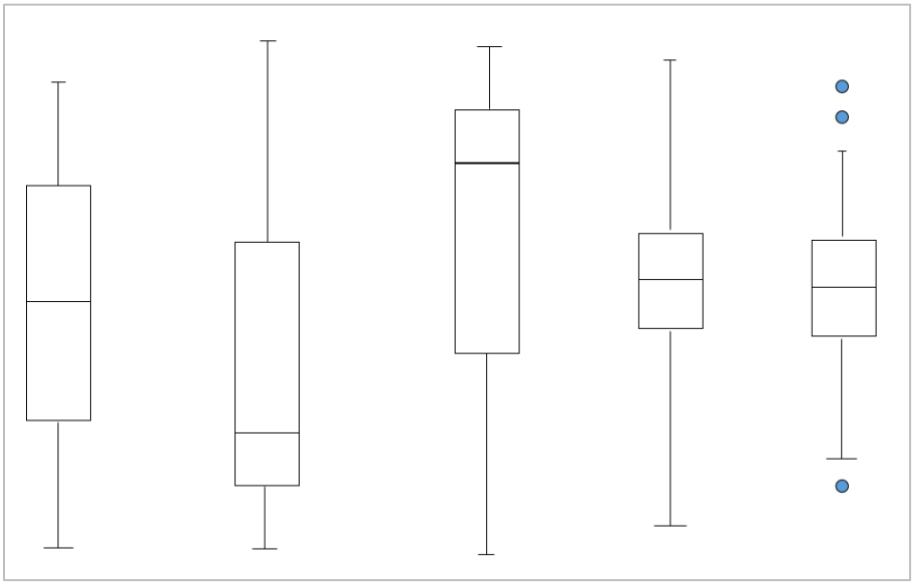
Curtosi → negativa (platicúrtica)

DIAGRAMA DE CAIXA

Ja contem amb tots els elements necessaris per a dibuixar el diagrama de caixa o *boxplot*. El diagrama de caixa és una representació visual del mínim, del màxim i dels quartils (Q1, Q2 = mediana, Q3)



Per calcular els extrems dels bigotis no es consideren els valors majors de Q3 + 1,5 IQR i els menors de Q1 – 1,5 IQR. Les dades que queden fora (valors extrems) es representen mitjançant punts (o diferents signes).



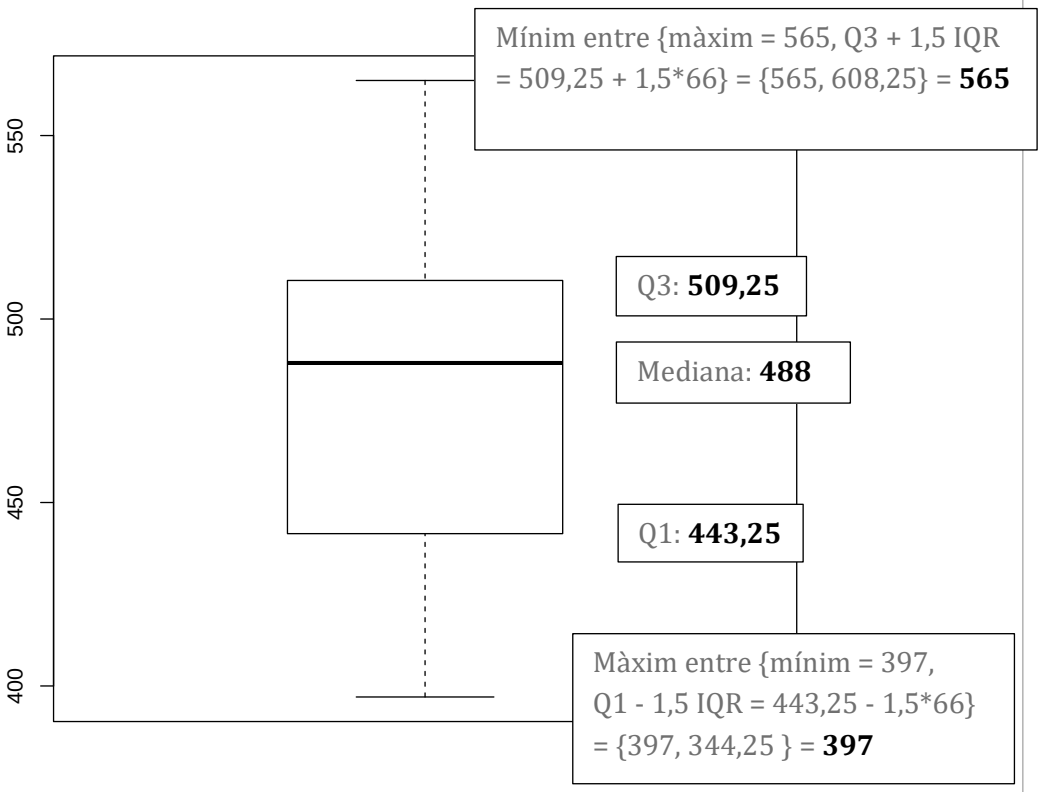
R-Commander: (amb les dades ja importades)

Gràfiques → Diagrama de caixa

Exemple (proteïnes en la llet de vaca)

```
> numSummary(Dades[,"Proteïna", drop=FALSE], statistics=c("mean", "sd",
  "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
     mean       sd IQR  0%    25% 50%    75% 100%  n
480.7143 46.10845  66 397 443.25 488 509.25  565 28
```

El valor mínim és 397; el valor màxim és 565; la mediana és 488; el primer quartil és 443,25; el tercer quartil és 509,25, i el rang interquartílic és IQR=66.



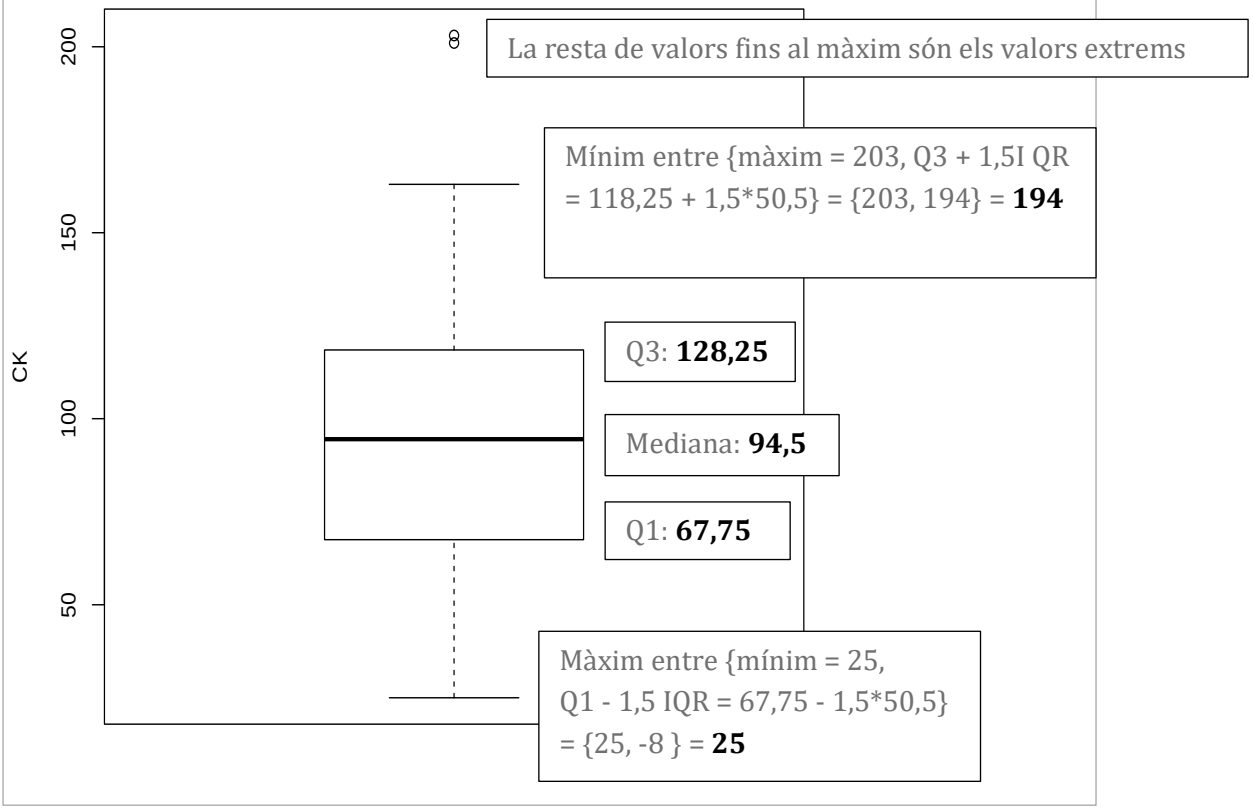
Exemple (fosfoquinasa creatina)

Les dades següents corresponen als nivells (en U/l) de fosfoquinasa creatina (CK) en la sang de 36 homes sans:

25	42	48	57	58	60	62	64	67	68
70	78	82	83	84	92	93	94	95	95
100	101	104	110	113	118	119	121	123	139
145	151	163	201	203					

```
> numSummary(Datos[,"CK", drop=FALSE], statistics=c("mean", "sd", "IQR",
+   "quantiles"), quantiles=c(0,.25,.5,.75,1))
     mean       sd  IQR 0%   25%  50%    75% 100%  n
 98.27778 40.38077 50.5 25 67.75 94.5 118.25  203 36
```

El valor mínim és 25; el valor màxim és 203; la mediana és 94,5; el primer quartil és 67,75; el tercer quartil és 118,25, i el rang interquartílic és IQR=50,5.



5. DESCRIPCIÓ NUMÈRICA D'UNA MOSTRA

Comencem aquest últim apartat del tema amb unes definicions bàsiques de successos i probabilitat.

Un **experiment aleatori** és un experiment en el qual no es pot predir el resultat exacte de cada realització o prova.

Una **variable aleatòria** és una característica d'interès observada durant l'experiment aleatori.

L'**espai mostral (Ω)** és el conjunt de possibles resultats que pot prendre la variable aleatòria.

Exemple

Experiment	Variable aleatòria	Espai mostral
Seleccionar a l'atzar un individu d'una població i mesurar-ne l'altura.	Altura	(0,∞)
Llançar un dau i veure el resultat.	Nombre	{1, 2, 3, 4, 5, 6}
Seleccionar a l'atzar un individu d'una població i determinar-ne el grup sanguini.	Grup sanguini	{A, B, AB, 0}

Un succés, E, és un subconjunt de l'espai mostral Ω. Els element de E comparteixen una mateixa característica.

La probabilitat del succés E és un valor numèric entre 0 i 1. Mesura el grau de facilitat que ocórrega el succés E. Quan l'espai mostral és finit i tots els seus elements tenen la mateixa probabilitat, la probabilitat de qualsevol succés E es pot calcular mitjançant la fórmula de Laplace:

$$P(E) = \frac{\text{nombre de casos favorables}}{\text{nombre de casos possibles}}.$$

Exemple

1. Considerem l'experiment de llançar una moneda a l'aire i siga E el succés "eixir cara". Si la moneda està equilibrada, P(C) = 0,5.
2. Ara llancem la moneda dues vegades. Com l'experiment ha canviat també canvia l'espai mostral (ara $\Omega = \{CC, XX, CX, XC\}$). Algunes probabilitats són P(CC) = 0,25 i P(XX) = 0,25.
3. Llancem un dau i observem el nombre que ix, per tant la variable és X = "nombre que ix en llançar el dau" i l'espai mostral és $\Omega = \{1, 2, 3, 4, 5, 6\}$. Algunes probabilitats són $P(X = 4) = 1/6$ i $P(2 \leq X \leq 4) = 3/6 = 0,5$.

A continuació estudiarem tres distribucions probabilístiques fonamentals en estadística: de Bernoulli, binomial i normal.

DISTRIBUCIÓ DE BERNOULLI

Les variables que poden prendre només dos valors es diuen dicotòmiques (exemple: sexe {mascle, femella}). Aquest valors poden representar-se mitjançant els valors {0,1}, de manera que:

- X = 1 si es compleix la propietat que interessa (èxit).
- X = 0 si no es compleix (fracàs).

La distribució de Bernoulli serveix per a descriure **probabilísticament les variables categòriques dicotòmiques**, X~ Ber(π), on π és la probabilitat d'èxit (quan la variable pren el valor 1). La funció de probabilitat és:

$$P(X = 1) = \pi \quad i \quad P(X = 0) = 1 - \pi.$$

La mitjana de X és π i la variància és $\pi(1 - \pi)$.

DISTRIBUCIÓ BINOMIAL

És una generalització de la distribució de Bernoulli. Suposem que es realitzen ***n* repeticions independents** d'un experiment amb dos resultats possibles (èxit i fracàs). La variable, X, que representa el nombre d'èxits en les *n* proves independents, té una distribució binomial, X~ Bi($n; \pi$), on π és la probabilitat d'èxit en cada prova. En aquest cas X pot prendre els valors {0, 1, 2, ... , n}. La funció de probabilitat per a qualsevol valor *k* de la variable, és:

$$P(X = k) = \frac{n!}{k!\,(n-k)!}\pi^k(1-\pi)^{n-k}, \qquad k = 0, 1, 2, \dots, n.$$

La mitjana de X és $n\pi$ i la variància és $n\pi(1 - \pi)$.

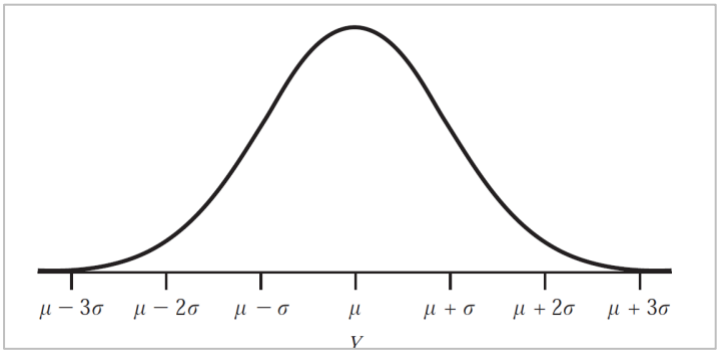
Exemple

Un cert medicament fa malbé en el ronyó en el 10% del pacients. Suposem un experiment amb cinc malalts que prenen el medicament. Aleshores, la variable, definida com X = "nombre de pacients amb danys pel medicament", es modelitza com una binomial, X~ Bi(5; 0,1).

DISTRIBUCIÓ NORMAL

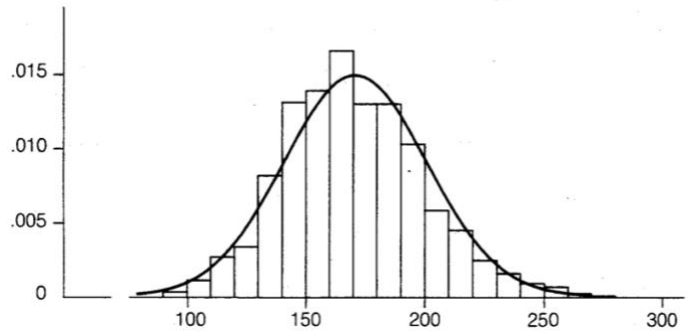
Una variable aleatòria continua X té una distribució normal, amb mitjana μ i desviació típica σ, $X \sim N(\mu; \sigma)$, si la funció de densitat és:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty.$$



Exemple

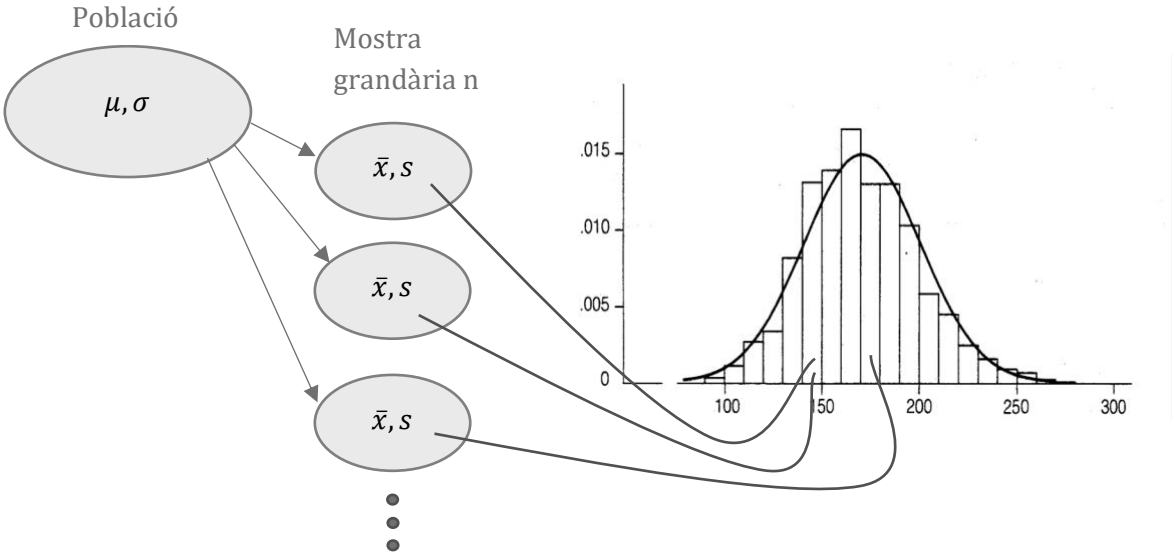
Com a part d'un estudi governamental de salut es va mesurar el nivell de colesterol sèric d'una gran mostra. La distribució per a xiquets de 14 anys pot ser aproximada per una corba normal amb mitjana $\mu = 170\ mg/dl$ i desviació típica $\sigma = 30\ mg/dl$.



TEOREMA CENTRAL DEL LÍMIT

La mitjana de la mostra $\bar{x}$ pot usar-se tant per a descriure les dades de la mostra com per estimar la mitjana poblacional μ. Però, com de prop està $\bar{x}$ de μ?

No podem respondre a aquesta pregunta per a $\bar{x}$, però podem respondre si pensem en termes del model de mostreig aleatori i veiem la mitjana mostral com una variable aleatòria $\bar{X}$. Com de prop està $\bar{X}$ de μ?



Elaborem la distribució en el mostreig de $\bar{X}$. En general, la mitjana de les mitjanes mostrals tendeix a la mitjana poblacional. La desviació típica de la distribució en el mostreig de $\bar{X}$ tendeix a la desviació típica poblacional dividida per l'arrel quadrada de la grandària de la mostra. A més, si la distribució poblacional X és normal, aleshores la distribució en el mostreig de $\bar{X}$ és també normal (independentment de la grandària de la mostra).

Teorema central del límit. Si *n* (la grandària mostral) és gran, aleshores la distribució en el mostreig $\bar{X}$ és aproximadament normal, encara que la distribució poblacional de X no siga normal

Per tant, podem afirmar que si la grandària de la mostra és gran (≥ 30), la mitjana mostral es comportarà com si procedira d'una distribució normal, amb la mateixa mitjana que la població però amb desviació típica dividida per $\sqrt{n}$:

$$\bar{X} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right).$$

6. RESUM

- La idea bàsica és transmetre al lector de manera clara la informació obtinguda de les dades. Cal decidir primer quin missatge es vol transmetre i assegurar-se que descriu de manera fidel les dades.
- Hem de fugir sempre de la sobreabundància de taules i gràfics.
- L'anàlisi exploratòria de les dades és el primer pas de tota anàlisi estadística, ja que ens permet resumir la informació obtinguda.
- Hem de tenir en compte si la variable o variables d'interès són categòriques o numèriques. El tipus de variable afecta la manera de descriure les dades observades a partir de la mostra.
- Podem donar informació mitjançant gràfiques i diagrames, però també podem utilitzar estadístics de tendència central, localització, dispersió i forma.
- Tenim eines informàtiques molt útils que ens proporcionen la informació necessària (gràfica i numèrica) de manera fàcil i segura. En particular, utilitzarem el programari R amb la interfície gràfica *R-Commander*.

ALGUNS EXERCICIS

1. Per a cadascuna de les situacions següents, descriu l'objectiu de l'estudi, la població i la mostra:
   i. Un analista ha de saber quants litres de sang utilitza un hospital de mitjana a la setmana. Disposa d'una llista amb tots els hospitals del país i ha decidit contactar amb aquells hospitals el número d'ordre dels quals en la llista correspon a una centena.
   ii. Un investigador vol saber si un nou producte curaria certa malaltia en humans. Com que ja se sap que el producte no és perjudicial, el prova amb una mostra aleatòria de 60 malalts d'una clínica de la seua ciutat.
   iii. Se suposa la mateixa situació que en l'apartat ii, però encara no se sap que el producte no siga perjudicial. Per tant, l'estudi es realitza primer amb 30 primats.
   iv. Un analista vol saber si un cert producte actua com s'esperava. Aleshores, pren una mostra aleatòria de 100 dels 1.256 individus que han provat aquest producte durant l'últim any.

2. En cadascun dels estudis següents, identifica la font o fonts de biaix en el mostreig, determina com podria afectar les conclusions de l'estudi i com es podria modificar el mètode de mostreig per evitar el biaix.
   i. Es van reclutar 800 voluntaris en clubs nocturns per participar en un experiment per a avaluar un nou tractament de l'ansietat social.
   ii. En un estudi sobre contaminació de l'aigua es van recollir espècimens d'aigua d'un rierol en quinze dies plujosos.

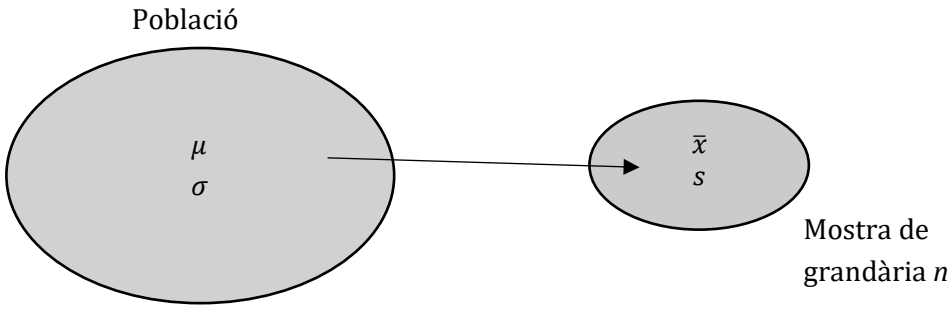
TEMA 2: INFERÈNCIA EN UNA POBLACIÓ

1. PARÀMETRES DE POBLACIÓ

Siga X la variable d'interès.

Una **població** és el conjunt de valors que resulta d'observar la variable X en **tots** els individus.

Un **paràmetre** és una mesura utilitzada per a descriure una certa característica de la població.



Tenim una mostra amb paràmetres desconeguts μ i σ, on μ és la mitjana poblacional i σ és la desviació típica de la població. El nostre objectiu és estimar el valor dels paràmetres μ i σ.

Un **estimador** és un estadístic mostral (és a dir, una funció de la mostra) utilitzat per a estimar un paràmetre de la població. El valor d'un estadístic proporciona una **estimació puntual** del paràmetre estudiat.

Per tal d'estimar el valor de μ i σ, seleccionem una mostra de grandària *n* i calculem la mitjana dels valors obtinguts i la seua desviació típica. D'aquesta manera, la mitjana mostral $\bar{x}$ és un estimador puntual de la mitjana poblacional μ i la desviació típica mostral s és un estimador puntual de la desviació típica poblacional σ.

2. ESTIMACIÓ DE LA MITJANA POBLACIONAL

Com hem vist, la mitjana mostral és el millor estimador de la mitjana poblacional. Per a una mostra particular, la mitjana mostral $\bar{x}$ és una estimació puntual de la mitjana poblacional μ. Però $\bar{x}$ no serà exactament igual a μ, hi ha una certa diferència entre les dues quantitats, coneguda com **error en el mostreig**.

Quan calculem una estimació puntual, assumim que hi ha algú error d'estimació. Per exemple, si triem una mostra diferent, de la mateixa de grandària que la primera mostra considerada, per a calcular l'estimació puntual, cada mostra proporcionarà possiblement una estimació puntual diferent de la mitjana poblacional μ. A més, podem calcular tantes

estimacions puntuals com mostres podem obtenir. De tots els valors que podem calcular, hi ha un valor millor que un altre? Els que estiguen més a prop de μ seran millors, però com que μ és desconegut, no podem saber-ho. Aleshores, com podem mesurar l'error que cometem quan calculem una estimació de la mitjana poblacional?

Per a qualsevol variable quantitativa o numèrica, X, on μ i σ són la mitjana i la desviació típica poblacional, respectivament, podem considerar que la població està constituïda per totes les possibles mostres de grandària n. Per tant, la mitjana poblacional per a les mitjanes mostrals és la mateixa que la mitjana poblacional per a la variable X, és a dir, $\mu_{\bar{X}} = \mu$, i la desviació típica per a les mitjanes mostral depèn de la grandària de la mostra i la desviació típica de X, $\sigma_{\bar{X}} = \sigma/\sqrt{n}$.

Com hem vist en el tema 1, la distribució de la mitjana mostral $\bar{X}$ estarà normalment distribuïda quan:

- X també siga una distribució normal. Aleshores, en aquest cas tenim que:

$$X \sim N(\mu, \sigma) \quad \Longrightarrow \quad \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Si X no té una distribució normal, s'hi pot aplicar el teorema central del límit (és a dir, n suficientment gran, $n \geq 30$).

Amb tot això, podem dir que la possible diferència entre la mitjana mostral i la poblacional està mesurada per la desviació típica de la mitjana mostral, $\frac{\sigma}{\sqrt{n}}$. Però, com que σ és també un paràmetre desconegut, l'estimem mitjançant el corresponent estadístic en la mostra, la desviació típica mostral *s*.

D'aquesta manera, definim l'error estàndard de la mitjana com:

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}.$$

Com hem vist, la mitjana poblacional μ, que volem estudiar, no serà exactament igual a la mitjana mostral $\bar{x}$, ja que hi ha un error associat. És convenient donar informació de com de segurs o confiats estem en la precisió d'aquesta estimació. Per a mesurar la precisió de l'estimació s'utilitza l'estimació per intervals, també anomenats **intervals de confiança.** Un interval de confiança per a μ és un interval [L1,L2] que inclou μ amb una probabilitat establerta.

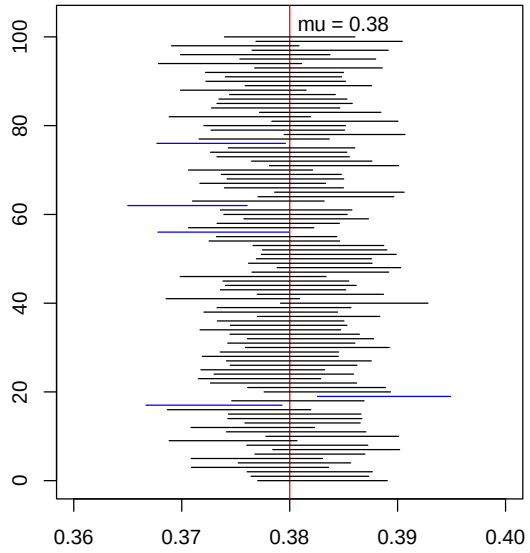
Per exemple, l'interval de confiança al 95% per a μ ha de complir que si es fa el mostreig diverses vegades, calculant cada vegada un interval de confiança, aproximadament el 95% dels intervals resultants contindrà μ. En la pràctica, en un experiment concret, només obtindrem una mostra i "confiem" en el fet que siga una de les mostres que estarien dins d'aquest 90%, però no sabrem amb certesa si és així o no.

Exemple (grossor de la closca dels ous)

En la producció comercial d'ous, el trencament és un dels problemes més importants. Per tant, la grossor de la closca és una variable d'interès. En un estudi s'observaren les grossors de closca dels ous produïts per una gran quantitat de gallines *White Leghorn,* i es va apreciar que les dades de la variable grossor eren compatibles amb les d'una distribució normal amb mitjana $\mu = 0{,}38\ mm$ i desviació típica $\sigma = 0{,}03\ mm$.

Construïm 100 mostres amb *R-Commander* i, per a cada mostra, calculem i dibuixem l'interval de confiança al 95% de confiança:

Observem que dels 100 intervals de confiança calculats, 5 no contenen la mitjana poblacional ($\mu = 0{,}38$) que considerem coneguda.



Com construïm l'interval de confiança?

Per a calcular els límits de l'interval de confiança utilitzem:

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

Com que σ és desconeguda, utilitzem *s* en la fórmula anterior. D'aquesta manera la distribució ja no és una normal i esdevé una **t de Student** amb *n-1* graus de llibertat:

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim T_{n-1}.$$

La distribució t de Student és una distribució contínua amb una forma similar a la normal estàndard, caracteritzada per un paràmetre que s'anomena graus de llibertat (gl). La dispersió en la t de Student és major que en la normal, però segons que augmenta els graus de llibertat, la distribució t de Student s'acosta a la normal estàndard.

Finalment, l'interval de confiança al $100(1-\alpha)\%$ per a μ serà:

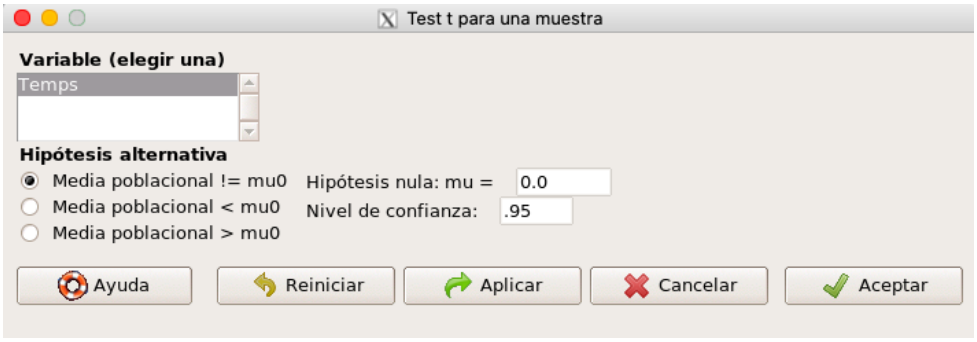
$$IC_{100(1-\alpha)\%}(\mu) = \left[\bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2}\frac{s}{\sqrt{n}}\right],$$

on α és el nivell de significació i $t_{1-\alpha/2}$ és el percentil d'ordre $1-\alpha/2$ d'una distribució t de Student amb *n-1* graus de llibertat. Per exemple, si $\alpha = 0{,}05$, aleshores $100(1-\alpha)\% = 95\%$ i $1-\alpha/2 = 0{,}975$, per tant hem de calcular el percentil $t_{0{,}975}$. Per sort, aquest càlcul el fa *R-Commander*.

R-Commander:

Estadístics → Mitjanes → Test t per a una mostra

Amb la selecció anterior ens apareix la finestra següent, on hem de seleccionar la variable de la qual volem calcular l'interval i indiquem el nivell de confiança (si estem calculant un interval de confiança al 95%, posem un nivell de confiança de ,95).



Exemple (picadures d'abella)

Temps (en minuts) d'aparició de la reacció en 40 malalts que van experimentar una reacció sistemàtica a la picadura d'una abella:

10,5	11,2	9,9	15,0	11,4	12,7	16,5	10,1	12,7	11,4	11,6
6,2	7,9	8,3	10,9	8,1	3,8	10,5	11,7	8,4	12,5	11,2
9,1	10,4	9,1	13,4	12,3	5,9	11,4	8,8	7,4	8,6	13,6
14,7	11,5	11,5	10,9	9,8	12,9	9,9				

Què podem dir sobre la mitjana de la població, μ? I sobre la desviació típica, σ?

Podem calcular les estimacions puntuals a partir del resum numèric.

Eixida *R-Commander*:

```
> numSummary(Picadures[,"Temps", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"),
  quantiles=c(0,.25,.5,.75,1))
    mean       sd   IQR  0%   25%  50%   75% 100%  n
10.5925 2.533609 2.825 3.8 9.025 10.9 11.85 16.5 40
```

Hem calculat, per tant, un estimador puntual per al temps mitjà d'aparició de reacció en tota la població,10,5925 minuts, amb una desviació típica de 2,533609 minuts.

Com és l'interval de confiança al 95% de la mitjana del temps d'aparició de la reacció sistemàtica a la picadura d'una abella?

Eixida *R-Commander:*

```
> with(Picadures, (t.test(Temps, alternative='two.sided', mu=0.0, conf.level=.95)))

        One Sample t-test

data:  Temps
t = 26.442, df = 39, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  9.782213 11.402787
sample estimates:
mean of x
  10.5925
```

L'interval de confiança al 95% és

$$IC_{95\%} = [9{,}782213;\ 11{,}402787].$$

Això vol dir que tenim una confiança del 95% que la mitjana poblacional estiga en aquest interval de confiança. "Confiem" que aquest siga l'interval que continga la mitjana poblacional de tots els que podríem construir.

3. CONTRASTOS D'HIPÒTESIS SOBRE LA MITJANA POBLACIONAL

Un contrast d'hipòtesis es basa en:

- Un nivell de significació α.
- Dues hipòtesis sobre les quals cal prendre una decisió:
  - Hipòtesi nul·la, H_0. És una hipòtesi conservadora i es formula amb l'únic objectiu de rebutjar-la. Connotació: igualtat.
  - Hipòtesi alternativa, H_A. És la hipòtesi complementaria de H_0. Per tant, és la hipòtesi "arriscada". És la hipòtesi d'interès.
- Una mostra de dades de la variable que es vol estudiar.

Hi ha dos tipus d'hipòtesis possibles:

Contrast bilateral	Contrast unilateral
$\begin{cases} H_0: \theta = \theta_0 \\ H_A: \theta \neq \theta_0 \end{cases}$	$\begin{cases} H_0: \theta \leq \theta_0 \\ H_A: \theta > \theta_0 \end{cases}$ $\quad$ $\begin{cases} H_0: \theta \geq \theta_0 \\ H_A: \theta < \theta_0 \end{cases}$

Ara, com podem prendre la decisió?

Un **test estadístic** és un procediment per a decidir sobre el rebuig o no de la hipòtesi nul·la a partir de les dades d'una mostra. Mitjançant el test, hem de decidir sobre la compatibilitat de les dades amb la hipòtesi nul·la. És a dir, hem de decidir si la diferència entre mostra i població pot ser deguda a l'atzar (error en el mostreig) o si la diferència és significativa i les dades proporcionen evidència a favor de la hipòtesi alternativa.

En particular, ens interessa el contrast sobre la mitjana poblacional:

Contrast bilateral	Contrast unilateral
$\begin{cases} H_0: \mu = \mu_0 \\ H_A: \mu \neq \mu_0 \end{cases}$	$\begin{cases} H_0: \mu \leq \mu_0 \\ H_A: \mu > \mu_0 \end{cases}$ $\quad$ $\begin{cases} H_0: \mu \geq \mu_0 \\ H_A: \mu < \mu_0 \end{cases}$

Per a resoldre el contrast que proposem, $H_0: \mu = \mu_0$, utilitzarem l'estadístic del test següent:

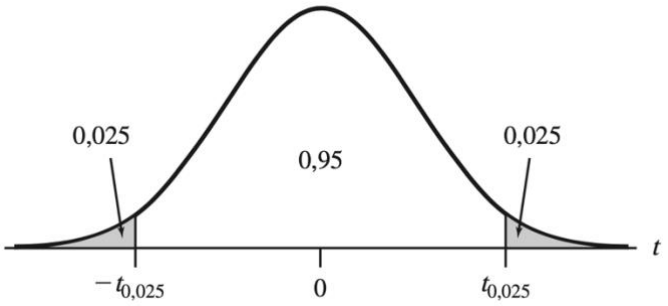
$$t_s = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Si H_0 és certa (és a dir, $\mu = \mu_0$), t_s té una distribució t de Student amb $n-1$ graus de llibertat. Però, com mesurem quantitativament si H_0 és certa o no? Utilitzarem el **p-valor** que és l'àrea, sota la corba de la t de Student, de les dues cues que estan més enllà de $-t_s$ i de $+t_s$. El p-valor mesura la compatibilitat de les dades amb H_0:

- Un p-valor gran (prop a 1) indica un valor de t_s proper a 0 (dades compatibles amb H_0).
- Un p-valor petit (prop a 0) indica que t_s està en una cua de les cues de la distribució (dades incompatible amb H_0).

Ara bé, què significa petit? Per contestar-ho, hem de definir que és el nivell de significació.

Per exemple, el valor $t_{0,025}$ es denomina **valor crític basat en el nivell de significació al 5% de dues cues** de la distribució t de Student i es defineix com el valor tal que l'interval entre $-t_{0,025}$ i $t_{0,025}$ conté el 95% de l'àrea sota la corba. L'àrea total ombrejada en la figura següent és igual 0,05. Noteu que aquesta part ombrejada està formada per dos "trossos", cadascun d'àrea 0,025.



NIVELL DE SIGNIFICACIÓ DEL TEST

A l'hora de rebutjar la hipòtesi nul·la hem de decidir quan considerem suficientment petit el p-valor. Per fer-ho, hem de fixar un valor límit, en l'escala del p-valor (recordem que el p-valor és un valor entre 0 i 1), de manera que per sota d'aquest valor direm que les dades són incompatibles amb H_0 i que per damunt seu considerarem que les dades són compatibles amb H_0. Aquest valor límit s'anomena **nivell de significació** del test i és designat pel símbol α. El valor de α més utilitzat és 0,05, encara que en alguns experiments també és habitual considerar 0,1, 0,01 o altres.

- Si el p-valor és menor que α, aleshores considerarem que les dades són incompatibles amb H_0; en aquest cas, rebutjarem H_0 i direm que les dades presenten evidència a favor de H_A.
- Si el p-valor és major o igual que α, no rebutjarem H_0 i la conclusió serà que no hi ha prou evidencia per afirmar que H_A és certa.

Però, en rebutjar o no la hipòtesi nul·la, ens podem trobar amb els errors següents:

	Decisió	
	No rebutgem H_0	Rebutgem H_0
H_0 és verdadera	Sense error: $(1-\alpha)$	Error de tipus I: (α)
H_0 és falsa	Error de tipus II: (β)	Sense error: $(1-\beta$= potència)

Un estudi ideal és aquell en què α i β són petits i la potència és gran. Però, sabem que el $\alpha\%$ de les vegades ens equivocarem i rebutjarem sense motius. L'únic error que controlarem en aquest curs és el de tipus I, ja que fixarem, a l'inici de l'experiment, el valor α tan petit com desitgem.

ESQUEMA DEL CONTRAST D'HIPÒTESIS

1. Establir el nivell de significació α (normalment 0,05).
2. Especificar les hipòtesis del contrast (plantejar contrast d'hipòtesis).
3. Calcular el valor de l'estadístic apropiat a partir de les dades de la mostra.
4. Calcular el p-valor.
5. Decidir si rebutgem o no la hipòtesi nul·la (resoldre el contrast d'hipòtesis). Si p-valor < α, rebutgem H_0.
6. Presentar les conclusions.

Exemple (pesos de corder en nàixer)

Una genetista va pesar 28 corders en nàixer. Tots els corders van nàixer a l'abril, tots eren de la mateixa raça (*Rambouillet*) i tots van ser naixements d'un sol corder (no hi havia bessons). La dieta i altres condicions experimentals van ser les mateixes per a tots els progenitors. Les dades següents es corresponen al pes de cada corder:

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

Estem interessats a comprovar si el pes mitjà en nàixer, de qualsevol corder de la població descrita, és igual o diferent de 4 kilograms (mitjana de la mostra). Denotem per X: "pes en nàixer d'un corder". Plantegem el contrast d'hipòtesis:

$$\begin{cases} H_0: \mu = 4 & (\text{el pes mitjà en nàixer per als corders, en general, és 4 quilograms}) \\ H_A: \mu \neq 4 & (\text{el pes mitjà en nàixer per als corders, en general, no és 4 quilograms}) \end{cases}$$

- Si la variable $\bar{X}$ segueix una distribució normal, apliquem el test de la **t de Student** d'una mostra per a resoldre el contrast. Si la variable $\bar{X}$ no segueix una distribució normal, apliquem un test no paramètric (**test de Wilcoxon**).

En aquest exemple, per a continuar, considerem que X segueix una distribució normal, $X \sim N(\mu, \sigma)$ (aquesta hipòtesi la comprovarem més endavant mitjançant un test adequat, el **test de Shapiro-Wilk**) i, per tant, $\bar{X}$ segueix una distribució normal i es compleix la condició de normalitat. Calculem el p-valor amb el **test t per a una mostra:**

```
> with(Corders, (t.test(Pes, alternative='two.sided', mu=4, conf.level=.95)))

        One Sample t-test

data:  Pes
t = 9.0923, df = 25, p-value = 0.00000000211
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 4.925208 5.467099
sample estimates:
mean of x
 5.196154
```

Conclusió: Com que el p-valor 0,00000000211 (molt petit) és menor que α = 0,05, rebutgem H_0. Hi ha prou evidència a favor de la H_A (el pes mitjà en nàixer per als corders no és 4 quilograms).

El contrast que hem vist en l'exemple s'anomena bilateral o no direccional, perquè no indiquem una direcció per a la hipòtesi alternativa. En algunes ocasions, però, la desviació de la mitjana només es pot donar en un sentit o només ens interessa demostrar que aquesta desviació es dona en un únic sentit. En aquests casos, utilitzarem una hipòtesi alternativa direccional per a indicar que rebutjarem la hipòtesi nul·la només si la diferència entre mostra i població és significativa en la direcció que proposa la hipòtesi alternativa. En qualsevol cas, hem d'elegir la forma de la hipòtesi alternativa abans d'obtenir les dades; altrament, estarem falsejant el significat del nivell de significació del test.

Exemple (bacteris)

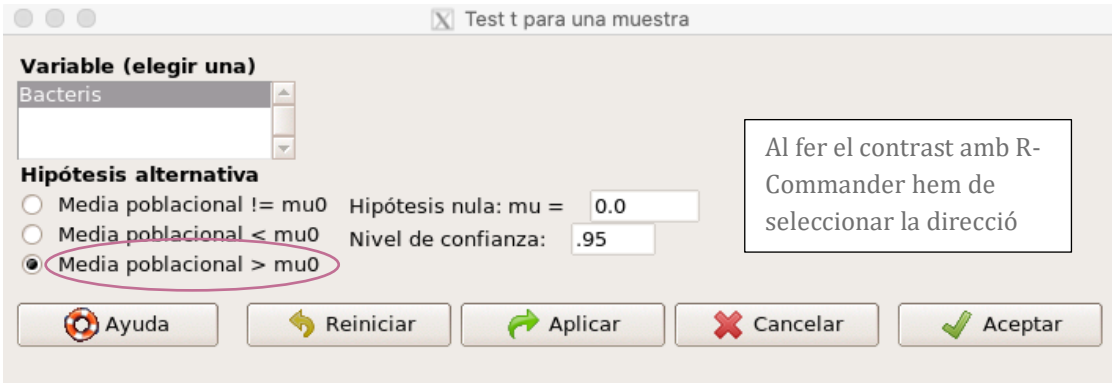
El Departament de Salut dels Estats Units ha fixat en 70 el nombre mitjà de bacteris per centímetre cúbic d'aigua per a les aigües en què es practica la recollida de cloïsses. Un nivell mitjà superior a 70 és perillós perquè menjar aquestes cloïsses podria causar hepatitis. S'ha pres una mostra de 9 observacions i s'han obtinguts els valors:

69 74 75 70 73 73 71 73 68

S'hauria de prohibir la recollida de cloïsses en aquestes aigües?

Només rebutjarem la hipòtesi nul·la si el resultat de l'estudi indica, amb un nivell de significació de α = 0,05, que el nombre mitjà de bacteris és superior a 70. És a dir, la hipòtesi alternativa és allò que volem demostrar:

$$\begin{cases} H_0: \mu \leq 70 \text{ (el nombre mitjà de bateris en l'aigua és menor o igual a 70)} \\ H_A: \mu > 70 \quad \text{(el nombre mitjà de bateris en l'aigua és superior a 70)} \end{cases}$$



Al fer el contrast amb R-Commander hem de seleccionar la direcció

```
> with(bacteris, (t.test(Bacteris, alternative='greater', mu=70, conf.level=.95)))

        One Sample t-test

data:  Bacteris
t = 2.132, df = 8, p-value = 0.03279
alternative hypothesis: true mean is greater than 70
95 percent confidence interval:
 70.21299      Inf
sample estimates:
mean of x
 71.66667
```

Apareix "*greater*" que significa major que

Conclusió: Com que el p-valor = 0,03279 és menor que α = 0,05, rebutgem H_0. Hi ha prou evidència a favor de la H_A (ha de prohibir-se la recollida de cloïsses).

4. CONDICIONS DE VALIDESA DEL TEST T DE STUDENT

S'han de complir les condicions següents, tant per al càlcul de l'interval de confiança per a la mitjana, com per a resoldre els contrast d'hipòtesi.

Condicions sobre el disseny de l'experiment.

- És raonable considerar les dades com una mostra aleatòria de la població d'interès. Per tant, les observacions de la mostra han de ser independents entre si.

Condicions sobre la distribució de la població. La distribució en el mostreig de mitjana mostral ($\bar{X}$) ha de ser normal. Tenim dues opcions:

- Si n és petita, la distribució de la població (X) ha de ser aproximadament normal (**test de normalitat de Shapiro-Wilk).**
- Si n és gran ($n \geq 30$), no cal que X siga aproximadament normal, perquè ho serà $\bar{X}$ (teorema central del límit).

TEST DE NORMALITAT DE SHAPIRO-WILK

R-Commander:

Estadístics → Resums → Test de normalitat de Shapiro-Wilk

Aquest test resol aquest contrast d'hipòtesi:

$$\begin{cases} H_0: \text{La distribució de la variable és normal} \\ H_A: \text{La distribució de la variable no és normal} \end{cases}$$

Per tant, si p-valor$< \alpha$, rebutgem la hipòtesi nul·la, rebutgem la normalitat. Si p-valor $\geq \alpha$, no rebutgem la normalitat (i considerarem que la variable segueix una distribució aproximadament normal).

Exemple (bacteris)

Fem el test de normalitat amb *R-Commander*:

```
> normalityTest(~Bacteris, test="shapiro.test", data=bacteris)

        Shapiro-Wilk normality test

data:  Bacteris
W = 0.96816, p-value = 0.8785
```

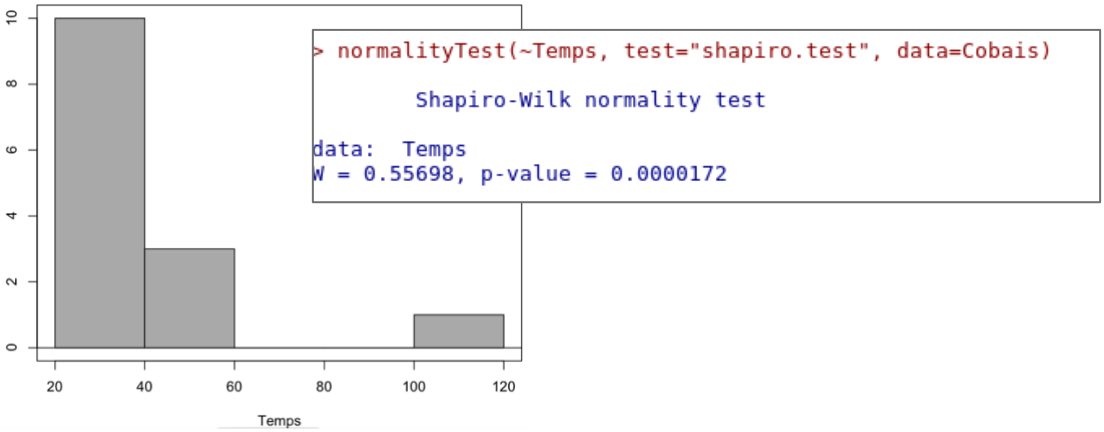
Com que el p-valor=0,8785 $\geq \alpha = 0{,}05$, no rebutgem H_0. Per tant, podem assumir la normalitat de la variable i aplicar el test T de Student.

Exemple (cobais)

En un experiment previ a provar-ho en humans, es va administrar a 14 cobais de laboratori un nou medicament que, com a efecte secundari, causa somnolència. El temps transcorregut, en minuts, entre la ingesta d'aquest producte i l'entrada en fase de son va ser:

44	27	24	24	36	36	44
44	120	29	36	36	36	36

Podem dir que les dades són normals?



Com que el p-valor=0,0000172 $< \alpha = 0{,}05$, rebutgem H_0. Per tant, no podem assumir la normalitat de la variable i **no podem** aplicar el test T de Student (ja que hi ha poques dades i no tenim normalitat).

5. PROVES NO PARAMÈTRIQUES: TEST DE WILCOXON

Si les dades no provenen d'una distribució normal (hem rebutjat la hipòtesi nul·la del test de Shapiro-Wilk) i a més no tenim una mostra prou gran, no es pot aplicar cap contrast basat en la t de Student, ni calcular intervals de confiança per a la mitjana poblacional. Aleshores, caldrà aplicar un test no paramètric basat en la mediana poblacional. Ho farem aplicant el test de Wilcoxon. El test de Wilcoxon és una prova no paramètrica basada en la mediana:

$$\begin{cases} H_0: \text{La mediana poblacional és igual a } \mathrm{m}_0 \ (Me = m_0) \\ H_A: \text{La mediana poblacional no és igual a } \mathrm{m}_0 \ (Me \neq m_0) \end{cases}$$

S'ha de destacar que el contrast pot ser unilateral.

R-Commander:

Estadístics → Tests no paramètrics → Test de Wilcoxon per a una mostra

Exemple (cobais)

Recordeu que en aquest exemple hem rebutjat la hipòtesi de normalitat. Per a la variable X= "temps transcorregut entre la ingesta d'un producte i l'entrada en fase de son", ens demanen contrastar si la mediana poblacional val 40 minuts.

$$\begin{cases} H_0: \text{La mediana poblacional de X és igual a 40 } (Me = 40) \\ H_A: \text{La mediana poblacional de X no és igual a 40 } (Me \neq 40) \end{cases}$$

Ara, fem el test de Wilcoxon amb *R-Commander*

```
> with(Cobais, wilcox.test(Temps, alternative='two.sided', mu=40))

        Wilcoxon signed rank test with continuity correction

data:  Temps
V = 29, p-value = 0.1365
alternative hypothesis: true location is not equal to 40
```

Com que el p-valor=0,1365 $> \alpha = 0{,}05$, no rebutgem H_0. Per tant, tenim evidència estadística per afirmar que la mediana del temps per entrar en fase de son és de 40 minuts.

Ara, com fem un interval de confiança per la mediana?

Executem en R el test de Wilcoxon per a una mostra, amb hipòtesi alternativa bilateral (alternative=two.sided) i, posteriorment, afegim en la finestra d'instruccions conf.int=TRUE i premem "Executa".

```
With(Cobais, wilcox.test(Temps, alternative=two.sided, mu=40, conf.int=TRUE))
```

Exemple (cobais)

```
> with(Cobais, wilcox.test(Temps, alternative='two.sided', mu=40, conf.int=TRUE))

        Wilcoxon signed rank test with continuity correction

data:  Temps
V = 29, p-value = 0.1365
alternative hypothesis: true location is not equal to 40
95 percent confidence interval:
 30.00003 40.00000
sample estimates:
(pseudo)median
      35.99995
```

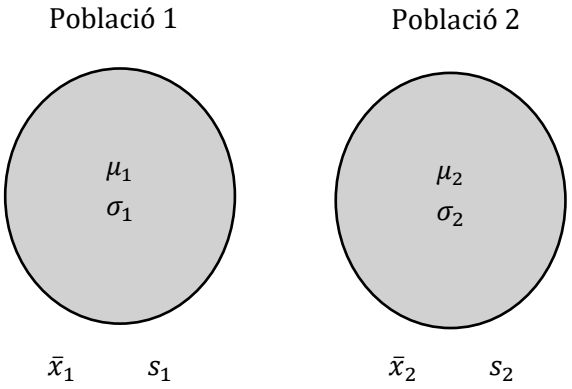
En aquest cas l'estimació puntual de la mediana poblacional és de 36,45 minuts, i l'interval de confiança al 95% per la mediana poblacional és [30,00003; 40] minuts.

TEMA 3: ANÀLISI ESTADÍSTICA DE DUES MOSTRES

En el tema 2 hem considerat l'anàlisi d'una mostra de dades numèriques per a fer inferència sobre la mitjana de la població de la qual s'havia obtingut la mostra. En moltes situacions pràctiques, la investigació involucra la comparació de dues o més mostres obtingudes de la mateixa població o de poblacions diferents. En aquest tema introduirem mètodes per a l'anàlisi i la comparació de dues mitjanes poblacionals. Podem trobar dos tipus de mostres.

Mostres relacionades/aparellades: cada individu de la segona mostra correspon a un individu de la primera. Per tant, la grandària de les dues mostres és la mateixa i tenim un conjunt de parells de dades.

Mostres independents: cada individu de la segona mostra no es relaciona amb cap individu en concret de la primera mostra. La grandària de les mostres pot ser diferent.

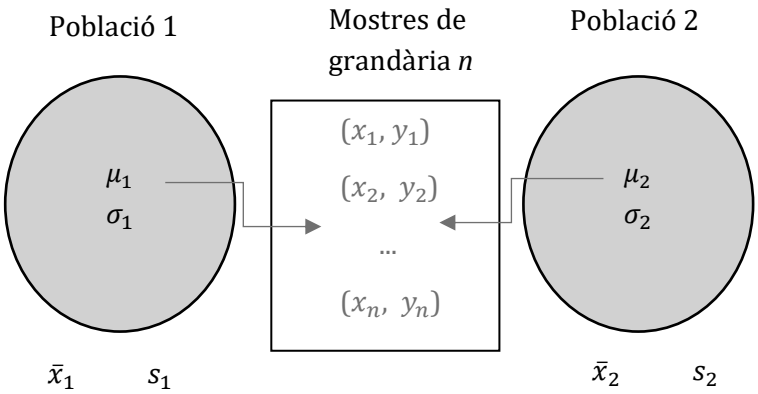


Amb dues mostres ens poden sorgir moltes preguntes:

- Entre quins valors, amb una certa confiança alta, es trobarà la diferència de les mitjanes poblacionals? Hem de calcular l'interval de confiança per a la diferència $\mu_1 - \mu_2$.
- Són les dades compatibles amb la hipòtesi segons la qual les mitjanes poblacionals de les mostes relacionades són iguals? Hem de resoldre un contrast d'hipòtesis per a $\mu_1 = \mu_2$.

En aquest tema veurem com construir intervals de confiança i com resoldre contrastos d'hipòtesis amb dues mostres, tant relacionades com independents.

1. MOSTRES RELACIONADES



En un disseny de mostres aparellades, les dades es presenten per parells. Per tant, les dues mostres tenen la mateixa grandària: n. Les unitats observacionals de cada parell de dades (x_1, x_2) estan relacionades entre si d'alguna manera, és a dir, tenen en comú alguna relació que no tenen amb els membres d'altres parells.

Com que les mesures es prenen sobre el mateix individu, podem definir la variable de la resta $X_d = X_1 - X_2$. Aquesta nova variable és la diferència entre les dues variables originals.

Ara, com calculem intervals de confiança o resolem un contrast d'hipòtesis?

- **Mètodes paramètrics:** estan basats en la distribució t de Student i ens permeten:
  - calcular intervals de confiança per a la diferència de les mitjanes poblacionals $\mu_1 - \mu_2$, i
  - resoldre contrastos d'hipòtesis per a la igualtat de les mitjanes poblacionals $\mu_1 = \mu_2$.
- **Mètodes NO paramètrics:** són una alternativa als mètodes paramètrics, si aquests no es poden utilitzar, i ens permeten:
  - calcular intervals de confiança per a la diferència de les medianes poblacionals, i
  - resoldre contrastos d'hipòtesis per a la igualtat de les distribucions (medianes), mitjançant el test de Wilcoxon.

La potència de les proves paramètriques és sempre major que la potència de les proves no paramètriques, per això els mètodes paramètrics són preferibles. Però, quan podem aplicar els mètodes paramètrics?

- Si la grandària de les mostres és menuda, hem d'exigir que la distribució de la variable diferència $X_d = X_1 - X_2$ siga normal.
- Si la grandària de la mostra és gran ($n \geq 30$), la distribució de la variable X_d no necessita ser normal, perquè assumim normalitat per a la seua mitjana mostral a partir del teorema central del límit.

Interval de confiança per a la diferència de dues mitjanes poblacionals

Per a fer inferència sobre la mitjana μ_d a partir de dues mostres relacionades utilitzarem la mostra de les diferències. És a dir, treballarem amb una única mostra.

L'interval de confiança al $100(1-\alpha)\%$ per a μ_d serà:

$$IC_{100(1-\alpha)\%}(\mu_d) = \left[\bar{x}_d - t_{1-\frac{\alpha}{2}}\frac{s_d}{\sqrt{n}}, \bar{x}_d + t_{1-\alpha/2}\frac{s_d}{\sqrt{n}}\right],$$

on $\bar{x}_d$ és la mitjana mostral de les diferències, s_d és la desviació típica mostral de les diferències, α és el nivell de significació i $t_{1-\alpha/2}$ és el percentil d'ordre $1-\alpha/2$ d'una distribució t de Student amb *n-1* graus de llibertat. Amb *R-Commander:*

R-Commander:

Estadístics → Mitjanes → Test t per a una mostra

Contrast d'hipòtesis

El contrast d'hipòtesis per a la comparació de dues mitjanes poblacionals aparellades es defineix de la manera següent:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_A: \mu_1 \neq \mu_2 \end{cases} \Leftrightarrow \begin{cases} H_0: \mu_d = 0 \\ H_A: \mu_d \neq 0 \end{cases}$$

L'estadístic de contrast utilitzat per a resoldre el contrast d'hipòtesis és $t_s = \frac{\bar{x}_d}{s_d/\sqrt{n}}$.

Criteri de decisió: rebutjar H_0 si el valor p < α.

Alternativa no paramètrica: test de Wilcoxon

Si no es compleixen les condicions de validesa dels mètodes basats en la t de Student, hem d'aplicar un test no paramètric per a mostres relacionades: **el test de Wilcoxon**, que ens permet comprovar si hi ha diferències entre les distribucions poblacionals de les dues mostres relacionades. És a dir, podem contrastar que:

$$\begin{cases} H_0: \text{La distribució de les variables } X_1 \text{ i } X_2 \text{ és la mateixa.} \\ H_A: \text{La distribució de les variables } X_1 \text{ i } X_2 \text{ NO és la mateixa.} \end{cases}$$

R-Commander:

Estadístics → Test no paramètrics → Test de Wilcoxon per a mostres aparellades

Podem calcular l'interval de confiança per a la mediana de la diferència, mitjançant el test de Wilcoxon. Per al seu càlcul és necessari utilitzar una H_A bilateral ("two.sided") i afegir l'opció **conf.int = TRUE.**

Exemple (diabetis)

La diabetis gestacional és una alteració en el metabolisme de la glucosa reconegut durant l'embaràs i que desapareix després del part. En un estudi es compararen els nivells de glucosa en sang (mg/dL) en 12 dones embarassades en la setmana 8 de gestació, abans i després de menjar-se una peça de fruita fresca. Volem saber si hi ha alguna diferència entre els nivells mitjans de glucosa abans i després de menjar fruita.

Abans	60	80	62	...
Després	97	66	116	...

Resum numèric *R-Commander*:

```
> numSummary(Diabetis[,c("abans", "després"), drop=FALSE],
  statistics=c("mean", "sd", "IQR",
  "quantiles"), quantiles=c(0,.25,.5,.75,1))
            mean       sd  IQR 0% 25%   50%   75% 100%  n
abans     74.5000 11.54044 12.5 60  66  74.5  78.5  103 12
després 119.0833 37.20083 60.5 66  92 112.5 152.5  170 12
```

Calculem la mostra de les diferències, que tindra grandària $n = 12$. Com que es tracta d'una mostra menuda, hi apliquem el test de normalitat per saber si podem utilitzar mètodes paramètrics:

$$\begin{cases} H_0: \text{La distribució de la diferència és normal} \\ H_A: \text{La distribució de la diferència no és normal} \end{cases}$$

Com que el valor p=0,9695 > 0,05, no rebutgem H_0 (normalitat). Per tant, utilitzem mètodes paramètrics.

```
> Diabetis$dif <- with(Diabetis, després- abans)

> normalityTest(~dif, test="shapiro.test", data=Diabetis)

        Shapiro-Wilk normality test

data:  dif
W = 0.97695, p-value = 0.9685
```

Com que assumim normalitat, podem calcular l'interval de confiança utilitzant la variable diferència.

L'interval ens indica que el nivell mitjà de glucosa en sang després de menjar una peça de fruita és major que el nivell mitjà de glucosa en sang abans. La diferència de nivells mitjans està entre 21,45 i 67,72 mg/dL, amb una confiança del 95%.

```
> with(Diabetis, (t.test(després, abans,
  alternative='two.sided', conf.level=.95,
  paired=TRUE)))

        Paired t-test

data:  després and abans
t = 4.2414, df = 11, p-value = 0.001386
alternative hypothesis: true difference
in means is not equal to 0

95 percent confidence interval:
 21.44766 67.71901
sample estimates:
mean of the differences
               44.58333
```

Exemple (diabetis). Continuació

També podem utilitzar el test t per a dues mostres relacionades que ens proporciona R-Commander per al càlcul de l'interval de confiança:

```
> with(Diabetis, (t.test(després, abans,
  alternative='two.sided', conf.level=.95,
  paired=TRUE)))

        Paired t-test

data:  després and abans
t = 4.2414, df = 11, p-value = 0.001386
alternative hypothesis: true difference
in means is not equal to 0

95 percent confidence interval:
 21.44766 67.71901
sample estimates:
mean of the differences
               44.58333
```

R-Commander:

Estadístics → Mitjanes → Test t per a dades relacionades

A continuació, volem saber si coincideixen o no els nivells mitjans de glucosa abans i després de menjar fruita.

$$\begin{cases} H_0: \text{Els nivells mitjans de glucosa abans i després de menjar són iguals} \\ H_A: \text{Els nivells mitjans de glucosa abans i després de menjar són distints} \end{cases}$$

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_A: \mu_1 \neq \mu_2 \end{cases} \qquad \begin{cases} H_0: \mu_d = 0 \\ H_A: \mu_d \neq 0 \end{cases}$$

```
> with(Diabetis, (t.test(després, abans,
  alternative='two.sided', conf.level=.95,
  paired=TRUE)))

        Paired t-test

data:  després and abans
t = 4.2414, df = 11, p-value = 0.001386
alternative hypothesis: true difference
in means is not equal to 0

95 percent confidence interval:
 21.44766 67.71901
sample estimates:
mean of the differences
               44.58333
```

```
> with(Diabetis, (t.test(dif,
  alternative='two.sided', mu=0.0,
  conf.level=.95)))

        One Sample t-test

data:  dif
t = 4.2414, df = 11, p-value = 0.001386
alternative hypothesis: true mean is
not equal to 0

95 percent confidence interval:
 21.44766 67.71901
sample estimates:
mean of x
 44.58333
```

Rebutgem H_0 amb un nivell de significació α = 0,05 (valor p = 0,001386 < 0,05). Tenim evidència estadística que els nivells mitjans de glucosa no coincideixen.

Exemple (cafeïna)

En un estudi sobre l'efecte de la cafeïna en el metabolisme muscular, 9 homes voluntaris feren diverses proves d'exercici de braços en dues ocasions separades. En la primera ocasió, els 9 voluntaris prengueren una càpsula de placebo i en la segona, una càpsula de cafeïna pura una hora abans de la prova. Durant cada prova es va mesurar la raó d'intercanvi respiratori (RIR) de l'individu. La RIR és la raó entre el diòxid de carboni produït i l'oxigen consumit, i indica si l'energia està obtenint-se a partir d'hidrats de carboni o a partir de greix. Volem saber si hi ha diferència en el RIR mitjà, en funció del fet que hagen pres primerament cafeïna o placebo.

Individu	1	2	3	4	5	6	7	8	9
Placebo	105	119	92	97	96	101	94	95	98
Cafeïna	96	99	89	95	88	95	88	93	88

Resum numèric *R-Commander*:

```
> numSummary(cafeina[,c("cafeina", "placebo"), drop=FALSE], statistics=c("mean", "sd", "IQR",
+   "quantiles"), quantiles=c(0,.25,.5,.75,1))
           mean       sd IQR 0% 25% 50% 75% 100% n
cafeina 92.33333 4.183300   7 88  88  93  95   99 9
placebo 99.66667 8.215838   6 92  95  97 101  119 9
```

Definim les varibles relaciones com a:

X_1: RIR en pacients amb placebo.

X_2: RIR en pacients amb cafeïna.

Calculem la mostra de les diferències, que tindrà una grandària $n = 9$. Com que es tracta d'una mostra menuda, comprovem la normalitat de la variable diferència.

```
> cafeina$diferencia <- with(cafeina, placebo - cafeina)

> normalityTest(~diferencia, test="shapiro.test", data=cafeina)

        Shapiro-Wilk normality test

data:  diferencia
W = 0.84907, p-value = 0.07279
```

Com que el valor p =0,07279 > α, no es rebutja H_0, la normalitat. És a dir, podem utilitzar mètodes paramètrics.

Volem resoldre el contrast següent:

$$\begin{cases} H_0: \mu_d = 0 \\ H_A: \mu_d \neq 0 \end{cases}$$

Rebutgem H_0 al nivell de significació $\alpha = 0{,}05$ (valor p = 0,004323 < 0,05). Tenim evidència estadística que els nivells mitjans de RIR són distints.

```
> with(cafeina, (t.test(diferencia, alternative='two.sided',
mu=0.0, conf.level=.95)))

        One Sample t-test

data:  diferencia
t = 3.9355, df = 8, p-value = 0.004323
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  3.036348 11.630318
sample estimates:
mean of x
 7.333333
```

Exemple (triglicèrids)

Els triglicèrids són components sanguinis que es pensa que intervenen en la malaltia coronària arterial. Per comprovar si l'exercici regular pot reduir els nivells de triglicèrids, els investigadors mesuraren la concentració de triglicèrids en sèrum sanguini de 7 homes voluntaris, abans i després d'un programa d'exercici de 10 setmanes. Els resultats es mostren en la taula següent.

Participant	1	2	3	4	5	6	7
Abans	0,87	1,13	3,14	2,14	2,98	1,18	1,6
Després	0,57	1,03	1,47	1,43	1,2	1,09	1,51

Resum numèric *R-Commander*:

```
> numSummary(Triglicerids[,c("abans", "després"), drop=FALSE],
 statistics=c("mean", "sd", "IQR",
 "quantiles"), quantiles=c(0,.25,.5,.75,1))
            mean        sd   IQR   0%   25% 50%  75% 100% n
abans   1.862857 0.9141611 1.405 0.87 1.155 1.6 2.56 3.14 7
després 1.185714 0.3312530 0.390 0.57 1.060 1.2 1.45 1.51 7
```

Calculem la mostra de les diferències, que tindrà una grandària $n = 7$. Com que es tracta d'una mostra menuda, hi apliquem el test de normalitat per saber si podem utilitzar mètodes paramètrics:

$$\begin{cases} H_0: \text{La distribució de la diferència és normal} \\ H_A: \text{La distribució de la diferència no és normal} \end{cases}$$

Com que el valor p=0,02421 < 0,05, rebutgem H_0 (normalitat). Per tant, utilitzem mètodes no paramètrics.

```
> Triglicerids$dif <- with(Triglicerids, abans- després)

> normalityTest(~dif, test="shapiro.test", data=Triglicerids)

        Shapiro-Wilk normality test

data:  dif
W = 0.77722, p-value = 0.02421
```

Plantegem el contrast d'hipòtesis:

$$\begin{cases} H_0: \text{El nivell de triglicèrids abans del programa és menor o igual que el nivell després} \\ H_A: \text{El nivell de triglicèrids abans del programa és major que el nivell després} \end{cases}$$

```
> with(Triglicerids, median(dif, na.rm=TRUE))
[1] 0.3

> with(Triglicerids, wilcox.test(dif, alternative='greater', mu=0.0))

        Wilcoxon signed rank exact test

data:  dif
V = 28, p-value = 0.007813
alternative hypothesis: true location is greater than 0
```

Per a un nivell de significació α = 0,05, rebutgem H_0. Hi ha suficient evidència per a afirmar que, en general, el nivell de triglicèrids abans del programa d'exercici és major que després.

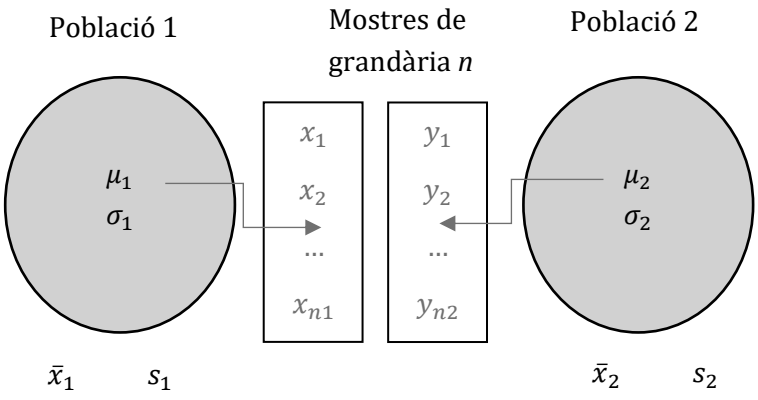
Exemple (triglicèrids). Continuació

Calculem l'interval de confiança per a la diferència mediana poblacional.

```
> with(Triglicerids, median(abans - després, na.rm=TRUE)) # median difference
[1] 0.3

> with(Triglicerids, wilcox.test(abans, després, alternative='two.sided',
 paired=TRUE, conf.int=TRUE))

        Wilcoxon signed rank exact test

data:  abans and després
V = 28, p-value = 0.01563
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.09 1.67
sample estimates:
(pseudo)median
        0.6075
```

Per a un nivell de confiança de 95%, esperem que la mediana poblacional de la diferència siga positiva i que els seus valors es troben entre 0,09 i 1,67.

2. MOSTRES INDEPENDENTS



En un disseny d'experiments amb observacions independents:

- Les observacions de cada mostra han de ser independents entre si.
- Les dues mostres han de ser independents l'una respecte de l'altra. És a dir, les unitats observacionals que formen la primera mostra no tenen cap relació amb les unitats observacionals de la segona mostra.

Per tant, la grandària de les mostres, n_1 i n_2, pot ser distinta.

Però, com calculem intervals de confiança o resolem un contrast d'hipòtesis?

Si volem aplicar mètodes paramètrics, fa falta que n_1 i n_2 siguen suficientment grans o comprovar la normalitat de la variable d'interès en ambdues poblacions. En aquest cas, els mètodes paramètrics (basats en la distribució t de Student) ens permeten:

- Calcular intervals de confiança per a la diferència de mitjanes.
- Contrastar hipòtesis sobre els valors de les dues mitjanes poblacionals.

Quan els mètodes paramètrics no són vàlids utilitzarem mètodes no paramètrics: **test de Wilcoxon per a mostres independents**, que ens permeten resoldre contrastos d'hipòtesis per a la igualtat de distribucions.

Sabem que la diferència entre les mitjanes mostrals, $\bar{x}_1 - \bar{x}_2$, és una estimació puntual de la diferència entre les mitjanes poblacionals. Aquesta estimació està subjecta a un error, l'**error estàndard**, que ens indica la precisió de l'estimació. Com calculem l'error estàndard per a la diferència de dues mitjanes mostrals? Aquest error dependrà de la variabilitat en les poblacions i de la grandària de les dues mostres. És a dir, l'estimació de l'error estàndard a partir de les dades de les dues mostres independents depèn de les variàncies poblacionals. Si les variàncies són iguals aplicarem el mètode combinat per al càlcul de l'error. En cas contrari, usarem el mètode no combinat. D'aquesta manera, i suposant que es compleixen les condicions de validesa dels tests paramètrics, l'interval de confiança per a la diferència de les mitjanes poblacionals ens indicarà d'una forma més senzilla quina és la precisió de l'estimació.

Definim l'error estàndard per a la diferència de dues mitjanes mostrals per a mostres independents com a:

$$SE_{\bar{x}_1-\bar{x}_2} = \begin{cases} \text{no combinat:} & \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \\ \text{combinat:} & \sqrt{\frac{s_c^2}{n_1}+\frac{s_c^2}{n_2}} \quad on \quad s_c = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \end{cases}$$

Si les desviacions típiques de les poblacions són iguals, $\sigma_1 = \sigma_2$, aplicarem el mètode combinat. En cas contrari, hem d'usar s_1 per a estimar σ_1 i s_2 per a estimar σ_2, és a dir, el mètode no combinat.

Com sabem si les desviacions típiques són iguals (o similars)?

Per a saber-ho utilitzarem el **test de Levene**:

$$\begin{cases} H_0: \text{Les desv. típiques de les dues poblacions són iguals } (\sigma_1 = \sigma_2) \\ H_A: \text{Les desv. típiques de les dues poblacions són diferents } (\sigma_1 \neq \sigma_2) \end{cases}$$

R-Commander:

Estadístics → Variàncies → Test de Levene

- Si valor p < α, rebutgem la igualtat de desv. típiques (H_0).
- Si valor p $\geq \alpha$, no rebutgem la igualtat de desv. típiques (H_0).

Interval de confiança per a la diferència de dues mitjanes poblacionals

L'interval de confiança al $100(1-\alpha)\%$ per a $\mu_1 - \mu_2$ es calcula utilitzant l'expressió:

$$IC_{100(1-\alpha)\%}(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{1-\frac{\alpha}{2}} SE_{\bar{x}_1-\bar{x}_2},$$

On $t_{1-\frac{\alpha}{2}}$ es correspon amb el percentil $1-\frac{\alpha}{2}$ d'una t de Student. L'error estàndard de la diferència de les mitjanes mostrals $SE_{\bar{x}_1-\bar{x}_2}$, s'obté utilitzant el mètode adequat en cada cas, tal com s'ha explicat abans (després d'aplicar el test de Levene).

R-Commander:

Estadístics → Mitjanes → Test t per a mostres independents *(on seleccionarem si les variàncies són iguals (var.equal = TRUE) o diferents (var.equal = FALSE))*

Contrast d'hipòtesis

El contrast (bilateral i unilateral) per a la comparació de dues mitjanes poblacionals es defineix de la manera següent:

Contrast bilateral	Constrast unilateral
$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_A: \mu_1 \neq \mu_2 \end{cases}$	$\begin{cases} H_0: \mu_1 \leq \mu_2 \\ H_A: \mu_1 > \mu_2 \end{cases}$ $\quad$ $\begin{cases} H_0: \mu_1 \geq \mu_2 \\ H_A: \mu_1 < \mu_2 \end{cases}$

L'estadístic de contrast utilitzat per a resoldre el contrast d'hipòtesis és $t_s = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}}$, on l'error estàndard de la diferència de les mitjanes mostrals s'obté utilitzant el mètode adequat (combinat o no combinat).

Criteri de decisió: rebutjar H_0 si el valor p < α.

Alternativa no paramètrica: test de Wilcoxon

Si no es compleixen les condicions de validesa dels mètodes basats en la t de Student, hem d'aplicar un test no paramètric per a mostres independents: **el test de Wilcoxon**, que ens permet comprovar si hi ha diferències entre les distribucions poblacionals de les dues mostres independents. És a dir, podem contrastar que:

$$\begin{cases} H_0: \text{La distribució de les variables } X_1 \text{ i } X_2 \text{ és la mateixa.} \\ H_A: \text{La distribució de les variables } X_1 \text{ i } X_2 \text{ NO és la mateixa.} \end{cases}$$

R-Commander:

Estadístics → Test no paramètrics → Test de Wilcoxon per a dues mostres

Podem calcular l'interval de confiança per a la mediana de la diferència, mitjançant el test de Wilcoxon. Per al seu càlcul és necessari utilitzar una H_A bilateral ("two.sided") i afegir l'opció **conf.int = TRUE.**

Exemple (dietes)

En un experiment per a comparar dues dietes dissenyades per a engreixar vaques productores de carn s'utilitzaren dos grups de 9 vaques del mateix estable, que foren alimentades respectivament amb les dues dietes. En la taula següent es mostra el pes guanyat pels animals durant un període de 70 dies.

Dieta	dieta1	dieta1	dieta1	dieta1	dieta1	dieta1	dieta1	dieta1	dieta1
Pes guanyat	596	422	524	454	538	552	478	564	556

Dieta	dieta2	dieta2	dieta2	dieta2	dieta2	dieta2	dieta2	dieta2	dieta2
Pes guanyat	498	460	468	458	530	482	528	598	456

Resum numèric *R-Commander*:

```
> numSummary(Dietes[,"Pes_guanyat", drop=FALSE],
  groups=Dietes$Dieta, statistics=c("mean",
  "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
           mean       sd IQR  0% 25% 50% 75% 100% Pes_guanyat:n
Dieta1 520.4444 57.11198  78 422 478 538 556  596             9
Dieta2 497.5556 47.28401  68 456 460 482 528  598             9
```

Com que les dues mostres són menudes, apliquem el test de normalitat a totes dues per saber si podem utilitzar mètodes paramètrics:

R-Commander:

Estadístics → Resums → Test de normalitat (opció: test per grups)

```
> normalityTest(Pes_guanyat ~ Dieta,
  test="shapiro.test", data=Dietes)

 --------
 Dieta = Dieta1

        Shapiro-Wilk normality test

data:  Pes_guanyat
W = 0.93523, p-value = 0.5327

 --------
 Dieta = Dieta2

        Shapiro-Wilk normality test

data:  Pes_guanyat
W = 0.84952, p-value = 0.07362
```

El valor p per al grup Dieta1 és 0,5327 > α, no es rebutja H_0 (assumim normalitat). El valor p per al grup Dieta2 és 0,07362 > α, no es rebutja H_0 (assumim normalitat).

Podem assumir que les dues distribucions del guany de pes són normals; utilitzarem mètodes paramètrics per a comparar-les.

Exemple (dietes)

Comprovem ara la hipòtesi d'igualtat de variàncies poblacionals:

```
> leveneTest(Pes_guanyat ~ Dieta, data=Dietes, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  0.5207 0.4809
      16
```

Com que el valor p = 0,4809 > α, assumim que les variàncies són iguals (error combinat).

```
> t.test(Pes_guanyat~Dieta, alternative='two.sided', conf.level=.95, var.equal=TRUE,
+   data=Dietes)

        Two Sample t-test

data:  Pes_guanyat by Dieta
t = 0.92611, df = 16, p-value = 0.3681
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -29.50493  75.28270
sample estimates:
mean in group Dieta1 mean in group Dieta2
            520.4444             497.5556
```

L'interval de confiança al 95% per a $\mu_1 - \mu_2$ és $IC_{95\%}(\mu_1 - \mu_2)$ = [-29,50, 75,28].

Ara, volem saber si la Dieta 1 és millor que la Dieta 2, respecte al guany de pes:

$$\begin{cases} H_0: \text{El pes guanyat amb la Dieta 1 no és major que amb la Dieta 2 } (\mu_1 \leq \mu_2) \\ H_A: \text{El pes guanyat amb la Dieta 1 és major que amb la Dieta 2 } (\mu_1 > \mu_2) \end{cases}$$

```
> t.test(Pes_guanyat~Dieta, alternative='greater', conf.level=.95, var.equal=TRUE,
+   data=Dietes)

        Two Sample t-test

data:  Pes_guanyat by Dieta
t = 0.92611, df = 16, p-value = 0.1841
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -20.26092       Inf
sample estimates:
mean in group Dieta1 mean in group Dieta2
            520.4444             497.5556
```

Com que el valor p = 0,1841 > α (= 0,05), no rebutgem H_0. No hi ha prou evidència per a afirmar que la dieta 1 és millor.

Exemple (BEH)

La β-endorfina humana (BEH) és una hormona secretada per la glàndula pituïtària en condicions d'estrès. Un fisiòleg mesurà (en pg/mL) la concentració de β-endorfina en sang en situació de repòs en dos grups d'homes. En el grup 1, els 11 homes havien estat fent jòguing regularment. El grup 2 consistia en 15 homes que acabaven d'entrar en un programa d'exercici.

Grup	Jòguing	Jòguing	Jòguing	Jòguing	...
Concentració	39	40	32	60	...
Grup	Principiants	Principiants	Principiants	Principiants	...
Concentració	70	47	54	27	...

Resum numèric *R-Commander*:

```
> numSummary(BEH[,"Concentració", drop=FALSE], groups=BEH$Grup,
  statistics=c("mean", "sd",
  "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean       sd  IQR 0% 25% 50%  75% 100% Concentració:n
Jòguing      35.72727 13.40217 10.5 13  30  37 40.5   60             11
Principiants 38.73333 16.06001 19.0  9  29  41 48.0   70             15
```

Definim les variables d'estudi com a:

X_1: Concentració de BEH en el grup jòguing.

X_2: Concentració de BEH en el grup de principiants.

Com que les dues mostres són menudes, apliquem el test de normalitat a totes dues per saber si podem utilitzar mètodes paramètrics:

```
> normalityTest(Concentració ~ Grup,
 test="shapiro.test", data=BEH)

 --------
 Grup = Jòguing

        Shapiro-Wilk normality test

data:  Concentració
W = 0.97431, p-value = 0.9264

 --------
 Grup = Principiants

        Shapiro-Wilk normality test

data:  Concentració
W = 0.99588, p-value = 1
```

Com que en els dos casos el valor p és major que α = 0,05, no rebutgem H_0. Podem assumir normalitat i per tant aplicar mètodes paramètrics.

Exemple (BEH)

Comprovem ara la hipòtesi d'igualtat de variàncies poblacionals:

```
> leveneTest(Concentració ~ Grup, data=BEH, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  0.4678 0.5005
      24
```

Com que el valor p = 0,5005 > α, no rebutgem H_0 i assumim que les variàncies són iguals (error combinat).

Ara, volem saber si hi ha diferències en la concentració de BEH depenent del grup:

$$\begin{cases} H_0: \text{La concentració mitjana de BEH és igual en els dos grups } (\mu_1 = \mu_2) \\ H_A: \text{Hi ha diferències en la concentració mitjana de BEH } (\mu_1 \neq \mu_2) \end{cases}$$

```
> t.test(Concentració~Grup, alternative='two.sided', conf.level=.95,
  var.equal=TRUE,  data=BEH)

        Two Sample t-test

data:  Concentració by Grup
t = -0.50452, df = 24, p-value = 0.6185
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.303374   9.291253
sample estimates:
     mean in group Jòguing mean in group Principiants
                  35.72727                   38.73333
```

Com que el valor p = 0,6185 > α (= 0,05), no rebutgem H_0. No hi ha prou evidència per a afirmar que hi ha diferències.

L'interval de confiança al 95% per a $\mu_1 - \mu_2$ és $IC_{95\%}(\mu_1 - \mu_2)$ = [-15,3; 9,29], que inclou el 0. Això indica que no podem rebutjar la hipòtesi nul·la.

Exemple (ratolins entrenats)

La farmacocinètica d'amikacina ha estat estudiada en humans i en diverses espècies d'animals, però el seu paper en animals durant l'exercici muscular és poc conegut. Per tal de determinar la farmacocinètica d'amikacina en ratolins amb entrenament físic i sense, es va fer un experiment amb 147 ratolins, dels quals 61 foren entrenats nadant 20 minuts diaris durant 7 setmanes (Grup 1: entrenat), mentre que els altres 86 ratolins quedaren sense entrenament (Grup 2: no entrenat).

El següent resum numèric mostra els estadístics descriptius del nivell enzimàtic d'aminotransferasa (AST) per als 147 ratolins.

```
> numSummary(Ratolins[,"AST", drop=FALSE], groups=Ratolins$Grup, statistics=c("mean", "sd",
+   "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean       sd       IQR         0%      25%      50%      75%     100% AST:n
entrenat    182.8275 21.43341  29.99833 132.364032 167.8479 180.2913 197.8463 236.6228    61
no entrenat 149.0654 73.89426 107.57938   1.524023 100.4310 138.1111 208.0104 337.7193    86
```

Definim les variables d'estudi com a:

X_1: Nivell d'AST en ratolins entrenats.

X_2: Nivell d'AST en ratolins no entrenats.

Com que les dues mostres són suficientment grans, podem utilitzar directament mètodes paramètrics.

Comprovem ara la hipòtesi d'igualtat de variàncies poblacionals:

```
> leveneTest(AST ~ Grup, data=Ratolins, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
       Df F value    Pr(>F)
group   1  57.269 4.042e-12 ***
      145
```

Com que el valor p és aproximadament 0 < α, rebutgem H_0 i assumim variànces distintes (mètode no combinat). Plantegem el contrast d'hipòtesis:

$$\begin{cases} H_0: \text{El nivell mitjà d'AST en ratolins entrenats és el mateix que en els no entrenats } (\mu_1 = \mu_2) \\ H_A: \text{El nivell mitjà d'AST en ratolins entrenats és diferent que en els no entrenats } (\mu_1 \neq \mu_2) \end{cases}$$

Resolem el contrast d'hipòtesis utilitzant el test t per a mostres independents sense igualtat de variàncies, també anomenat **test de Welch**:

```
> t.test(AST~Grup, alternative='two.sided', conf.level=.95,
  var.equal=FALSE, data=Ratolins)

        Welch Two Sample t-test

data:  AST by Grup
t = 4.0062, df = 104.28, p-value = 0.000116
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 17.05048 50.47371
sample estimates:
   mean in group entrenat mean in group no entrenat
                 182.8275                 149.0654
```

Com que el valor p=0,0001 < 0,05, rebutgem H_0, hi ha evidències que el nivell mitjà d'AST és igual en els dos grups. L'interval de confiança al 95% per a la diferència de mitjanes és [17,05; 50,47].

Exemple (niacina)

En un experiment sobre l'alimentació de corders es va observar el guany en pes durant dues setmanes de 10 animals alimentats amb una dieta normal (Grup 2: Normal) i altres 10 alimentats amb una ració extra de niacina (Grup 1: Niacina). S'espera que la niacina augmente el guany de pes. No és possible pensar que puga disminuir-lo. Els resultats foren aquests:

```
> numSummary(Dades[,"Guany", drop=FALSE], groups=Dades$Dieta,
  statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
        mean       sd IQR 0%   25%  50%   75% 100% Guany:n
Niacina 14.2 5.202563 4.5  4 13.25 16.5 17.75   19      10
Normal  10.0 3.231787 4.5  5  8.00  9.5 12.50   15      10
```

Observem que les estimacions puntuals de les mitjanes poblacionals cauen en la direcció de l'objectiu de l'estudi, 14,2 > 10.

Com que les dues mostres són menudes, apliquem el test de normalitat a totes dues per saber si podem utilitzar mètodes paramètrics:

```
> normalityTest(Guany ~ Dieta, test="shapiro.test", data=Dades)

 --------
 Dieta = Niacina

        Shapiro-Wilk normality test

data:  Guany
W = 0.80846, p-value = 0.01837

 --------
 Dieta = Normal

        Shapiro-Wilk normality test

data:  Guany
W = 0.96418, p-value = 0.8323
```

.

Per al Grup 1, el valor p = 0,018 < α, rebutgem H_0, no tenim normalitat. Per al Grup 2, el valor p = 0,832 > α, no es rebutja H_0,, podem assumir normalitat. Com que no podem assumir que la distribució de la variable "guany de pes" siga normal en les dues poblacions, utilitzarem mètodes no paramètrics per a la seua comparació.

$$\begin{cases} H_0: \text{El guany de pes té la mateixa distribució en les dues poblacions} \\ H_A: \text{La niacina és eficaç per a incrementar el guany de pes.} \end{cases}$$

```
> wilcox.test(Guany ~ Dieta, alternative="greater", data=Dades)

        Wilcoxon rank sum test with continuity correction

data:  Guany by Dieta
W = 77, p-value = 0.02238
alternative hypothesis: true location shift is greater than 0
```

Per a un nivell de significació α = 0,05, rebutgem H_0. Hi ha evidència estadística que la niacina augmenta el guany de pes.

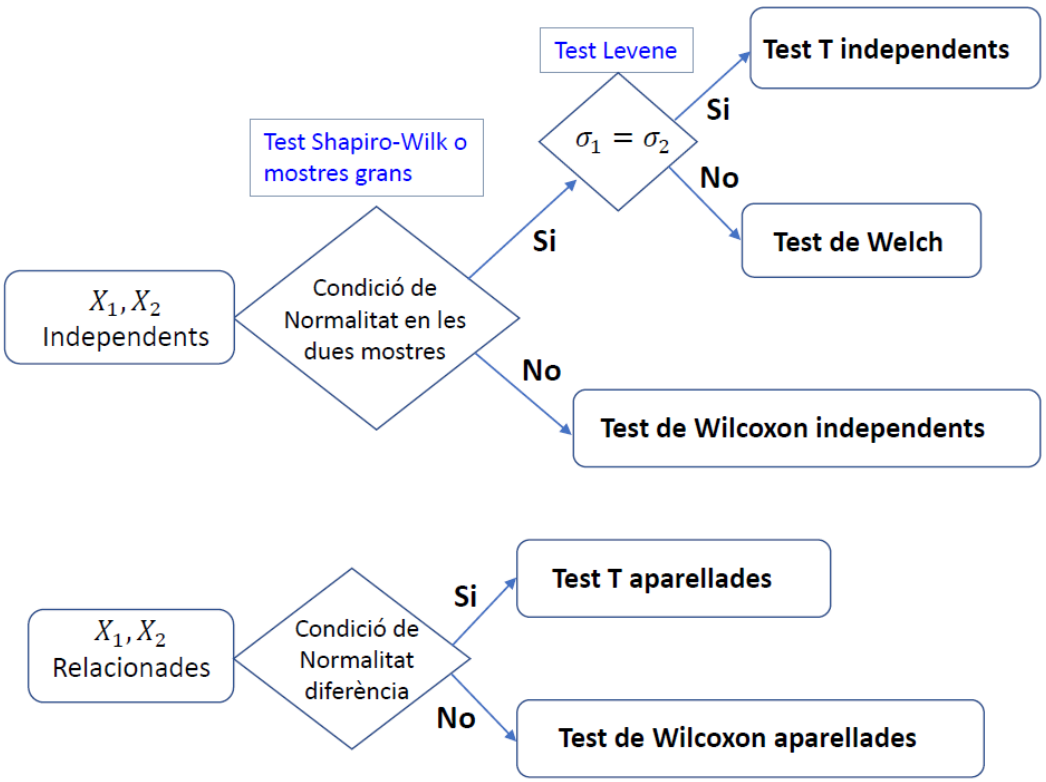
Exemple (niacina)

Per a poder calcular l'interval de confiança per a la diferència de les medianes poblacionals de dues mostres independents cal utilitzar **alternative = two.sided,** i afegir la instrucció **conf.int=TRUE**.

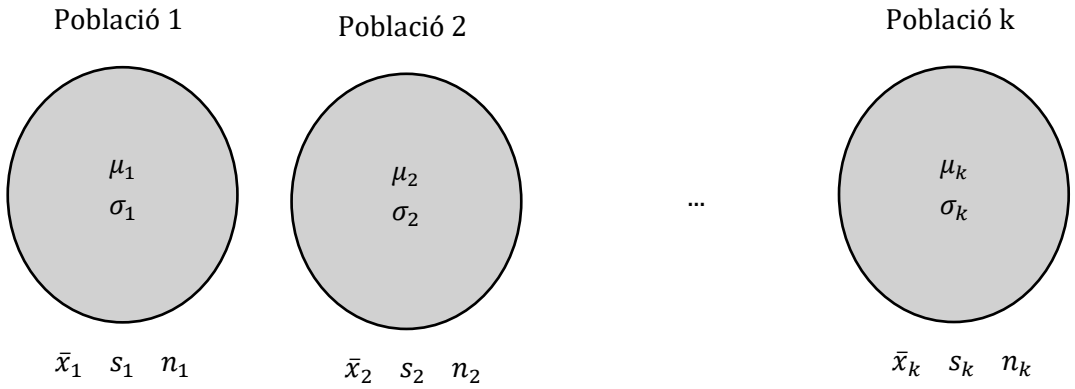
```
> wilcox.test(Guany ~ Dieta, alternative="two.sided", data=Dades,
+ conf.int=TRUE)

        Wilcoxon rank sum test with continuity correction

data:  Guany by Dieta
W = 77, p-value = 0.04475
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.000008320377 8.999986044230
sample estimates:
difference in location
              5.000064
```

L'eixida de *R-Commander* ens informa que la diferència dels valors de les medianes mostrals és de 5,000064. A més, amb un nivell de confiança del 95%, proporciona un interval de confiança per a la diferència de medianes poblacionals, [0,999957; 8,99999], i és sempre positiu, fet que indica major guany en el cas de la dieta enriquida amb niacina.

3. RESUM



TEMA 4: ANÀLISI ESTADÍSTICA DE K MOSTRES INDEPENDENTS



En un disseny d'experiments amb observacions independents:

- les mostres de cada grup han de poder considerar-se mostres aleatòries de les seues respectives poblacions;
- les mostres han de ser totes independents entre si.

En aquest tema veurem com resoldre contrastos d'hipòtesis amb k mostres independents.

1. COMPARACIÓ DE MITJANES: ANOVA

Ens preguntem: són les dades compatibles amb la hipòtesi que les mitjanes poblacionals de les k poblacions són iguals? És a dir, el contrast que volem resoldre és de la forma:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A: \text{No es compleix } H_0 \end{cases}$$

El procediment incorrecte seria comparar les mitjanes poblacionals dels grups dos a dos amb un test t, ja que perdem informació de la interacció global.

Per poder aplicar **ANOVA** s'ha de complir:

- Totes les distribucions poblacionals han de complir la normalitat (totes normals o mostres prou grans) i han de tenir la mateixa variància (o desviació típica).
- Si les distribucions poblacionals compleixen la normalitat, però no es compleix la condició d'igualtat de les variàncies, utilitzarem el **test de Welch**.
- Si alguna de les distribucions poblacionals no compleix la normalitat, utilitzarem un mètode no paramètric per a comparar les k poblacions, el **test de Kruskal-Wallis**.

Per comprovar si les variàncies (desviacions típiques) són semblants, apliquem el **test de Levene**:

$$\begin{cases} H_0: \text{Les desv. típiques de totes les poblacions són iguals } (\sigma_1 = \sigma_2 = \cdots = \sigma_k) \\ H_A: \text{No es compleix } H_0 \end{cases}$$

R-Commander:

Estadístics → Variàncies → Test de Levene

ANOVA

Indiquem a continuació la notació per al càlcul de l'estadístic que utilitza ANOVA.

k: nombre de grups.

n_i: nombre de observacions en el grup i, i=1,2,...,k.

$n = \sum_{i=1}^{k} n_i$

y_{ij}: observació j del grup i, on j=1, 2,..., n_i.

$\bar{y}_i$: mitjana mostral del grup i.

$\bar{y}$: mitjana mostral de totes les observacions

$s_i = \sqrt{\frac{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{n_i-1}}$ desviació típica mostral del grup i.

Per tant, a partir de la variabilitat dins dels grups i la variabilitat entre els grups podem calcular la variabilitat total:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k}\left(\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2\right) + \sum_{i=1}^{k}\left(\sum_{j=1}^{n_i}(\bar{y}_i-\bar{y})^2\right)$$

D'aquesta manera, la informació de la variabilitat que obtenim de totes les dades sol expressar-se en forma de taula, coneguda com taula ANOVA:

	Variabilitat. Suma de quadrats (SQ)	Graus de llibertat (gl)	Quadrats mitjans (QM)
Entre grups	$\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2$	$k-1$	SQ(entre grups)/gl
Intragrups	$\sum_{i=1}^{k}(n_i-1)s_i^2$	$n-k$	SQ(intragrups)/gl
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$	$n-1$	

Passos que cal seguir per resoldre un contrast amb ANOVA.

1. Seleccionem el nivell de significació, α.
2. Calculem el valor de l'estadístic del contrast: $F_S = \frac{QM(entre)}{QM(intra)}$.
    - Si H_0 és certa, aleshores $F_s \approx 1$.

- Si H_0 és falsa, aleshores $F_s \gg 1$.

3. Calculem el p-valor del contrast.
   - Si p-valor $< \alpha$, rebutgem H_0.

R-Commander:

Estadístics → Mitjanes → ANOVA d'un factor

En *R-Commander* tenim l'opció de realitzar comparacions dos a dos de les mitjanes poblacionals. Aquesta opció només té sentit si rebutgem la hipòtesi nul·la d'igualtat de mitjanes poblacionals. Per fer aquestes comparacions apliquem el **test de Tukey**.

R-Commander:

Estadístics → Mitjanes → ANOVA d'un factor (seleccionar "comparacions dos a dos de les mitjanes").

Test de Welch

Si les desviacions típiques de les poblacions són iguals, $\sigma_1 = \sigma_2 = \cdots = \sigma_k$, aplicarem ANOVA. En cas contrari, aplicarem el test de Welch.

R-Commander:

Estadístics → Mitjanes → ANOVA d'un factor (seleccionar "Welch F-test").

En aquest cas, per a fer les comparacions dos a dos utilitzarem la instrucció següent:

pairwise.t.test(Dataset$Variable,Dataset$Grupos, pool.sd=FALSE)

Alternativa no paramètrica: Test de Kruskal-Wallis

Si no es compleixen les condicions de validesa dels mètodes basats en la t de Student, hem d'aplicar un test no paramètric: **el test de Kruskal-Wallis**, que ens permet comprovar si hi ha diferències entre les distribucions poblacionals de les mostres. És a dir, podem contrastar que:

$$\begin{cases} H_0: \text{La distribució en totes les variables és la mateixa.} \\ H_A: \text{La distribució en totes les variables NO és la mateixa.} \end{cases}$$

R-Commander:

Estadístics → Test no paramètrics → Test de Kruskal-Wallis

En aquest cas, per a fer les comparacions dos a dos utilitzarem la instrucció següent:

pairwise.wilcox.test(Dataset$Variable,Dataset$Grupos, p.adjust="bonf")

Exemple (pes guanyat en corders)

En la taula següent es presenta el pes (en lliures) guanyat en dues setmanes per corders alimentats amb tres dietes diferents:

Dieta 1	8	16	9		
Dieta 2	9	16	21	11	18
Dieta 3	15	10	17	6	

Resum numèric *R-Commander*:

```
> numSummary(Pes_Corders[,"Guany", drop=FALSE], groups=Pes_Corders$Dieta,
  statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
        mean       sd IQR 0%  25%  50%  75% 100% Guany:n
Dieta 1   11 4.358899 4.0  8  8.5  9.0 12.5   16       3
Dieta 2   15 4.949747 7.0  9 11.0 16.0 18.0   21       5
Dieta 3   12 4.966555 6.5  6  9.0 12.5 15.5   17       4
```

Les grandàries de les mostres són: $n_1 = 3, n_2 = 5, n_3 = 4$. Per tant, hem de comprovar-ne la normalitat, en totes. Fem el test de Shapiro-Wilk:

```
Dieta = Dieta1

        Shapiro-Wilk normality test

data:  Pes
W = 0.84211, p-value = 0.2196
```

```
Dieta = Dieta2

        Shapiro-Wilk normality test

data:  Pes
W = 0.95322, p-value = 0.7602
```

```
Dieta = Dieta3

        Shapiro-Wilk normality test

data:  Pes
W = 0.9516, p-value = 0.7262
```

En els tres casos el p-valor > 0,05 i, per tant, no rebutgem H_0, podem assumir normalitat.

Una vegada contrastada la normalitat en totes les mostres, passem a la segona comprovació: podem assumir que totes les variàncies (o desviacions típiques) poblacionals són iguals?

$$\begin{cases} H_0: \sigma_1 = \sigma_2 = \sigma_3 \\ H_A: \text{No es compleix } H_0 \end{cases}$$

```
> leveneTest(Guany ~ Dieta, data=Pes_Corders, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  2  0.1378 0.8731
       9
```

Com que el p-valor = 0,8731 > 0,05, no es rebutja H_0 i, per tant, assumim igualtat de variàncies. Podem aplicar el procediment ANOVA.

Exemple (pes guanyat en corders)

Definim la variable X: pes guanyat pels corders. Es vol analitzar estadísticament l'efecte dieta. És a dir, volem contrastar:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_A: \text{No es compleix } H_0 \end{cases}$$

```
> AnovaModel.1 <- aov(Guany ~ Dieta, data=Pes_Corders)

> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value Pr(>F)
Dieta        2     36   18.00   0.771  0.491
Residuals    9    210   23.33
```

Com que p-valor = 0,491 > 0,05, no hi ha prou evidència estadística per a rebutjar H_0; per tant, **assumim que les tres dietes tenen el mateix efecte.**

Exemple (β-endorfina)

La β-endorfina humana és una hormona segregada per la glàndula pituïtària sota condicions d'estrès. Un fisiòleg de l'exercici va mesurar la concentració de β-endorfina en sang en situació de repòs en tres grups d'homes. El grup 1 consistia en 11 homes que havien estat fent jòguing regularment durant un cert període de temps. El grup 2 consistia en 15 homes que anaven a iniciar un programa d'exercici. El grup 3 estava format per 10 homes sedentaris. Els resultats, en pg/ml, van ser els següents:

Jòguing	39	40	32	60	19			
	52	41	32	13	37	28		
Principiants	70	47	54	27	31	42	37	
	41	9	18	33	23	49	41	59
Sedentaris	46	39	41	59	62			
	78	41	48	52	43			

Resum numèric *R-Commander*:

```
> numSummary(BEH[,"beta.endorfina", drop=FALSE], groups=BEH$grup,
  statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean       sd   IQR 0%  25% 50%   75% 100% beta.endorfina:n
Jòguing      35.72727 13.40217 10.50 13 30.0  37 40.50   60               11
Principiants 38.73333 16.06001 19.00  9 29.0  41 48.00   70               15
Sedentaris   50.90000 12.27871 15.75 39 41.5  47 57.25   78               10
```

Les grandàries de les mostres són: $n_1 = 11, n_2 = 15, n_3 = 10$. Per tant, hem de comprovar-ne la normalitat, en totes. Fem el test de Shapiro-Wilk:

En els tres casos el p-valor > 0,05 i, per tant, no rebutgem H_0 i assumim normalitat.

```
> normalityTest(beta.endorfina ~ grup,
 test="shapiro.test", data=BEH)

 --------
 grup = Jòguing

        Shapiro-Wilk normality test

data:  beta.endorfina
W = 0.97431, p-value = 0.9264

 --------
 grup = Principiants

        Shapiro-Wilk normality test

data:  beta.endorfina
W = 0.99588, p-value = 1

 --------
 grup = Sedentaris

        Shapiro-Wilk normality test

data:  beta.endorfina
W = 0.8684, p-value = 0.09576
```

Exemple (β-endorfina)

Una vegada contrastada la normalitat en totes les mostres, passem a la segona comprovació: podem assumir que totes les variàncies (o desviacions típiques) poblacionals són iguals?

$$\begin{cases} H_0: \sigma_1 = \sigma_2 = \sigma_3 \\ H_A: \text{No es compleix } H_0 \end{cases}$$

```
> leveneTest(beta.endorfina ~ grup, data=BEH, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  2  0.4339 0.6516
      33
```

Com que el p-valor = 0,6516 > 0,05, no es rebutja H_0 i, per tant, assumim igualtat de variàncies. Podem aplicar el procediment ANOVA.

Definim les variables de l'estudi:

Y_1: concentració de β-endorfina en sang en persones que han estat fent jòguing regularment durant un temps, $Y_1 \sim N(\mu_1, \sigma)$.

Y_2: concentració de β-endorfina en sang en persones que acaben de començar un programa d'exercici físic, $Y_2 \sim N(\mu_2, \sigma)$.

Y_3: concentració de β-endorfina en sang en persones sedentàries, $Y_3 \sim N(\mu_3, \sigma)$.

Són les dades compatibles amb la hipòtesi que les mitjanes de les tres poblacions són iguals ($H_0: \mu_1 = \mu_2 = \mu_3$)?

```
> AnovaModel.2 <- aov(beta.endorfina ~ grup, data=BEH)

> summary(AnovaModel.2)
            Df Sum Sq Mean Sq F value Pr(>F)
grup         2   1362   680.8   3.322 0.0485 *
Residuals   33   6764   205.0
---
```

Per a un nivell de significació $\alpha = 0{,}1$, rebutgem H_0. Hi ha prou evidència per afirmar que les concentracions mitjanes de β-endorfina no són les mateixes en els tres grups.

Exemple (β-endorfina)

A continuació fem les comparacions dos a dos per decidir quins grups són els que poden tenir la mateixa mitjana poblacional.

```
        Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = beta.endorfina ~ grup, data = BEH)

Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
Principiants - Jòguing == 0    3.006      5.683   0.529   0.8575
Sedentaris - Jòguing == 0     15.173      6.255   2.426   0.0531 .
Sedentaris - Principiants == 0   12.167      5.845   2.082   0.1090
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Podem observar que a nivell 0,1, hi ha diferències significatives entre el grup 1 i el grup 3 (jòguing i sedentaris, respectivament).

L'eixida del test ANOVA també ens proporciona estimacions puntuals i intervals de confiança al 95% de les diferències de les mitjanes poblacionals dos a dos:

```
        Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = beta.endorfina ~ grup, data = BEH)

Quantile = 2.4536
95% family-wise confidence level


Linear Hypotheses:
                             Estimate lwr      upr
Principiants - Jòguing == 0    3.0061 -10.9380  16.9501
Sedentaris - Jòguing == 0     15.1727  -0.1755  30.5209
Sedentaris - Principiants == 0  12.1667  -2.1740  26.5073
```

Exemple (pressió arterial)

Es vol avaluar l'eficàcia de distintes dosis d'un fàrmac contra la hipertensió arterial, comparant-la amb la d'una dieta pobre en sal. Per fer-ho, se seleccionen a l'atzar 25 hipertensos i es distribueixen aleatòriament en 5 grups. Als del primer grup, no se'ls subministra cap tractament; als del segon, una dieta amb un contingut pobre en sal; als del tercer, una dieta sense sal; als del quart, una dosi determinada del fàrmac; i als del cinquè una dosi més elevada del fàrmac. Les pressions arterials sistòliques dels 25 individus, en acabar el tractament, són:

Control (grup 1)	180	178	179	182	181
Poca sal (grup 2)	172	152	167	160	180
Sense sal (grup 3)	163	170	158	162	170
Fàrmac (grup 4)	158	155	160	161	157
Més fàrmac (grup 5)	147	152	143	155	160

Resum numèric *R-Commander*:

```
> numSummary(Dades[,"Pressió", drop=FALSE], groups=Dades$Grup,
  statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
       mean        sd IQR  0% 25% 50% 75% 100% Pressió:n
Grup1 180.0  1.581139   2 178 179 180 181  182         5
Grup2 166.2 10.779610  12 152 160 167 172  180         5
Grup3 164.6  5.272571   8 158 162 163 170  170         5
Grup4 158.2  2.387467   3 155 157 158 160  161         5
Grup5 151.4  6.655825   8 143 147 152 155  160         5
```

Com que la grandària de les 5 mostres és petita (menor que 30), hem de comprovar-ne la normalitat:

```
> normalityTest(Pressió ~ Grup, test="shapiro.test", data=Dades)

 --------
 Grup = Grup1

        Shapiro-Wilk normality test

data:  Pressió
W = 0.98676, p-value = 0.9672

 --------
 Grup = Grup2

        Shapiro-Wilk normality test

data:  Pressió
W = 0.99509, p-value = 0.9941

Grup = Grup3

        Shapiro-Wilk normality test

data:  Pressió
W = 0.88255, p-value = 0.321

 --------
 Grup = Grup4

        Shapiro-Wilk normality test

data:  Pressió
W = 0.97378, p-value = 0.8989

 --------
 Grup = Grup5

        Shapiro-Wilk normality test

data:  Pressió
W = 0.98773, p-value = 0.9712
```

Com que tots els p-valors són majors que 0,05, no rebutgem la hipòtesi nul·la i assumim normalitat.

Exemple (pressió arterial)

Una vegada contrastada la normalitat en totes les mostres, passem a la segona comprovació: podem assumir que totes les variàncies (o desviacions típiques) poblacionals són iguals?

$$\begin{cases} H_0: \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma_5 \\ H_A: \text{No es compleix } H_0 \end{cases}$$

```
> leveneTest(Pressió ~ Grup, data=Dades, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value  Pr(>F)
group  4  3.7991 0.01866 *
      20
```

Com que el p-valor és menor que 0,05, rebutgem H_0. Per tant, no podem assumir variàncies iguales i les considerem distintes. Hem d'aplicar el test de Welch.

Definim les variables de l'estudi:

Y_1: Pressió arterial sistòlica dels individus sense tractament, $Y_1 \sim N(\mu_1, \sigma_1)$.

Y_2: Pressió arterial sistòlica dels individus amb dieta pobre en sal, $Y_2 \sim N(\mu_2, \sigma_2)$.

Y_3: Pressió arterial sistòlica dels individus amb dieta sense sal, $Y_3 \sim N(\mu_3, \sigma_3)$.

Y_4: Pressió arterial sistòlica dels individus amb dosi 1 del fàrmac, $Y_4 \sim N(\mu_4, \sigma_4)$.

Y_5: Pressió arterial sistòlica dels individus amb dosi 2 del fàrmac, $Y_5 \sim N(\mu_5, \sigma_5)$.

Són les dades compatibles amb la hipòtesi que les mitjanes de les tres poblacions són iguals ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$)?

```
> oneway.test(Pressió ~ Grup, data=Dades) # Welch test

        One-way analysis of means (not assuming equal variances)

data:  Pressió and Grup
F = 72.79, num df = 4.000, denom df = 9.231, p-value = 0.0000005723
```

Per a un nivell de significació 0,05, rebutgem H_0. Hi ha prou evidència per afirmar que les pressions arterials mitjanes no són les mateixes en tots els grups.

A continuació, realitzem comparacions dos a dos per tal de veure quins grups sí que poden tenir la mateixa mitjana poblacional.

```
> pairwise.t.test(Dades$Pressió, Dades$Grup, pool.sd=FALSE)

        Pairwise comparisons using t tests with non-pooled SD

data:  Dades$Pressió and Dades$Grup

      Grup1     Grup2  Grup3  Grup4
Grup2 0.2249    -      -      -
Grup3 0.0151    0.7760 -      -
Grup4 0.0000064 0.3483 0.2249 -
Grup5 0.0039    0.2182 0.0634 0.2522

P value adjustment method: holm
```

Per $\alpha = 0{,}05$, rebutgem:

$H_0: \mu_1 = \mu_3$

$H_0: \mu_1 = \mu_3$

$H_0: \mu_1 = \mu_3$

Exemple (calci)

La taula següent mostra el contingut en calci (mg./100 grams de producte) de 17 productes lactis agrupats en tres categories: formatge, llet i iogurt.

Formatge	295	740	623,46	838,47	714,77	809,01	
Llet	120	114	110	183			
Iogurt	100	85	96	120	96	127	131

Resum numèric amb *R-Commander*:

```
> numSummary(Dades[,"Calci", drop=FALSE], groups=Dades$Grup, statistics=c("mean",
 "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5, .75,1))
              mean        sd    IQR  0%      25%     50%      75%   100% Calci:n
Formatge 670.1183 198.69633 145.47 295 646.2875 727.385 791.7575 838.47       6
Iogurt   107.8571  17.86457  27.50  85  96.0000 100.000 123.5000 131.00       7
Llet     131.7500  34.41293  22.75 110 113.0000 117.000 135.7500 183.00       4
```

Les grandàries de les mostres són: $n_1 = 4, n_2 = 6, n_3 = 7$. Per tant, hem de comprovar-ne la normalitat, en totes. Fem el test de Shapiro-Wilk:

```
> normalityTest(Calci ~ Grup, test="shapiro.test", data=Dades)

 --------
 Grup = Formatge

        Shapiro-Wilk normality test

data:  Calci
W = 0.82579, p-value = 0.09899

 --------
 Grup = Iogurt

        Shapiro-Wilk normality test

data:  Calci
W = 0.89697, p-value = 0.313

 --------
 Grup = Llet

        Shapiro-Wilk normality test

data:  Calci
W = 0.73695, p-value = 0.02893
```

Com que hi ha un p-valor (0,029) menor que 0,05, es rebutja la normalitat per a la població de llet. En aquest cas, no fa falta comprovar la igualtat de variàncies. Com que una distribució no és normal, utilitzarem el test no paramètric de Kruskal-Wallis per resoldre el contrast següent:

$$\begin{cases} H_0: \text{El contingut de calci en els tres grups és el mateix.} \\ H_A: \text{El contingut de calci en els tres grups NO és el mateix.} \end{cases}$$

Exemple (calci)

```
> Tapply(Calci ~ Grup, median, na.action=na.omit, data=Dades) # medians by group
Formatge   Iogurt     Llet
 727.385  100.000  117.000

> kruskal.test(Calci ~ Grup, data=Dades)

        Kruskal-Wallis rank sum test

data:  Calci by Grup
Kruskal-Wallis chi-squared = 11.494, df = 2, p-value = 0.003192
```

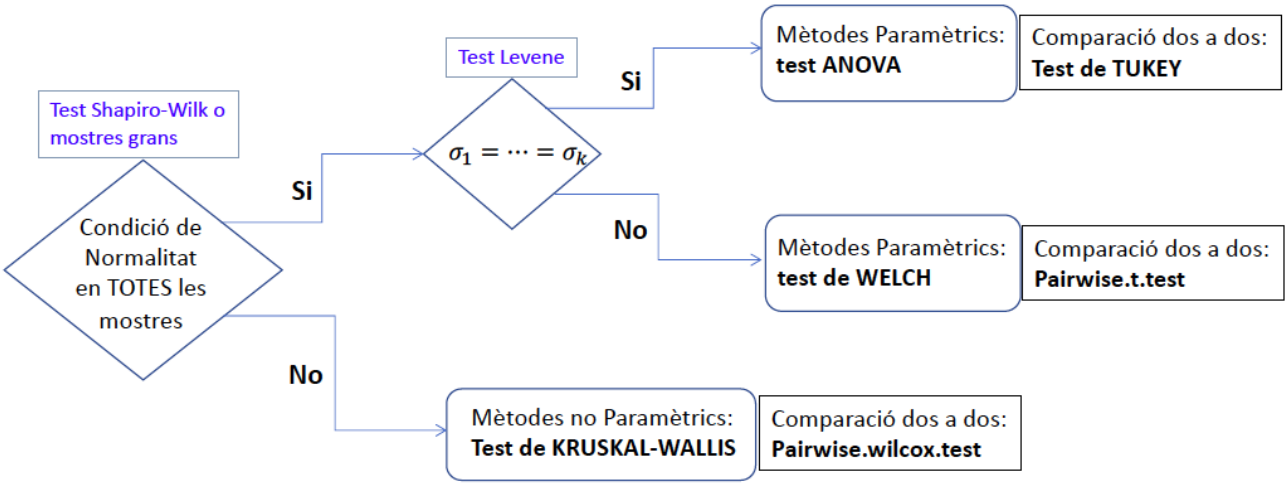
Per a un nivell de significació de 0,05, rebutgem H_0. Hi ha prou evidència per afirmar que el contingut de calci en els tres grups no és el mateix.

Ara bé, hi haurà semblances del contingut de calci si comparem els productes dos a dos?

```
> pairwise.wilcox.test(Dades$Calci, Dades$Grup, p.adjust="bonf")

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  Dades$Calci and Dades$Grup

       Formatge Iogurt
Iogurt 0.010    -
Llet   0.029    1.000

P value adjustment method: bonferroni
```

Per a $\alpha = 0{,}05$, el contingut de calci del grup “formatge” té una distribució distinta de la dels grups “llet” i “iogurt”.

2. RESUM



TEMA 5: ANÀLISI DE DADES CATEGÒRIQUES

Les dades categòriques representen atributs o categories a les quals pot pertànyer l'individu considerat. Aquest tipus de dades apareixen quan observem una variable categòrica X amb un nombre finit de possibles categories. Per exemple:

- Sexe: {M, F}.
- Fumador: {Sí, No}.
- Grup sanguini: {A, B, 0, AB}.
- Evolució d'un malalt: {millora, estable, empitjora}.

Per tant, podem considerar la població dividida en categories, cadascuna de les quals representa una proporció del total.

1. ANÀLISI D'UNA PROPORCIÓ POBLACIONAL

En particular, una variable dicotòmica és una variable categòrica que únicament té dos possibles valors, els quals solen representar-se mitjançant els valors {0, 1}, de manera que:

- X = 1 si es compleix la propietat que ens interessa (èxit).
- X = 0 si no es compleix (fracàs).

Com vam veure en el tema 1, la **distribució de Bernoulli** serveix per a descriure probabilísticament les variables dicotòmiques. $X \sim Ber(\pi)$, on π és la probabilitat d'èxit, és a dir, que la variable pren el valor 1. La funció de probabilitat és: $P(X = 1) = \pi$ i $P(X = 0) = 1 - \pi$. La mitjana és X π i la variància $\pi(1 - \pi)$.

Ens interessa la proporció d'individus en la població que compleixen una certa característica, és a dir, $P(X = 1) = \pi$. Aquesta proporció π és desconeguda i, per tant, haurem de fer una inferència estadística sobre ella. Com π és la mitjana de la variable X en la població, la mitjana mostral serà, per tant, una estimació puntual de π. Siga $\{x_1, x_2, \dots, x_n\}$ una mostra aleatòria de grandària n, en què es donen r observacions iguals a 1 i n-r observacions iguals a 0. La mitjana mostral coincideix amb la proporció de la categoria d'interès en la mostra:

$$\text{Estimació puntual de } \pi = \hat{\pi} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{r}{n}.$$

Com tota estimació puntual, l'estimació de la proporció de la categoria d'interès en la població estarà subjecta a un error, **l'error estàndard** de la proporció mostral.

$$SE_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

Per al càlcul de l'interval de confiança per a la proporció poblacional π utilitzarem el fet que r = "nombre d'èxits observats en la mostra" segueix una distribució binomial, $r \sim Bi(n, \pi)$.

R-Commander:

Estadístics → proporcions → test de proporcions per a una mostra

També podem escriure directament en R:

binom.test (r, n, alternative = 'two.sided', p = p0, conf.level = cl)

on r és el nombre d'èxits, n és el nombre total de proves, p0 és la proporció que es vol comparar (canviarem aquest valor per als contrastos d'hipòtesi) i conf.level és el nivell de confiança establert per al càlcul de l'interval de confiança.

Exemple (cefotaxima)

Tenim una mostra de 70 malalts amb ferides infectades per bacteris que foren tractats amb l'antibiòtic cefotaxima. La resposta bacteriològica (desaparició dels bacteris de la ferida) fou considerada satisfactòria en 59 malalts.

Quina és la probabilitat que l'antibiòtic siga efectiu?

Definim la variable aleatòria X, resposta bacteriològica, i considerem X = 1 les respostes bacteriològiques satisfactòries, és a dir:

X: "resposta bacteriològica amb el tractament cefotaxima",

$X \sim Ber(\pi)$ on π = P(èxit) = P(resposta satisfactòria) = P(X = 1).

Volem estimar P(X = 1) = π.

Una estimació puntual de π és

$$\hat{\pi} = \frac{r}{n} = \frac{59}{70} = 0{,}843.$$

Ara, entre quins valors, amb una confiança del 95%, es troba el valor de π?

```
> binom.test(59,70, alternative='two.sided', p=.5, conf.level=.95)

        Exact binomial test

data:  59 and 70
number of successes = 59, number of trials = 70, p-value = 0.000000004466
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.7362002 0.9188562
sample estimates:
probability of success
             0.8428571
```

L'interval de confiança del 95% per a π és [0,74; 0,92].

Exemple (proporció de neutròfils)

Per fer un recompte de glòbuls blancs s'estén uniformement una gota de sang sobre un suport de vidre, s'aplica tinta Wright i s'examina pel microscopi. Per a un determinat malalt, dels 200 glòbuls blancs comptabilitzats, 125 n'eren neutròfils (uns glòbuls blancs que es produeixen en la medul·la òssia, la funció dels quals és, en part, eliminar agents infecciosos en la sang).

Calcula un interval de confiança del 95% per a la proporció de neutròfils del malalt.

```
> binom.test(125,200, alternative='two.sided', p=.5, conf.level=.95)

        Exact binomial test

data:  125 and 200
number of successes = 125, number of trials = 200, p-value = 0.0004994
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5539486 0.6922862
sample estimates:
probability of success
                 0.625
```

La proporció de neutròfils d'aquest malalt està entre 0,55 i 0,69, amb una confiança del 95%.

Exemple (consum d'aliments amb soia)

Per estimar el percentatge d'individus de certa població que consumeix productes amb soia s'ha elegit una mostra de grandària 100 i s'ha observat que 18 dels individus de la mostra eren consumidors d'aquests productes.

Què podem dir sobre el percentatge de consumidors en la població? Podria ser del 30%?

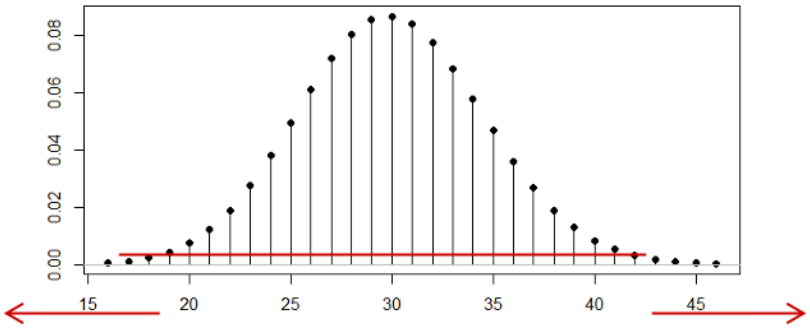
Estem interessats en una pregunta concreta sobre el paràmetre d'interès: proporció d'individus en la població que consumeixen productes de soia. Volem plantejar el contrast d'hipòtesis següent:

$$\begin{cases} H_0: \pi = 0{,}3 \\ H_A: \pi \neq 0{,}3 \end{cases}$$

Recordem que el p-valor es calcula com la probabilitat d'observar en la hipòtesi nul·la un succés igual o menys probable que el succés que hem observat.

En aquest cas, hem observat un nombre d'èxits r = 18 en una mostra de grandària n = 100, i volem contrastar $H_0: \pi = 0{,}3$. Per tant, en H_0 el valor més probable és 30.

Els successos iguals o menys probables que r = 18 són r = 0, ... , 18 i r = 42, ... ,100.



Per tant, el p-valor es calcula com la probabilitat P(r = 0, ... , 18, 42, ... , 100).

```
> binom.test(18,100, alternative='two.sided', p=.3, conf.level=.95)

        Exact binomial test

data:  18 and 100
number of successes = 18, number of trials = 100, p-value = 0.00849
alternative hypothesis: true probability of success is not equal to 0.3
95 percent confidence interval:
 0.1103112 0.2694771
sample estimates:
probability of success
                  0.18
```

El p-valor és 0,00849 < 0,05 i, per tant, rebutgem H_0. Tenim proves del fet que el percentatge de consumidors de productes amb soia és diferent de 30%.

Exemple (al·lèrgia primaveral)

Una empresa de productes farmacèutics afirma que un dels seus medicaments redueix considerablement els símptomes de l'al·lèrgia primaveral en el 90% de la població. Una associació de consumidors ha provat aquest fàrmac en una mostra de 200 persones, de les quals 170 han millorat notablement.

Podem concloure que l'afirmació de l'empresa farmacèutica és certa?

$$\begin{cases} H_0: \pi \geq 0{,}9 \\ H_A: \pi < 0{,}9 \end{cases}$$

```
> binom.test(170,200, alternative='less', p=.9, conf.level=.95)

        Exact binomial test

data:  170 and 200
number of successes = 170, number of trials = 200, p-value = 0.01633
alternative hypothesis: true probability of success is less than 0.9
95 percent confidence interval:
 0.0000000 0.8899186
sample estimates:
probability of success
                  0.85
```

El p-valor és 0,01633 < 0,05 i, per tant, rebutgem H_0. L'afirmació no és certa.

2. ANÀLISI DE BONDAT D'AJUST

En aquesta secció treballarem amb una variable categòrica X amb k categories (k ≥ 2). Treballarem amb una mostra aleatòria de grandària n d'una població, els elements de la qual pertanyen a una de les k categories en què està dividida la població: c_1, c_2, ... , c_k. Cada categoria c_i té associada una proporció poblacional π_i. **L'anàlisi de bondat d'ajust** ens permet contrastar les proporcions poblacionals (probabilitats) de totes les categories com una única hipòtesi:

$$\begin{cases} H_0\text{: } \pi_i = \pi_{i0} \text{ per a } i = 1, 2, \dots , \quad d \\ H_A\text{: hi ha almenys un } j\text{, tal que } \pi_j \neq \pi_{j0} \end{cases}$$

L'estadístic del test es basa en les diferències entre les freqüències observades i les esperades per a totes les categories:

$$\chi^2_s = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

on O_i és la freqüència observada en la categoria i, E_i és la freqüència esperada de la categoria i la suma s'estén a totes les categories.

Si la hipòtesi nul·la és certa, l'estadístic χ^2_s segueix una distribució Khi-quadrat amb k - 1 graus de llibertat, on k és el nombre de categories. Un valor "gran" de l'estadístic indica que les dades són incompatibles amb H_0.

Si el p-valor de les dades és menor que α, aleshores considerem que les dades són incompatibles amb H_0 i la rebutjarem.

La condició d'aplicabilitat del test Khi-quadrat és que **almenys el 80% de les freqüències esperades siguen majors que 5.**

Exemple (llavors de lli)

Les llavors de lli són un potent anticancerós, contenen una gran quantitat de fibra i àcids grassos omega 3 i tenen un efecte beneficiós en moltes malalties: inflamatòries, cardiovasculars, al·lèrgies, etc. S'ha fet un estudi sobre una mutació de llavors de lli amb la qual s'espera obtenir oli per a ús medicinal. La quantitat d'àcid palmític en la llavor {baixa, mitjana, alta} i el color d'aquesta {marró, jaspiada} eren factors importants de la investigació. D'acord amb les lleis de Mendel, les sis combinacions possibles haurien d'ocórrer en una proporció 3:6:3:1:2:1 (la suma és 16).

Color	Nivell àcid	Observat	Esperat
Marró	Baix	15	15,1875
Marró	Mitjà	26	30,375
Marró	Alt	14	15,1875
Jaspiada	Baix	6	5,0625
Jaspiada	Mitjà	8	10,1258
Jaspiada	Alt	12	5,0625
TOTAL		**81**	**81**

Per calcular les freqüències esperades calculem primer la probabilitat segons el model de Mendel: P(marró i baix) = 3/16 = 0,1875, i després multipliquem el valor pel total de llavors.

Freqüència esperada = 81*0,1875 = 15,1875.

Fixa't que és com fer una regla de tres.

Volem contrastar una hipòtesi nul·la que especifique les probabilitats poblacionals d'aquestes categories:

$$\begin{cases} H_0: \pi_1 = \frac{3}{16}, \pi_2 = \frac{6}{16}, \pi_3 = \frac{3}{16}, \pi_4 = \frac{1}{16}, \pi_5 = \frac{2}{16}, \pi_6 = \frac{1}{16} \\ H_A: hi\ ha \text{ almenys un } j, \text{tal que } \pi_j \neq \pi_{j0} \end{cases}$$

En aquest cas es compleix la condició d'aplicabilitat del test Khi-quadrat, ja que totes les freqüències esperades són majors que 5. Per calcular el p-valor escrivim en la finestra d'instruccions:

```
> chisq.test(c(15,26,14,6,8,12), p=c(3/16,6/16,3/16,1/16,2/16,1/16))

        Chi-squared test for given probabilities

data:  c(15, 26, 14, 6, 8, 12)
X-squared = 10.852, df = 5, p-value = 0.0544
```

El p-valor = 0,0544 > 0,05, per tant no rebutgem H_0. Hi ha una prova significativa del fet que les dades són consistents amb el model mendelià.

Exemple (excés d'ingesta)

Recomanem prendre un medicament experimental a un grup de 50 malalts amb problemes per excés d'ingesta. Després d'un cert temps es classifica la condició de cada malalt en una de les quatre categories següents: sense resposta, resposta moderada, resposta marcada i remissió. Els resultats són:

Sense resposta	Resposta moderada	Resposta marcada	Remissió
15	7	4	24

El tractament es considera efectiu en aquesta primera fase d'investigació si les proporcions de les categories són compatibles amb les proporcions 1:2:2:5.

A la vista dels resultats obtinguts, podem concloure que el tractament és efectiu?

En cas contrari, quins valors esperaríem observar en cada categoria si és certa la hipòtesi nul·la?

Volem contrastar una hipòtesi nul·la que especifique les probabilitats poblacionals d'aquestes categories:

$$\begin{cases} H_0: \pi_1 = 0{,}1, \pi_2 = 0{,}2, \pi_3 = 0{,}2, \pi_4 = 0{,}5 \\ H_A: hi\ ha \text{ almenys un } j, \text{tal que } \pi_j \neq \pi_{j0} \end{cases}$$

En aquest cas es compleix la condició d'aplicabilitat del test Khi-quadrat, ja que totes les freqüències esperades són majors o iguals que 5.

```
> chisq.test(c(15,7,4,24), p=c(0.1,0.2,0.2,0.5))$expected
[1]  5 10 10 25
```

Calculem el p-valor

```
> chisq.test(c(15,7,4,24), p=c(0.1,0.2,0.2,0.5))

        Chi-squared test for given probabilities

data:  c(15, 7, 4, 24)
X-squared = 24.54, df = 3, p-value = 0.00001927
```

El p-valor és aproximadament zero < 0,05, per tant rebutgem H_0. Hi ha una prova suficient del fet que les proporcions de les categories NO són compatibles amb les ràtios 1:2:2:5.

Els valors esperats en cada categoria si H_0 fóra certa són E_1=50*0,1 = 5, E_2 = 50*0,2 = 10, E_3 = 50*0,2 = 10, E_4 = 50*0,5 = 25. Hem comprovat que hi ha una diferència significativa entre els valors observats i els que esperaríem observar si H_0 fóra certa.

3. ANÀLISI DE TAULES DE CONTINGÈNCIA

Fins ara hem considerat l'anàlisi d'una única mostra de dades categòriques. En aquest apartat ampliarem l'estudi sobre dades categòriques a diverses poblacions. Ens trobem en dos situacions

- Situació 1: volem saber si una determinada característica (variable categòrica) té la mateixa distribució en diverses poblacions (**test d'homogeneïtat**).

$$\begin{cases} H_0: \text{té la mateixa distribució} \\ H_A: \text{no té la mateixa distribució} \end{cases}$$

- Situació 2: volem saber si dues característiques diferents (dues variables categòriques) estan relacionades entre si o són independents (t**est d'independència**).

$$\begin{cases} H_0: \text{són independentes} \\ H_A: \text{estan relacionades} \end{cases}$$

En ambdós casos i de manera anàloga s'utilitza una taula de contingència per a representar les dades i donar resposta al contrast plantejat mitjançant un test basat en la distribució Khi-quadrat.

	Opció 1 columna	Opció 2 columna	....
Opció 1 fila	Freqüències observades		
Opció 2 fila			
....			

En general, la freqüència esperada de la cel·la de la fila i, columna j, és:

$$E_{ij} = E(fila\ i, columna\ j) = \frac{(total\ fila\ i) \times (total\ columna\ j)}{total\ general}$$

L'estadístic del test es calcula a partir dels valors de les freqüències observades i les esperades en totes les cel·les de la taula:

$$\chi_s^2 = \sum_{j=1}^{k} \sum_{i=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

on O_{ij} és la freqüència observada en la combinació de categories ij (cel·la (i, j) de la taula de contingència), E_{ij} és la freqüència esperada de la categoria ij. La suma s'estén a totes les cel·les de la taula. Si la hipòtesi nul·la H_0 és certa, l'estadístic segueix una distribució Khi-quadrat amb graus de llibertat (r - 1) x (k-1), on r és el nombre de files i k el nombre de columnes. Un valor gran de l'estadístic indica que les dades són incompatibles amb H_0.

Per poder aplicar els test Khi-quadrat, almenys el 80% de les freqüències esperades han de ser majors o iguals a 5.

R-Commander:

Estadístics → taules de contingència → introduir i analitzar una taula de doble entrada

ODDS RATIO

En taules de contingència 2 x 2 podem utilitzar el concepte **ODDS ratio (OR)** que indica la direccionalitat de l'associació.

ODDS: és la raó entre l'ocurrència d'un succés i la no ocurrència del succés.

ODDS ratio (OR): és el quocient entre la ODDS d'un succés en un grup d'individus exposats i la mateixa ODDS sobre el grup de no exposats.

- Si OR = 1, no hi ha associació entre el factor i el succés.
- Si OR > 1, l'associació és positiva (factor de risc), el factor s'associa amb la major ocurrència del succés.
- Si OR < 1, es considera que l'associació és negativa (factor protector), el factor s'associa amb una menor ocurrència del succés.

	Exposat	No exposat
Casos	a	b
No casos	c	d

$$OR = \frac{a/c}{b/d}$$

El **test exacte de Fisher** és una alternativa al test de la Khi-quadrat que ens permet analitzar taules de contingència 2 x 2 directament. El test calcula l'ODDS ratio i resol el contrast d'hipòtesis següent:

$$\begin{cases} H_0: OR = 1 \\ H_A: \mathrm{OR} \neq 1 \end{cases}$$

Exemple (timolol)

En un estudi per a avaluar l'efectivitat de la droga timolol per a prevenir els atacs en malalts d'angina de pit es tria a l'atzar un grup de malalts als quals s'administra durant 28 setmanes una dosi de timolol i a la resta dels malalts s'administra un placebo. Es van obtenir els resultats següents:

	Timolol	Placebo	TOTAL
Sense atacs	44	19	63
Amb atacs	116	128	244
TOTAL	160	147	307

Població 1: pacients tractats amb timolol.

Població 2: pacients tractats amb placebo.

Variable categòrica: AMB atacs i SENSE.

Té la mateixa distribució la variable categòrica en les dues poblacions?

$$\begin{cases} H_0\text{: els atacs es donen de la mateixa manera en les dues poblacions} \\ H_A\text{: els atacs no es donen de la mateixa manera en les dues poblacions} \end{cases}$$

Calculem ara una freqüència esperada:

- Calculem la probabilitat de no tenir atacs considerant tota la mostra:

$$P(\text{no tenir atacs}) = \frac{\text{pacients sense atacs}}{\text{pacients en total}} = \frac{63}{307} = 0{,}205.$$

- Aleshores, d'un total de 160 malalts que han pres timolol, és possible esperar que aproximadament un 20,5% d'aquests estiguen lliure d'atacs:

$$E(\text{timolol i sense atacs}) = 160 \times 0{,}205 = \frac{160 \times 63}{307} = 32{,}83$$
$$= \frac{(total\ fila\ 1) \times (total\ columna\ 1)}{total\ general}.$$

D'aquesta manera podem calcular la taula de freqüències esperades:

Esperades	Timolol	Placebo
Sense atacs	32,83	30,17
Amb atacs	127,17	116,83

Aquest càlcul pot fer-se directament amb *R-Commander*:

```
> .Test$expected # Expected Counts
                columns
rows              Timolol   Placebo
  Sense atacs  32.83388  30.16612
  Amb atacs   127.16612 116.83388
```

Comprovem que els casos esperats són majors o iguals a 5 per a totes les categories. Per tant, podem aplicar el test Khi-quadrat.

Exemple (timolol)

```
> .Table  # Counts
             columns
rows          Timolol Placebo
  Sense atacs      44      19
  Amb atacs       116     128

> colPercents(.Table) # Column Percentages
             columns
rows          Timolol Placebo
  Sense atacs    27.5    12.9
  Amb atacs      72.5    87.1
  Total         100.0   100.0
  Count         160.0   147.0

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 9.9782, df = 1, p-value = 0.001584
```

El p-valor = 0,001584 $< \alpha$ = 0,05, així que rebutgem H_0. Hi ha una prova estadística suficent per a concloure que els atacs no es donen per igual en ambdues poblacions.

Podem calcular també les probabilitats d'estar lliure d'atacs per als tractats amb timolol i per als tractats amb placebo.

- Dels malalts que reberen timolol, aproximadament el 28% van estar lliures d'atacs, P(sense atacs | timolol) = 44 / 160 = 0,275.
- Dels malalts que reberen un placebo, aproximadament el 13% van estar lliures d'atacs, P(sense atacs | placebo) = 19 / 147 = 0,129.
- Podem concloure que els malalts tractats amb timolol tenen una probabilitat menor de patir atacs que els tractats amb placebo.

En aquest cas, com estem treballant amb una taula 2 x 2, podem calcular també l'ODDs ratio:

$$OR = \frac{44/116}{19/128} = 2{,}556.$$

Una OR = 2,556 indica que el tractament amb timolol és efectiu. La seua ODDS –sobre no patir atacs– és el doble que la corresponent al grup placebo. El factor –prendre timolol – s'associa a una major ocurrència del succés –sense atacs–. Podem fer el test de Fisher com alternativa al Khi-quadrat. Hem de resoldre el contrast següent:

$$\begin{cases} H_0: OR = 1 \\ H_A: \mathrm{OR} \neq 1 \end{cases}$$

Exemple (timolol)

```
> .Table  # Counts
             columns
rows          Timolol Placebo
  Sense atacs      44      19
  Amb atacs       116     128

> fisher.test(.Table)

        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.001785
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.365721 4.902015
sample estimates:
odds ratio
  2.547687
```

El p - valor = 0,001785 < α, rebutgem H_0.

Podem concloure que prenre timolol influeix en el risc de patir atacs ajudant a reduir l'ocurrència d'aquests. A més, l'ODDS ratio és 2,55. El risc de patir atacs entre els qui no prenen timolol s'estima que és més del doble que entre els que en prenen.

Exemple (daltonisme i sexe)

En un estudi sobre daltonisme un grup d'investigadors examinà un gran nombre d'estudiants noruecs d'ensenyament primari i obtingueren els resultats següents:

	Xiquets	Xiquetes	TOTAL
Daltònics	725	40	765
No daltònics	8.324	9.032	17.356
TOTAL	9.049	9.072	18.121

Variable 1: xiquet, xiqueta.

Variable 2: daltònic, no daltònic.

Està relacionant el daltonisme amb el sexe?

$$\begin{cases} H_0\text{: el daltonisme és independent del sexe} \\ H_A\text{: el daltonisme està relacionat amb el sexe} \end{cases}$$

Fem el test Khi-quadrat.

```
> .Test$expected # Expected Counts
                columns
rows              Xiquets  Xiquetes
  Daltonics     382.0145  382.9855
  No_daltonics 8666.9855 8689.0145
```

```
> .Table  # Counts
              columns
rows           Xiquets Xiquetes
  Daltonics        725       40
  No_daltonics    8324     9032

> colPercents(.Table) # Column Percentages
              columns
rows           Xiquets Xiquetes
  Daltonics          8      0.4
  No_daltonics      92     99.6
  Total            100    100.0
  Count           9049   9072.0

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 642.22, df = 1, p-value < 2.2e-16
```

Comprovem que les freqüències esperades són majors o iguals a 5 per a totes les categories. Podem aplicar el test Khi-quadrat.

El p-valor és aproximadament $0 < \alpha = 0{,}05$, per tant rebutgem H_0. Podem concloure que el daltonisme està relacionat amb el sexe. Com?

$$\begin{cases} H_0: OR = 1 \\ H_A: \text{OR} \neq 1 \end{cases}$$

Exemple (daltonisme i sexe)

```
> fisher.test(.Table)

        Fisher's Exact Test for Count Data

data:  .Table
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 14.27301 27.79658
sample estimates:
odds ratio
  19.66864
```

El p-valor és aproximadament 0 < α, rebutgem H_0. El sexe sí afecta el daltonisme.

L'ODDS ratio és 19,67, el risc de patir daltonisme entre els xiquets s'estima que és quasi vint vegades més gran que entre les xiquetes.

Exemple (tipus sanguini i malaltia)

En la taula següent es presenta la distribució observada entre la severitat d'una malaltia i els tipus sanguini en una població.

	Cap	Moderat	Sever	TOTAL
A	543	44	28	615
B	211	22	9	242
AB	90	8	7	105
O	476	31	31	538
Total	1.320	105	75	1.500

Són les dades compatibles amb la hipòtesi de la independència entre el tipus sanguini i el grau d'afecció de la malaltia?

$$\begin{cases} H_0: \text{el grau d'afecció NO està relacionat amb el grup sanguini} \\ H_A: \text{el grau d'afecció està relacionat amb el grup sanguini} \end{cases}$$

Els resultats del test de la Khi-quadrat amb R-Commander són:

```
> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 5.1163, df = 6, p-value = 0.529


> .Test$expected # Expected Counts
    columns
rows    Cap Moderat Sever
  A  541.20   43.05 30.75
  B  212.96   16.94 12.10
  AB  92.40    7.35  5.25
  0  473.44   37.66 26.90
```

Comprovem que les freqüències esperades són majors o iguals a 5 per a totes les categories. Podem aplicar el test Khi-quadrat.

El p-valor és 0,529 > α = 0,05, per tant no rebutgem H_0. No tenim una prova estadística suficient per a dir si el grau d'afecció de la malaltia està relacionat amb el grup sanguini. El grup sanguini és independent del grau d'afecció de la malaltia.

4. RESUM

- Per a una variable categòrica dicotòmica podem fer una inferència sobre la probabilitat d'èxit (π) mitjançant el test binomial exacte: interval de confiança i contrast d'hipòtesi.
- Per a variables categòriques amb més de tres categories podem fer un contrast d'hipòtesis per contrastar la bondat d'ajust respecte d'una determinada distribució de probabilitat: test Khi-quadrat de bondat d'ajust.
- L'anàlisi de dues variables categòriques (test d'independència) o d'una variable categòrica en diverses poblacions (test d'homogeneïtat) es fa mitjançant el test Khi-quadrat per a taules de contingència.
- En taules de contingència 2 x 2 podem utilitzar el concepte d'ODDS ratio (OR), que indica la direccionalitat de l'associació, mitjançant el test de Fisher.
- Per poder aplicar el test de la Khi-quadrat hem d'assegurar-nos que el 80% de les freqüències esperades siguen majors o iguals que 5.

TEMA 6: REGRESSIÓ LINEAL

En aquest tema estudiarem mètodes per a analitzar la relació entre dues variables quantitatives, X i Y. Les dades són n parells d'observacions $(x_i, y_i), i = 1, 2, \dots, n$, on els x_i són els valors que pren la variable X, explicativa, i els y_i representen els valors de la variable Y, explicada, per a cadascun dels n individus de la mostra.

La recta de regressió lineal i l'anàlisi de correlació són tècniques basades en l'ajust d'una recta a les dades.

1. DESCRIPCIÓ DE LA RELACIÓ LINEAL ENTRE DUES VARIABLES NUMÈRIQUES

El diagrama de dispersió és una representació gràfica bidimensional de les observacions. Permet confirmar visualment l'existència d'una relació lineal entre les variables X i Y. Nosaltres ens ocuparem de la relació més senzilla, la **relació lineal**. Per tant, estudiarem com s'assembla el núvol de punts a una recta.

R-Commander:

Gràfiques → diagrames de dispersió

Recordem que tenim una mostra de n parelles en què cada parella representa les mesures de dues variables, X i Y. Si el diagrama de dispersió mostra una tendència lineal, mesurarem la força de l'associació lineal utilitzant la covariància i el coeficient de correlació. Com hem vist en el tema 1, les variàncies mostrals de les variables X i Y són:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}SS_x,$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}SS_y.$$

De la mateixa manera, la **covariància** es defineix com:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}SS_{xy}.$$

El signe de la covariància indica el sentit de la relació lineal:

- Si és positiu, la relació és directa.
- Si és negatiu, la relació és inversa.

Si la covariància és aproximadament 0, no hi ha relació lineal. El "problema" de la covariància és que depèn de les unitats de X i Y, i, per tant, no permet quantificar el grau d'associació entre les variables.

El **coeficient de correlació de Pearson** es defineix com:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Aquest valor no depèn de les unitats de les variables, $r \in [-1, 1]$, per la qual cosa permet quantificar el grau d'associació entre les variables X i Y,

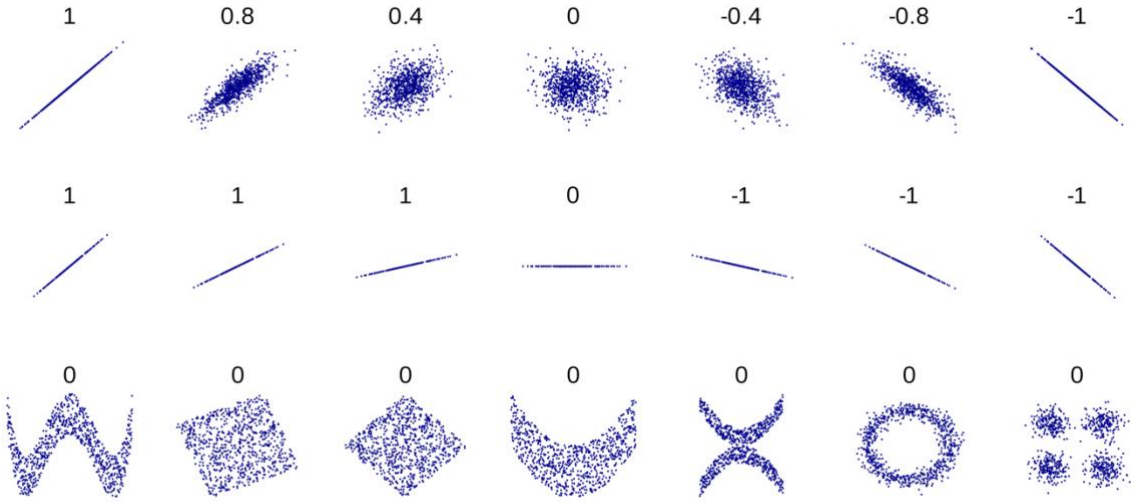
- Si r > 0, la relació és directa i major com major és r. En particular, si r = 1, la relació és directa perfecta.
- Si r < 0, la relació és inversa i major com menor és r. En particular, si r = -1, la relació és inversa perfecta.
- Si r = 0, no hi ha relació lineal.

Cal destacar que hi ha un contrast d'hipòtesis: la hipòtesi nul·la amb l'absència de relació lineal en la població, és a dir:

$$\begin{cases} H_0: \ \rho = 0 \\ H_A: \ \rho \neq 0 \end{cases}$$

on ρ és el coeficient de correlació lineal <u>en la població.</u>

Una correlació zero no significa necessàriament que no hi haja relació entre X i Y, significa que **no hi ha relació lineal** entre X i Y:



Per calcular els coeficients de correlació lineal fem amb R-Commander el següent:

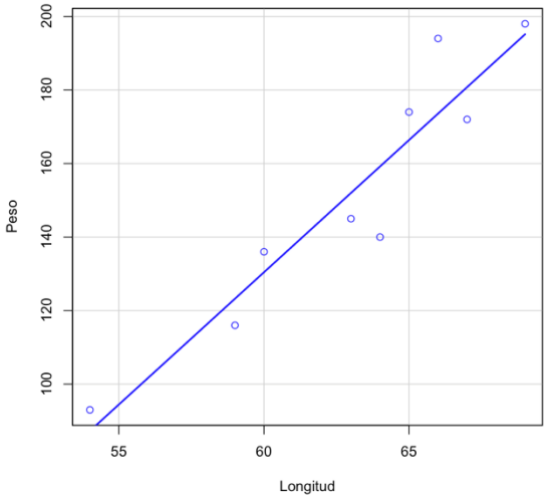
R-Commander:

Estadístics → resums → matriu de correlacions

- Cal seleccionar totes les variables que es desitja estudiar.
- S'ha de seleccionar l'opció "p-valors aparionats" (pairwise).

2. RECTA DE REGRESSIÓ

L'objectiu és explicar la variable Y, variable explicada o dependent, a partir de la informació de la variable X, variable explicativa o independent. La relació més senzilla entre ambdues variables és la lineal $Y = b_0 + b_1 X$. Aquesta recta és la més pròxima al núvol de punts en el diagrama de dispersió i s'anomena **recta de regressió**.



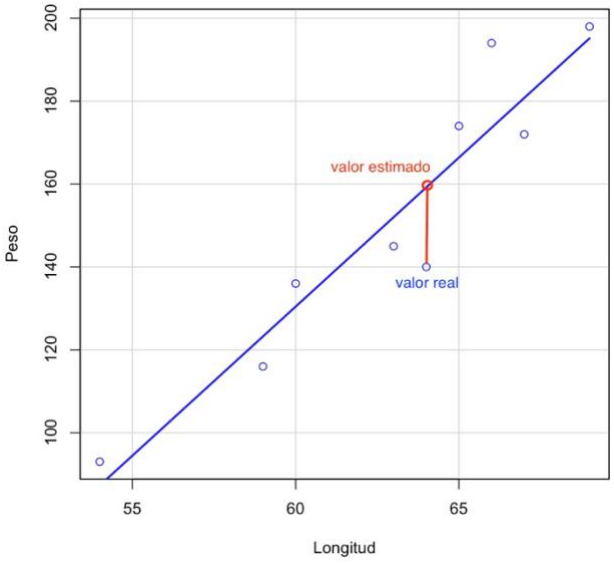
El mètode per determinar els coeficients de la recta b_0 i b_1 és el de **mínims quadrats.**

Definim la **distància** d'un punt (x_i, y_i), el valor real, a una recta $y = b_0 + b_1 x$ com la **distància vertical**: $|y_i - \hat{y}_i|$ amb $\hat{y}_i = b_0 + b_1 x_i$, el valor estimat.

La **distància de mínims quadrats** es defineix com:

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

on $\sum_{i=1}^{n} e_i^2$ es denomina **suma de quadrats residual**.



La recta de mínims quadrats és la que minimitza la suma de quadrats residual. És a dir, volem minimitzar l'error comès quan utilitzem el valor estimat mitjançant la recta, en compte del valor observat. D'aquesta manera, el mínim s'obté per a:

$$b_0 = \bar{y} - b_1 \bar{x} \qquad\qquad b_1 = \frac{SS_{xy}}{SS_x}.$$

Interpretació dels coeficients de la recta:

- b_0 és la constant, també coneguda com intercepte. Estimació de y per a x = 0 (de vegades no té sentit).

- b_1, el pendent. Increment en Y per a cada unitat d'increment en X. Serà positiu si la relació és creixent i negatiu si és decreixent.
- Observem que el pendent té el mateix signe que la covariància. Per tant, també que el coeficient de correlació.
- Si no hi ha relació, el pendent val zero.

R-Commander:

Estadístics → ajust de models → regressió lineal

3. COEFICIENT DE DETERMINACIÓ

La recta de regressió serà útil per a predir si conèixer X disminueix la nostra incertesa sobre Y.

El **coeficient de determinació (R^2)** és una mesura de la disminució de la incertesa. És la proporció de variabilitat (variància) de Y explicada per la regressió. Es defineix com:

$$R^2 = 1 - \frac{SS_e(variància\ dels\ residus)}{SS_y}, \quad 0 \leq R^2 \leq 1.$$

- Si $R^2 = 0$, el model no explica res de Y a partir de X. Un valor pròxim a 0 indica poca capacitat explicativa del model.
- Si $R^2 = 1$, el model proporciona un ajust perfecte. Un valor pròxim a 1 indica una gran capacitat explicativa del model.
- En regressió lineal es pot comprovar que $R^2 = r^2$.

4. INFERÈNCIES ESTADÍSTIQUE EN EL MODEL DE REGRESSIÓ LINEAL

El valor dels coeficients de la recta de regressió depèn de les dades (x_i, y_i) observades en l'experiment, és a dir, cada mostra ens donarà una equació de la recta distinta. Però, el model (relació "biològica" que volem caracteritzar) és invariable i universal, encara que siga un model desconegut.

Una relació lineal universal per a les variables aleatòries X i Y es formularia mitjançant el **model de regressió lineal simple normal homocedàstic**:

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- Els coeficients β_0 i β_1 són paràmetres desconeguts que cal estimar.
- El terme ϵ s'anomena error aleatori i correspon a les petites desviacions respecte de la fórmula que es donen en l'observació de les variables.
- Per poder fer inferències sobre els paràmetres del model (β_0 i β_1) mitjançant una mostra necessitem que es complisquen certes condicions: linealitat, independència, normalitat i homocedasticitat.

La formulació anterior del model de regressió lineal implica disposar dels estadístics necessaris per a calcular intervals de confiança i contrastos d'hipòtesis (inferència)

sobre els paràmetres (β_0 i β_1), que són útils per a obtenir informació sobre el model de regressió.

Paràmetre	Estimador	DT estimada	Estadístic del contrast
$\boldsymbol{\beta_0}$	$b_0 = \bar{y} - b_1\bar{x}$	$s_{\beta_0} = \frac{s_e}{\sqrt{n}}\sqrt{1 + \left(\frac{\bar{x}}{s_x}\right)^2}$	$\frac{(\beta_0 - \hat{\beta}_0)}{s_{\beta_0}} \sim t_{n-2}$
$\boldsymbol{\beta_1}$	$b_1 = \frac{SS_{xy}}{SS_x}$	$s_{\beta_0} = \frac{s_e}{s_x\sqrt{n}}$	$\frac{(\beta_1 - \hat{\beta}_1)}{s_{\beta_1}} \sim t_{n-2}$

Intervals de confiança

En molts estudis el valor β_1 és un paràmetre amb sentit biològic i l'objectiu principal de l'anàlisi de dades és estimar-lo. Es pot construir un interval de confiança per a β_1 pel mètode habitual basat en la desviació típica estimada i la distribució t de Student. L'interval de confiança al $(1-\alpha)$ 100% pel pendent es defineix com:

$$IC(\beta_1) = b_1 \pm t_{\alpha/2}\, s_{\beta_1,}$$

on $t_{\alpha/2}$ es determina a partir de la distribució t de Student amb n-2 graus de llibertat. L'interval de confiança per a β_0 es construeix de forma anàloga:

$$IC(\beta_0) = b_0 \pm t_{\alpha/2}\, s_{\beta_0,}$$

Per obtenir els intervals de confiança, després d'ajustar el model lineal elegim en el menú de *R-Commander*, l'opció "Intervals de confiança".

R-Commander:

Models → intervals de confiança

Contrastos d'hipòtesis

Contrast sobre la constant de la recta	Contrast de linealitat (sobre el pendent)
$\begin{cases}\boldsymbol{H_0}: \ \boldsymbol{\beta_0 = 0} \\ \boldsymbol{H_A}: \ \boldsymbol{\beta_0 \neq 0}\end{cases}$	$\begin{cases}\boldsymbol{H_0}: \ \boldsymbol{\beta_1 = 0} \\ \boldsymbol{H_A}: \ \boldsymbol{\beta_1 \neq 0}\end{cases}$

El contrast sobre el pendent es diu contrast de linealitat. És molt rellevant, ja que l'acceptació de la hipòtesi nul·la (és a dir, la possibilitat que el pendent siga 0) indica la no utilitat del model. En canvi la inferència sobre la constant no sol utilitzar-se.

L'estimador del pendent pot expressar-se en termes del coeficient de correlació mostral:

$$b_1 = \frac{SS_{xy}}{SS_x} = r\frac{s_y}{s_x},$$

que estima el coeficient de correlació poblacional ρ (mesura de la relació lineal entre X i Y poblacional, és a dir, que no depén de la mostra concreta que s'haja observat). Per tant:

$$\begin{cases}H_0: \ \beta_1 = 0 \\ H_A: \ \beta_1 \neq 0\end{cases} \quad \text{és equivalent a} \quad \begin{cases}H_0: \ \rho = 0 \\ H_A: \ \rho \neq 0\end{cases}$$

Per tant, el contrast de linealitat és equivalent al contrast sobre l'existència de relació lineal entre les variables X i Y. No rebutjar la hipòtesi nul·la no significa necessàriament que no hi haja relació entre les dues variables, només que si n'hi ha, la relació no és lineal.

5. PREDICCIÓ I MODEL LINEAL

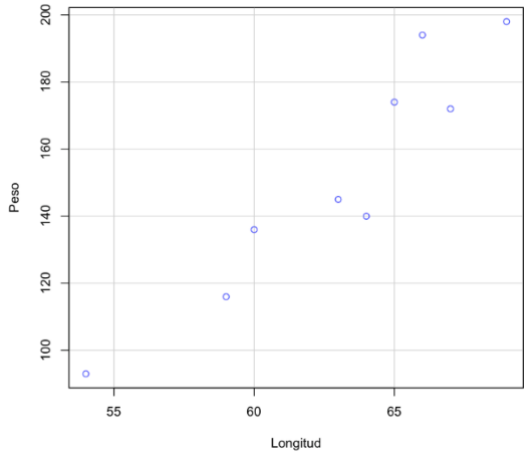
Considerem el cas de recórrer a l'altura, X, per predir el pes, Y, en un grup gran d'homes joves el pes mitjà dels quals és 80 kg. Suposem que es tria aleatòriament un home jove i hem de predir-ne el pes.

- Si no sabem res sobre l'altura de l'home, llavors el millor estimador que podem donar del seu pes és el pes mitjà global, 80 kg.
- Suposem que sabem que l'altura de l'home és de 175 cm. Si sabem que el pes mitjà de tots els homes del grup amb altura de 175 cm és de 78 kg, llavors podem utilitzar la mitjana, 78 kg, com a predicció del pes de l'home. Esperem que aquesta predicció siga més exacta que la predicció donada en l'apartat anterior.
- Suposem que sabem que l'altura de l'home és de 175 cm i que també sabem que el model de regressió és Y = -189 + 1,56 X, amb R^2 = 0,87. Llavors podem utilitzar el valor x = 175 per a fer una predicció: -189 + 1,56 x 175 = 84 kg.

Exemple (serps)

En un estudi de la població salvatge de la serp *Vipera bertis*, uns investigadors van capturar i mesurar nou femelles adultes. Les longituds dels seus cossos i els seus pesos es mostren en la taula següent.

Longitud X (cm.)	Pes Y (g.)
60	136
69	198
66	194
64	140
54	93
67	172
59	116
65	174
63	145



El diagrama de dispersió indica una clara tendència ascendent. Diem que el pes mostra una associació **positiva** amb la longitud, la qual cosa indica que les longituds majors estan associades amb els pesos majors.

```
> rcorr.adjust(serpientes[,c("Longitud","Peso")],
 type="pearson", use="complete")

 Pearson correlations:
         Longitud   Peso
Longitud   1.0000 0.9437
Peso       0.9437 1.0000

 Number of observations: 9

 Pairwise two-sided p-values:
         Longitud Peso
Longitud          0.0001
Peso     0.0001
```

Hi ha una relació lineal directa entre la longitud i el pes de les serps, r = 0,9437.

Aquesta relació és molt significativa:

p-valor = 0,0001

Exemple (serps)

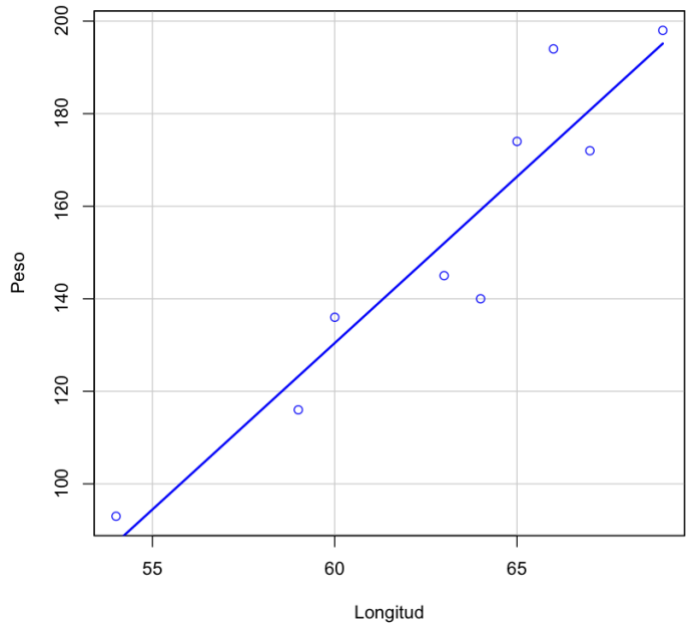
```
> RegModel.1 <- lm(Peso~Longitud, data=serpientes)

> summary(RegModel.1)

Call:
lm(formula = Peso ~ Longitud, data = serpientes)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -301.0872    60.1885  -5.002 0.001561 **
Longitud       7.1919     0.9531   7.546 0.000132 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.5 on 7 degrees of freedom
Multiple R-squared:  0.8905, Adjusted R-squared:  0.8749
F-statistic: 56.94 on 1 and 7 DF,  p-value: 0.0001321
```

La recta de regressió és: Pes = -301,09 + 7,19 x Longitud.

Interpretació b_1. Per cada cm més de longitud d'una serp s'estima un increment de 7,19 grams en el seu pes.

Interpretació b_0. A una serp amb longitud 0 cm correspondria un pes de -301,09.

Aquest és un d'aqueixos casos en els quals la interpretació de la constant no té sentit.



Exemple (serps)

```
> RegModel.1 <- lm(Peso~Longitud, data=serpientes)

> summary(RegModel.1)

Call:
lm(formula = Peso ~ Longitud, data = serpientes)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -301.0872    60.1885  -5.002 0.001561 **
Longitud       7.1919     0.9531   7.546 0.000132 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.5 on 7 degrees of freedom
Multiple R-squared:  0.8905,  Adjusted R-squared:  0.8749
F-statistic: 56.94 on 1 and 7 DF,  p-value: 0.0001321
```

El percentatge de variància de Y explicat per X és molt alt, 89,05% (quantitat en unitats relatives, percentatge). La longitud explica un 89,05% del pes d'una serp.

A continuació definim els contrastos d'hipòtesis per als paràmetres β_0 i β_1:

Contrast sobre la constant de la recta	**Contrast de linealitat (sobre el pendent)**
$\begin{cases} \boldsymbol{H_0}: \ \boldsymbol{\beta_0 = 0} \\ \boldsymbol{H_A}: \ \boldsymbol{\beta_0 \neq 0} \end{cases}$	$\begin{cases} \boldsymbol{H_0}: \ \boldsymbol{\beta_1 = 0} \\ \boldsymbol{H_A}: \ \boldsymbol{\beta_1 \neq 0} \end{cases}$
Com el p-valor = 0,001561 < 0,05, rebutgem $\boldsymbol{H_0}$ i l'ordenada en l'origen és significativa en el model	Com el p-valor = 0,000132 < 0,05, rebutgem H_0 i el pendent és significatiu en el model

Com podem acceptar que $\boldsymbol{\beta_1 \neq 0}$ en un nivell de confiança elevat, la relació lineal entre longitud i pes és significativa.

```
> RegModel.1 <- lm(Peso~Longitud, data=serpientes)

> summary(RegModel.1)

Call:
lm(formula = Peso ~ Longitud, data = serpientes)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -301.0872    60.1885  -5.002 0.001561 **
Longitud       7.1919     0.9531   7.546 0.000132 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.5 on 7 degrees of freedom
Multiple R-squared:  0.8905,  Adjusted R-squared:  0.8749
F-statistic: 56.94 on 1 and 7 DF,  p-value: 0.0001321
```

Exemple (serps)

Calculem els intervals de confiança:

```
> Confint(RegModel.1, level=0.95)
                Estimate       2.5 %      97.5 %
(Intercept) -301.08721 -443.410309 -158.764110
Longitud       7.19186    4.938183    9.445538
```

L'interval de confiança pel pendent és (4,94, 9,45).

Per a la recta de regressió Pes = -301,09 + 7,19 x longitud, quin seria el pes d'una serp amb una longitud de 63 cm?

Pes = -301,09 + 7,19 x 63 = 151,88 grams.

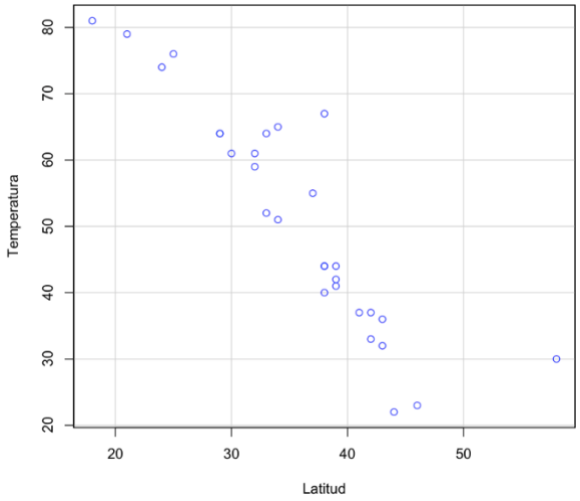
En el cas d'una serp de 80 cm podríem estimar-ne el pes?

En el gràfic de dispersió podem veure que el rang de les dades observades no inclou el valor 80 cm. Per tant, no es pot fer l'estimació.

Exemple (temperatura)

La latitud és la distància angular entre la línia equatorial (l'equador) i un punt determinat de la Terra. Mesura al llarg del meridià en el qual es troba aquest punt. Volem estudiar la possible relació entre latitud i temperatura. Disposem de dades sobre la temperatura màxima en gener (en graus Celsius) i la latitud (en graus sexagesimals) de diverses ciutats dels EUA..

Ciutat	Latitud X	Temperatura Y
Mobile, Ala	30	16
Montgomery, Ala	32	15
Juneau, Alaska	58	-1
Phoenix, Ariz	33	18
Litle Rock, Ark	34	11
Los Angeles, Cal	34	18
San Francisco, Cal	37	13
Denver, Col	39	6
...	...	...



El diagrama de dispersió indica una tendència decreixent. Diem que hi ha una associació **negativa** entre les dues variables, és a dir, que quan una augmenta l'altra disminueix. La gràfica indica que a més latitud la temperatura màxima en gener és més baixa.

```
> rcorr.adjust(temperatura[,c("Latitud","Temperatura")],
 type="pearson", use="complete")

 Pearson correlations:
             Latitud Temperatura
Latitud       1.0000     -0.8955
Temperatura -0.8955      1.0000

 Number of observations: 29

 Pairwise two-sided p-values:
             Latitud Temperatura
Latitud              <.0001
Temperatura <.0001
```

Hi ha relació lineal inversa entre la latitud i la temperatura màxima en gener, r = - 0,8955.

Aquesta relació és molt significativa :

p-valor < 0,0001

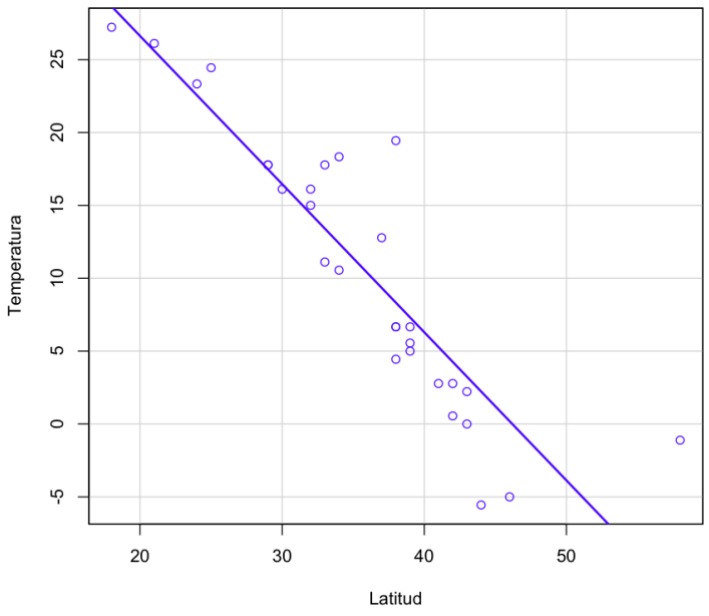
Exemple (temperatura)

```
> RegModel.1 <- lm(Temperatura~Latitud, data=temperatura)

> summary(RegModel.1)

Call:
lm(formula = Temperatura ~ Latitud, data = temperatura)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7868 -2.2988 -1.0253  0.7772 11.1157

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.94596    3.56963   13.15 2.97e-13 ***
Latitud     -1.01624    0.09719  -10.46 5.41e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.228 on 27 degrees of freedom
Multiple R-squared:  0.8019, Adjusted R-squared:  0.7946
F-statistic: 109.3 on 1 and 27 DF,  p-value: 5.41e-11
```

La recta de regressió és: Temperatura 46,95 – 1,02 x latitud.

Interpretació b_1. Per cada grau més de latitud de la localització d'una ciutat s'estima una disminució de la temperatura de 1,02 graus Celsius.

Interpretació b_0. La temperatura màxima en gener per a una ciutat localitzada a l'Equador seria aproximadament de 46,95ºC.



Exemple (temperatura)

```
> RegModel.1 <- lm(Temperatura~Latitud, data=temperatura)

> summary(RegModel.1)

Call:
lm(formula = Temperatura ~ Latitud, data = temperatura)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7868 -2.2988 -1.0253  0.7772 11.1157

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.94596    3.56963   13.15 2.97e-13 ***
Latitud     -1.01624    0.09719  -10.46 5.41e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.228 on 27 degrees of freedom
Multiple R-squared:  0.8019,  Adjusted R-squared:  0.7946
F-statistic: 109.3 on 1 and 27 DF,  p-value: 5.41e-11
```

El percentatge de variància de Y explicat per X és alt, 80,19%. La latitud explica un 80,19% de la temperatura màxima en gener.

A continuació definim els contrastos d'hipòtesis per als paràmetres β_0 i β_1:

Contrast sobre la constant de la recta	**Contrast de linealitat (sobre el pendent)**
$\begin{cases} \boldsymbol{H_0}: \ \boldsymbol{\beta_0 = 0} \\ \boldsymbol{H_A}: \ \boldsymbol{\beta_0 \neq 0} \end{cases}$	$\begin{cases} \boldsymbol{H_0}: \ \boldsymbol{\beta_1 = 0} \\ \boldsymbol{H_A}: \ \boldsymbol{\beta_1 \neq 0} \end{cases}$
Com el p-valor = 2,97e-13 < 0,05, rebutgem $\boldsymbol{H_0}$ i l'ordenada en l'origen és significativa en el model	Com el p-valor = 5,41e-11< 0,05, rebutgem H_0 i el pendent és significatiu en el model

La recta de regressió és: Temperatura = 46,95 – 1,02 latitud.
Per cada grau més de latitud de la localització d'una ciutat s'estima una disminució de la temperatura de 1,02 graus Celsius. Aquesta disminució és **significativa** (p-valor = 5,41e-11, aleshores $\boldsymbol{\beta_1 \neq 0}$.

```
> RegModel.1 <- lm(Temperatura~Latitud, data=temperatura)

> summary(RegModel.1)

Call:
lm(formula = Temperatura ~ Latitud, data = temperatura)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7868 -2.2988 -1.0253  0.7772 11.1157

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.94596    3.56963   13.15 2.97e-13 ***
Latitud     -1.01624    0.09719  -10.46 5.41e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.228 on 27 degrees of freedom
Multiple R-squared:  0.8019,  Adjusted R-squared:  0.7946
F-statistic: 109.3 on 1 and 27 DF,  p-value: 5.41e-11
```

Exemple (temperatura)

Calculem els intervals de confiança:

```
> Confint(RegModel.1, level=0.95)
             Estimate    2.5 %     97.5 %
(Intercept) 46.945963 39.62169 54.2702411
Latitud     -1.016244 -1.21567 -0.8168182
```

L'interval de confiança en el 95% pel pendent (latitud) és (-1,22, -0,82).

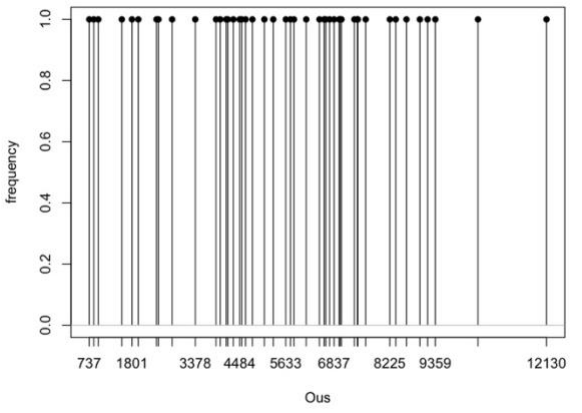
PROBLEMES TEMA 1

EXERCICI 1

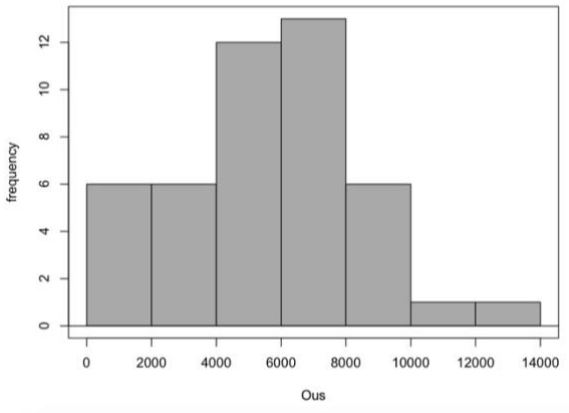
Una de les variables d'interès en l'estudi del cranc *Xantido* (petit cranc que habita prop de Gloucester Point, Virgínia) és el nombre d'ous posts per individu. Les observacions següents corresponen al nombre d'ous obtinguts per a 45 crancs.

1959	4534	7020	6725	6964	7428	9359	9166	2802	2462
4000	3378	7343	4189	8973	4327	2412	7624	1548	4801
737	5321	849	5749	6837	8639	7417	6982	10421	962
3894	1801	5099	6627	4484	5633	4148	6588	5837	4632
6472	8372	8225	6142	12130					

a) Descriu la població, la mostra estudiada (indicant la grandària de la mostra) i classifica la variable observada.
b) Quina gràfica et sembla més apropiada per a aquestes dades, la gràfica 1 o la gràfica 2. Per què? Quina informació podries extraure de cadascuna?
c) Quins estadístics s'han calculat en la taula 1? Calcula tots els estadístics de la variable **Ous** (que representa el nombre d'ous) que pugues obtenir fàcilment a partir dels anteriors. Interpreta el significat dels percentils calculats.



Gràfica 1



Gràfica 2

```
> numSummary(Problema1[,"Ous", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
     mean       sd  IQR  0%  25%  50%  75%  100%  n
 5578.044 2669.322 3343 737 4000 5749 7343 12130 45
```

Taula 1

EXERCICI 2

En un estudi morfomètric dels coloms domèstics, es va mesurar, entre altres, l'ample intraorbital. Per a una mostra de 40 coloms domèstics, s'han obtingut les dades següents:

12,2	12,9	11,8	11,9	11,6	11,1	12,3	12,2	11,8	11,8
10,7	11,5	11,3	11,2	11,6	11,9	13,3	11,2	10,5	11,1
12,1	11,9	10,4	10,7	10,8	11	11,9	10,2	10,9	11,6
10,8	11,6	10,4	10,7	12	12,4	11,7	11,8	11,3	11

a) Descriu la població, la mostra estudiada (indicant la grandària de la mostra) i classifica la variable observada.
b) Amb la informació estadística disponible en la taula 2, dibuixa el diagrama de caixa corresponent.

```
> numSummary(Problema2[,"Amplitud", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.95,1))
    mean        sd   IQR   0%    25%  50%    95% 100%  n
 11.4775 0.6933706 0.925 10.2 10.975 11.6 12.425 13.3 40
```

Taula 2

EXERCICI 3

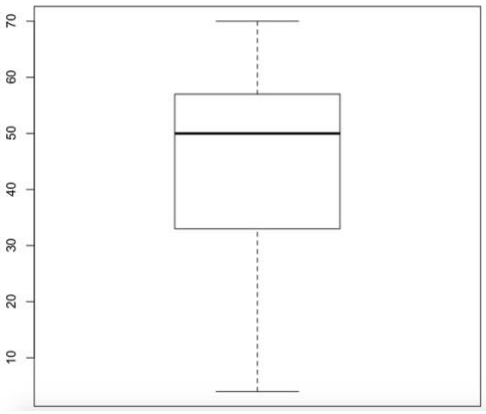
En un estudi sobre el comportament de la mosca del vinagre *Drosophila melanogaster*, un biòleg va mesurar el temps en segons que una mosca passava netejant-se en un període de 6 minuts. Els temps de neteja observats per a 20 mosques diferents van ser:

34	24	10	16	52	76	33	31	46	24
18	26	57	32	25	48	22	48	29	19

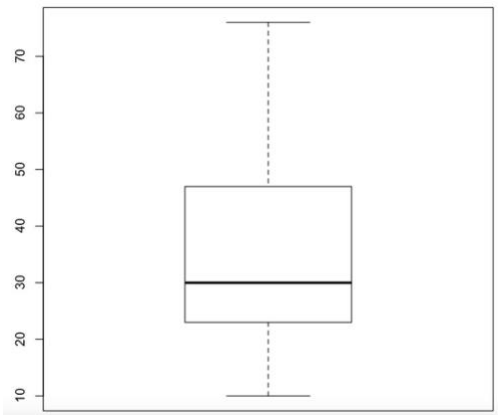
a) Raona quin diagrama de caixa seria el correcte, dels representats en les gràfiques 3, 4 i 5, tenint en compte la informació proporcionada en la taula 3.
b) És possible que un mosca tarde entre 60 i 70 segons a netejar-se?
c) A partir dels estadístics de la taula 3, quant val la variància, la mediana, el valor màxim i el valor mínim? Explica el significat de cada quantitat.

```
> numSummary(Problema3[,"Temps", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
 mean       sd IQR 0%  25% 50%  75% 100%  n
 33.5 16.31435  23 10 23.5  30 46.5   76 20
```

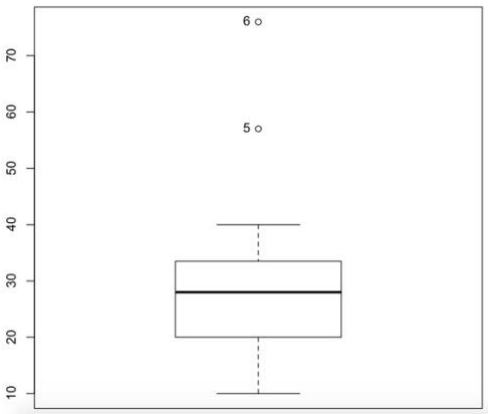
Taula 3



Gràfica 3



Gràfica 4



Gràfica 5

EXERCICI 4

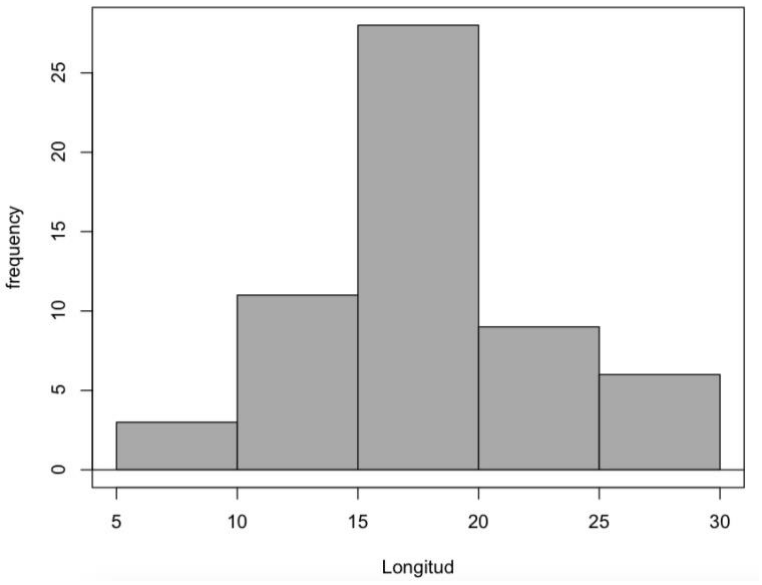
En un experiment per a estudiar l'efecte d'un fertilitzant sobre el creixement dels raves, es compara el creixement en dos grups de raves: el grup de control, format per 23 plantes de rave que no han estat tractades amb fertilitzant, i el grup experimental, format per 34 plantes tractades amb aquest fertilitzant. Les dades següents corresponen a la longitud, en mm, d'un cotilèdon de cada una de les plantes considerades.

Sense fertilitzant:	18	19	12	16	14	17	13	16
	15	14	20	17	10	8	17	28
	19	12	16	16	19	17	12	
Amb fertilitzant:	29	17	21	21	30	19	17	17
	19	11	18	20	19	19	27	14
	25	19	20	26	9	11	24	24
	27	15	24	25	16	19	16	23
	18	23						

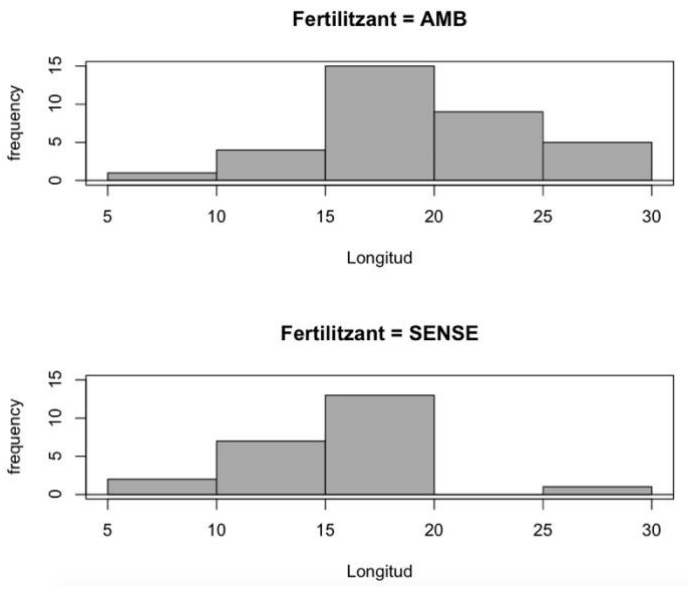
Comenta el significat de la taula 4 i de cada gràfica.

```
> numSummary(Problema4[,"Longitud", drop=FALSE], groups=Problema4$Fertilitzant, statistics=c("mean", "sd", "IQR", "quantiles"),
+     quantiles=c(0,.25,.5,.75,1))
          mean       sd IQR 0%  25% 50%  75% 100% Longitud:n
AMB   20.05882 5.074841   7  9 17.0  19 24.0   30         34
SENSE 15.86957 4.048618   4  8 13.5  16 17.5   28         23
```

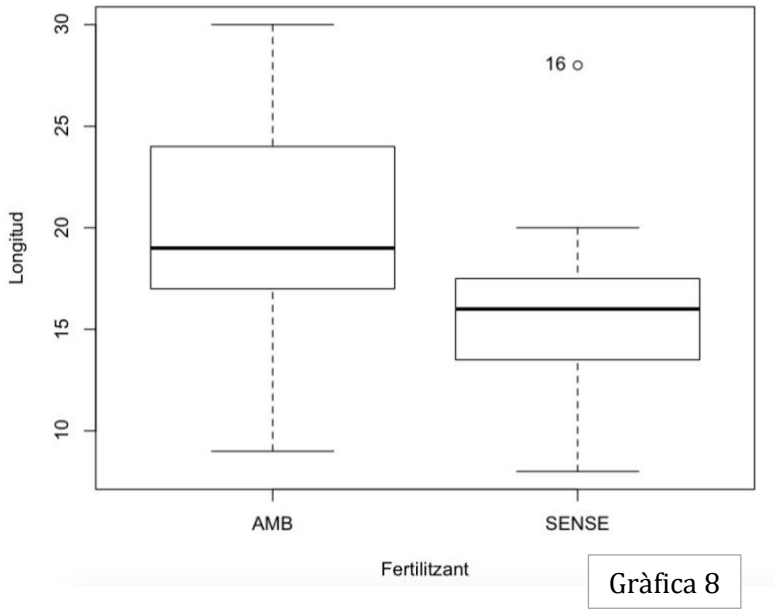
Taula 4



Gràfica 6



Gràfica 7



Gràfica 8

EXERCICI 5

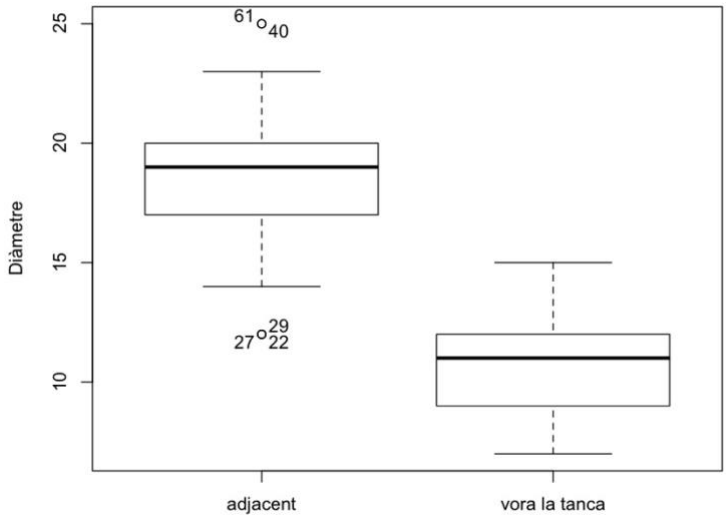
Certes plantes de *Ranunculus acris* que creixen prop d'una tanca apareixen a la vista amb flors clarament més petites que les habituals de l'espècie. Per obtenir una confirmació quantitativa d'aquesta impressió visual, es van mesurar els diàmetres de les flors de les plantes que creixen vora la tanca i d'altres plantes adjacents amb flors normals. Els diàmetres, en mm, van ser aquests:

Vora la tanca:	15	9	13	8	13	9	7	10	11	9
	10	11	11	12	12	11				
Adjacents:	15	19	14	20	17	12	21	17	14	16
	12	19	12	17	19	19	19	19	17	19
	19	19	20	25	18	20	21	22	16	19
	19	23	21	20	19	18	18	15	20	15
	19	20	18	21	25	19				

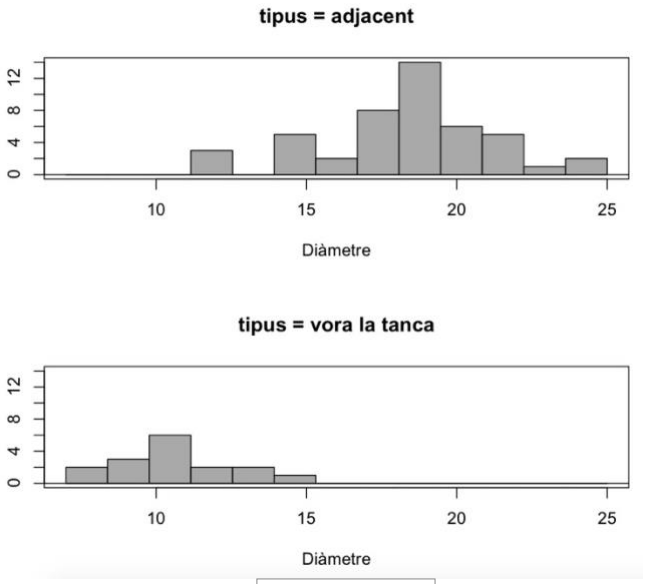
a) Confirmen les representacions gràfiques la impressió visual quant a la diferència de grandària?
b) Què pot deduir-se de les gràfiques sobre la simetria d'ambdues distribucions?
c) A partir dels estadístics descrits en la taula 5, què podem dir sobre la simetria?

```
> numSummary(Problema5[,"Diàmetre", drop=FALSE], groups=Problema5$tipus, statistics=c("mean", "sd", "IQR", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
                mean       sd IQR 0% 25% 50% 75% 100% Diàmetre:n
adjacent      18.3913 2.924823   3 12  17  19  20   25         46
vora la tanca 10.6875 2.056494   3  7   9  11  12   15         16
```

Taula 5



Gràfica 9



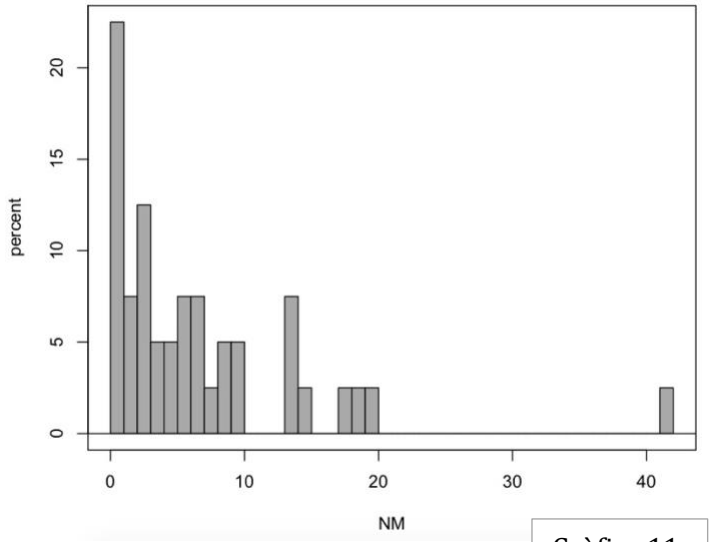
Gràfica 10

EXERCICI 6

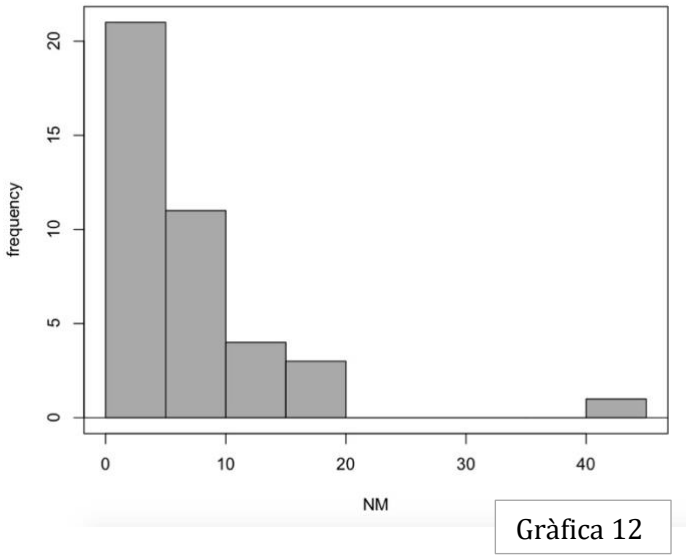
Es va parasitar amb àcars rates blanques infectades amb el cuc filarial *Litomosoides carinii*. Posteriorment, es van dissecar i es va comptabilitzar NM=”Nre. de microfilàries en cada àcar". Els resultats de la dissecció de 40 àcars van ser:

3	3	1	8	0	7	2	0	10	15	3
19	1	2	42	3	4	7	0	9	0	18
4	6	6	10	1	1	9	14	5	7	5
14	20	6	1	2	14	3	9	13	8	13
9	7	10	11	9						

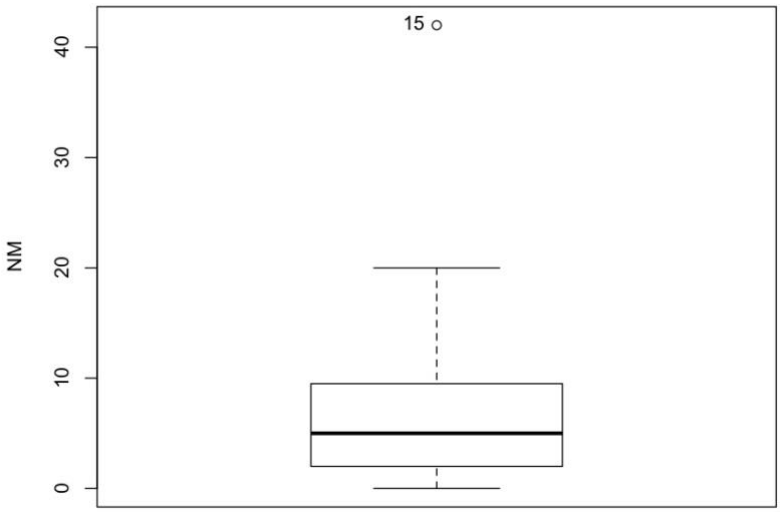
a) Quina diferència hi ha entre la gràfica 11 i la gràfica 12? Quina penses que pot ser més adequada per a descriure aquestes dades?
b) Comenta si les dades podrien provenir d’una distribució normal.
c) Explica el significat del punt que apareix en la gràfica de caixa i del núm. 15 que hi ha al costat.



Gràfica 11



Gràfica 12



Gràfica 13

```
> numSummary(Problema6[,"NM", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles", "skewness", "kurtosis"),
+   quantiles=c(0,.25,.5,.75,1), type="2")
  mean       sd  IQR skewness kurtosis 0% 25% 50%  75% 100%  n
 7.125 7.936244 7.25 2.462343  8.60996  0   2   5 9.25   42 40
```

Taula 6

EXERCICI 7

En un experiment d'*Aanes* es va administrar a 14 conills porquins de laboratori un producte químic que causa somnolència. El temps transcorregut, en minuts, entre la ingestió d'aquest producte i l'entrada en la fase de son va ser:

42	27	24	24	36	36	44
44	120	29	36	36	36	36

Per tal de valorar com afecta als estadístics de la mostra i a les gràfiques la presència d'un possible valor extrem, hem preparat dues mostres: la primera amb totes les dades observades i la segona amb una dada menys. Així, la taula 7 i les gràfiques 14 i 16 corresponen a les dades originals, mentre que la taula 8 i les gràfiques 15 i 17 corresponen a la mostra obtinguda després d'eliminar la dada amb el major valor observat de la mostra original, el valor extrem (ara la mostra només té 13 observacions).

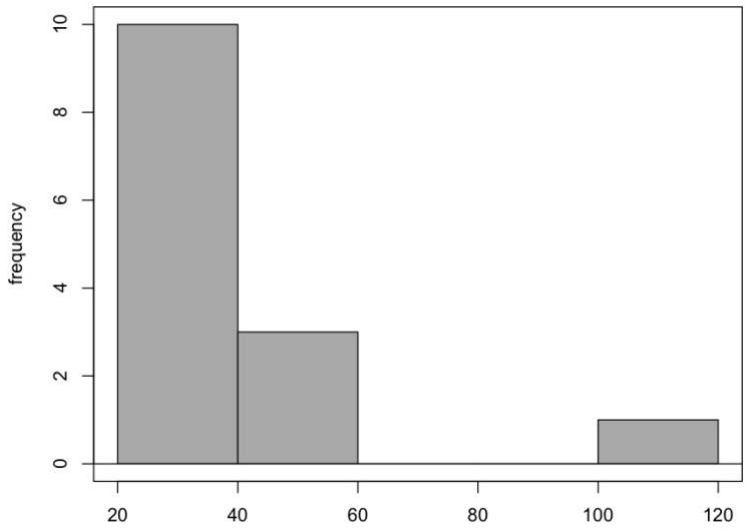
a) Com varien els estadístics calculats d'una mostra a l'altra? Quins es veuen més afectats? I quins menys?
b) Què pot deduir-se de les gràfiques sobre la simetria d'ambdues distribucions?
c) Què pot deduir-se dels diagrames de caixa?

```
> numSummary(Problema7[,"Nous", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
    mean       sd IQR 0% 25% 50% 75% 100%  n NA
33.84615 7.197756   9 24  27  36  36   44 13  1
```

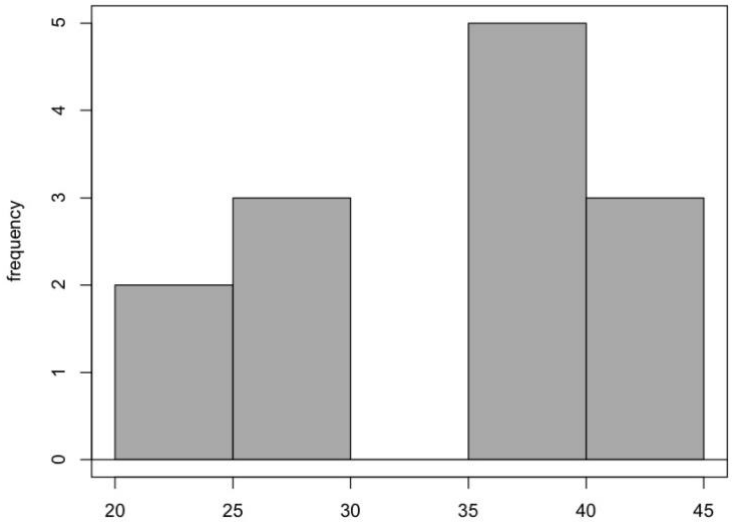
Taula 7

```
> numSummary(Problema7[,"Originals", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
mean       sd IQR 0%  25% 50%  75% 100%  n
  40 24.04163  13 24 27.5  36 40.5  120 14
```

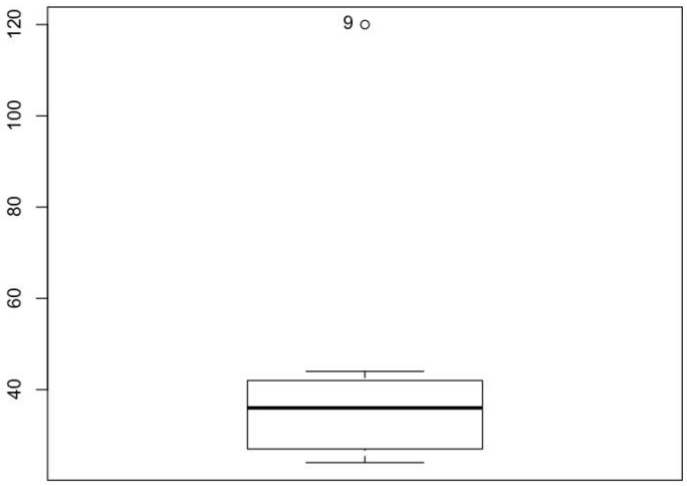
Taula 8



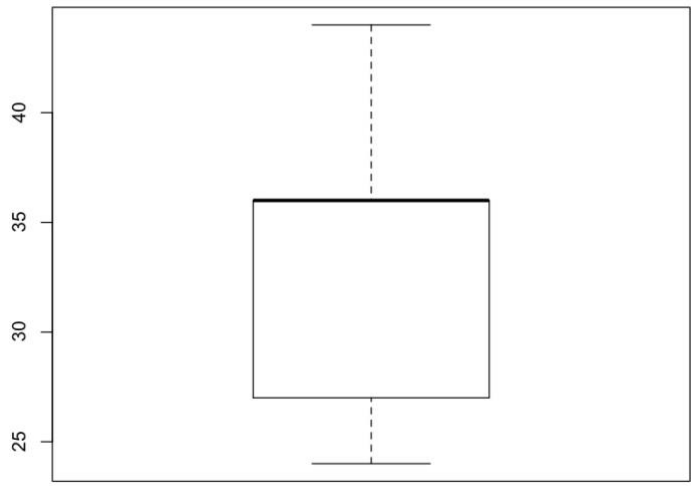
Gràfica 14



Gràfica 15



Gràfica 16



Gràfica 17

EXERCICI 8

Les observacions següents corresponen al nombre de caparres *Ixodes trianguliceps* al cos de 44 ratolins:

0	2	0	0	2	2	0	0	1	1	3
0	0	1	0	0	1	0	1	4	0	0
1	4	2	0	0	1	0	0	2	2	1
1	0	6	0	5	1	3	0	1	0	1

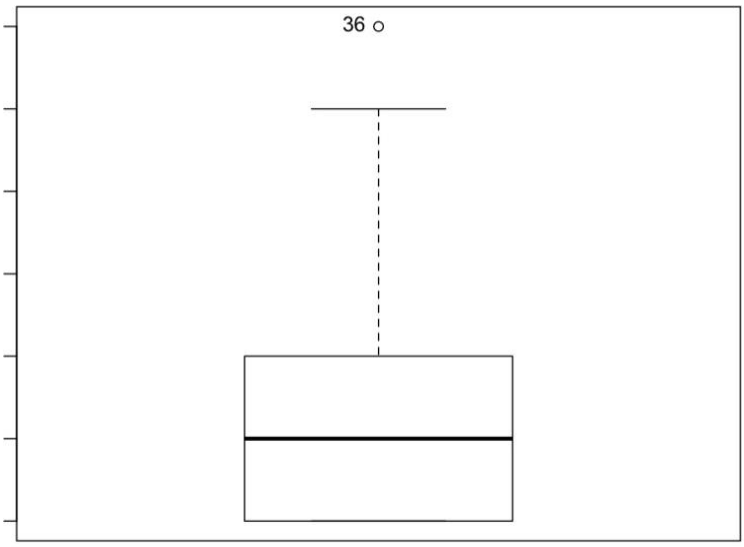
a) Determina la variable observada (i classifica-la), la població i la grandària de la mostra estudiada.
b) Amb l'ajuda de la taula 9, completa l'escala de l'eix vertical de la gràfica 18. Interpreta la taula i la gràfica.

c) A partir de les dades recollides en la taula 10, calcula el percentatge de ratolins que tenen almenys una caparra. Quina és la probabilitat de tenir exactament 4 caparres?

d) Quina és la gràfica més adequada per a aquestes dades? Dibuixa-la.

```
> numSummary(Problema8[,"Caparres", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
     mean       sd IQR 0% 25% 50% 75% 100%  n
 1.113636 1.466144   2  0   0   1   2    6 44
```

Taula 9



Gràfica 18

```
> Problema8 <- within(Problema8, {
+   Caparres <- as.factor(Caparres)
+ })

> local({
+   .Table <- with(Problema8, table(Caparres))
+   cat("\ncounts:\n")
+   print(.Table)
+   cat("\npercentages:\n")
+   print(round(100*.Table/sum(.Table), 2))
+ })

counts:
Caparres
 0  1  2  3  4  5  6
20 12  6  2  2  1  1

percentages:
Caparres
    0     1     2     3     4     5     6
45.45 27.27 13.64  4.55  4.55  2.27  2.27
```

Taula 10

EXERCICI 9

Les dades següents corresponen als nivells (en U/l) de *creatina-fosfocinasa* en la sang de 36 homes sans, ordenades de menor a major:

25	42	48	57	58	60	62	64	67	68	70
78	82	83	84	92	93	94	95	95	100	101
104	110	110	113	118	119	121	123	139	145	151
163	201	203								

a) Calcula el percentatge d'observacions que es troben com a màxim a una distància de:
   - Una desviació típica de la mitjana.
   - Dues desviacions típiques de la mitjana.
   - Tres desviacions típiques de la mitjana.

b) Calcula la probabilitat que un home tinga més de 94 U/l de *creatina-fosfocinasa*. I menys de 82 U/l?

```
> numSummary(Problema9[,"Fosfoquinasa", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
     mean       sd  IQR 0%   25%  50%    75% 100%  n
 98.27778 40.38077 50.5 25 67.75 94.5 118.25  203 36
```

Taula 11

EXERCICI 10

La investigació que feren Rodríguez i Humberto (2014) tenia per objectiu determinar la prevalença en el compliment del tractament farmacològic de malalts adults amb tuberculosi i identificar quina és la dependència o associació del compliment terapèutic amb factors que tenen a veure amb el nivell socioeconòmic, l'equip de professionals de la salut, amb el tractament contra la malaltia i el mateix malalt. En aquest estudi participaren 90 malalts amb tuberculosi atesos al centre de salut. Les dades que es presenten recullen els resultats corresponents als factors servei farmacèutic i escolaritat.

		Compliment terapèutic	
		Sí	No
Considera que les recomanacions del servei farmacèutic són adequades?	Sí	16	42
	No	2	30

		Compliment terapèutic	
		Sí	No
Nivell d'estudis	Sense estudis	6	20
	Primària	2	12
	Batxillerat	10	40

a) Identifica les variables, poblacions i mostres que intervenen en l'estudi.
b) Indica la proporció de malalts amb compliment terapèutic.
c) Indica les proporcions de compliment terapèutic en funció del nivell d'estudis.
d) Fes una descripció gràfica de la variable compliment terapèutic.

PROBLEMES TEMA 2

EXERCICI 1

Se sap que la mitjana del període absolutament refractari dels nervis de les rates no enverinades és d'1,3 ms. L'enverinament hauria de retardar la recuperació del nervi i, per tant, augmentar aquest període. Les mesures fetes en 12 rates van donar els resultats següents (en mil·lisegons)

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

a) Identifica els elements (població, mostra i variable) i l'objectiu de l'estudi.
b) Identifica el mètode inferencial (paramètric/no paramètric) que resulta adequat per a analitzar aquesta mostra, i justifica la teua elecció.
c) Planteja i resol el contrast d'hipòtesi corresponent. Indica clarament el test que apliques i raona la resposta.
d) Explica les conclusions que es desprenen de l'anàlisi.

```
> normalityTest(~Temps, test = "shapiro.test", data = Problema1)

        Shapiro-Wilk normality test

data:  Temps
W = 0.95033, p-value = 0.6418
```

Taula 1

```
> with(Dataset, (t.test(Temps, alternative = "two.sided", mu = 1.3, conf.level = 0.95)))

        One Sample t-test

data:  Temps
t = 3.6283, df = 11, p-value = 0.003969
alternative hypothesis: true mean is not equal to 1.3
95 percent confidence interval:
 1.401623 1.715044
sample estimates:
mean of x
 1.558333
```

Taula 2

```
> with(Dataset, (t.test(Temps, alternative = "greater", mu = 1.3, conf.level = 0.95)))

        One Sample t-test

data:  Temps
t = 3.6283, df = 11, p-value = 0.001984
alternative hypothesis: true mean is greater than 1.3
95 percent confidence interval:
 1.430466      Inf
sample estimates:
mean of x
 1.558333
```

Taula 3

EXERCICI 2

En un experiment per a estudiar la regulació de la secreció d'insulina, es van prendre espècimens de sang de 7 gossos abans i després de l'estimulació elèctrica del nervi vague. Els valors següents mostren, per cada animal, l'increment ("després'" menys "abans") en la concentració d'insulina immunoreactiva en el plasma venós pancreàtic:

30 100 60 30 130 1060 30

A partir dels resultats obtinguts mitjançant *R-Commander*, és possible afirmar, a nivell $\alpha = 0{,}02$, que l'estimulació del nervi vague augmenta la concentració d'insulina immunoreactiva?

```
> normalityTest(~Concentració, test = "shapiro.test", data = Dataset)

        Shapiro-Wilk normality test

data:  Concentració
W = 0.54045, p-value = 0.00005251
```

Taula 4

```
> with(Problema2, wilcox.test(Increment, alternative = "greater", mu = 0))

        Wilcoxon signed rank test with continuity correction

data:  Increment
V = 28, p-value = 0.01077
alternative hypothesis: true location is greater than 0
```

Taula 5

EXERCICI 3

S'està realitzant una investigació sobre l'eficàcia d'una nova dieta per a gallines, que hauria de reduir la quantitat de colesterol dels ous que ponen. Se sap, per experiència, que, amb una dieta estàndard, la quantitat mitjana de colesterol als ous és de 270 mg. Després d'utilitzar la dieta durant algun temps, s'obté una mostra aleatòria de 18 ous i se'n mesura el contingut en colesterol, i resulten els valors següents:

280 241 259 265 260 266 261 255 272
260 260 264 242 250 271 281 276 266

a) Proporciona dues estimacions per interval de confiança per a la mitjana de colesterol, amb un nivell de confiança del 95% i del 99%, respectivament. Quina diferència hi ha entre els intervals?
b) Es redueix el contingut mitjà de colesterol usant la dieta nova?

```
> normalityTest(~Colesterol, test = "shapiro.test", data = Problema3)

        Shapiro-Wilk normality test

data:  Colesterol
W = 0.95747, p-value = 0.5537
```

Taula 6

```
> with(Problema3, (t.test(Colesterol, alternative = "two.sided", mu = 0, conf.level = 0.99)))

        One Sample t-test

data:  Colesterol
t = 99.032, df = 17, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 255.0335 270.4110
sample estimates:
mean of x
 262.7222
```

Taula 7

```
> with(Problema3, (t.test(Colesterol, alternative = "two.sided", mu = 270, conf.level = 0.95)))

        One Sample t-test

data:  Colesterol
t = -2.7433, df = 17, p-value = 0.01386
alternative hypothesis: true mean is not equal to 270
95 percent confidence interval:
 257.1251 268.3194
sample estimates:
mean of x
 262.7222
```

Taula 8

```
> with(Problema3, (t.test(Colesterol, alternative = "less", mu = 270, conf.level = 0.95)))

        One Sample t-test

data:  Colesterol
t = -2.7433, df = 17, p-value = 0.00693
alternative hypothesis: true mean is less than 270
95 percent confidence interval:
     -Inf 267.3372
sample estimates:
mean of x
 262.7222
```

Taula 9

EXERCICI 4

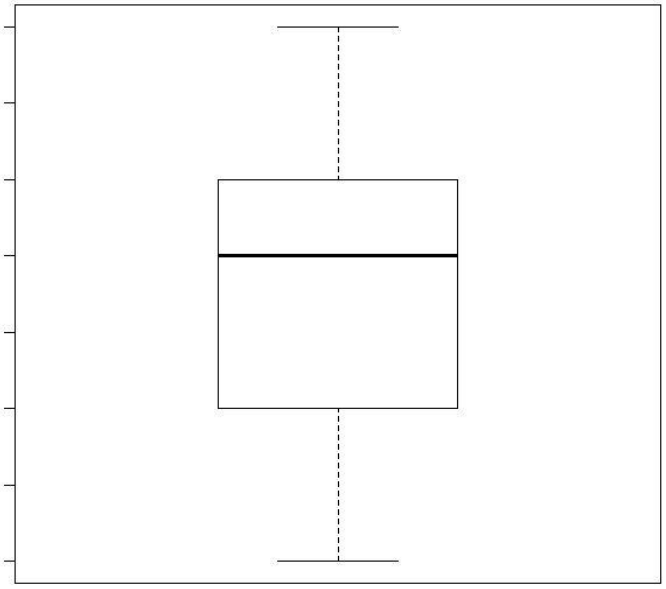
El Departament de Salut dels Estats Units ha fixat en 70, el nombre mitjà de bacteris per centímetre cúbic d'aigua que constitueix un nivell màxim acceptable per a les aigües en què es practica la recollida de cloïsses. Un nivell mitjà superior a 70 és perillós, perquè menjar cloïsses pescades en aquestes aigües pot causar hepatitis. Per a decidir si s'ha de prohibir la pesca de cloïsses en unes aigües concretes, s'ha pres una mostra aleatòria, de 9 observacions, que ha donat els valors següents:

69 74 75 70 72 73 71 73 68

Completa les dades del diagrama de caixa i fes l'anàlisi completa del problema que es planteja. Indica també la decisió que cal prendre per a aquestes aigües.

```
> numSummary(Problema4[,"Bacteris", drop=FALSE], statistics=c("mean", "sd", "IQR", "quantiles"),
+     quantiles=c(0,.25,.5,.75,1))
     mean       sd IQR 0% 25% 50% 75% 100% n
 71.66667 2.345208   3 68  70  72  73   75 9
```

Taula 10



Taula 11

```
> normalityTest(~Bacteris, test="shapiro.test", data=Problema4)

        Shapiro-Wilk normality test

data:  Bacteris
W = 0.96816, p-value = 0.8785
```

Taula 12

```
> with(Problema4, (t.test(Bacteris, alternative='greater', mu=70, conf.level=.95)))

        One Sample t-test

data:  Bacteris
t = 2.132, df = 8, p-value = 0.03279
alternative hypothesis: true mean is greater than 70
95 percent confidence interval:
 70.21299      Inf
sample estimates:
mean of x
 71.66667
```

Taula 13

EXERCICI 5

El calibre d'un arbre és el diàmetre mesurat 6 polzades per damunt de terra. Per a poder afirmar que la mida mitjana dels arbres, amb alçades entre 12 i 14 peus, és proporcional a la resistència del tronc, el calibre mitjà hauria de ser 2 polzades. S'ha obtingut una mostra de 16 arbres, entre 12 i 14 peus d'alçada, cultivats en un viver particular i s'ha determinat el calibre de cada un, i han resultat els valors següents (en polzades):

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

a) Calcula un interval de confiança al 95% del calibre mitjà dels arbres cultivats al viver.
b) És possible concloure que la mida mitjana dels arbres cultivats al viver, amb alçades entre 12 i 14 peus, és proporcional a la resistència del tronc?

```
> normalityTest(~Calibre, test="shapiro.test", data=Problema5)

        Shapiro-Wilk normality test

data:  Calibre
W = 0.97842, p-value = 0.9498
```

Taula 14

```
> with(Problema5, (t.test(Calibre, alternative='two.sided', mu=2.0, conf.level=.95)))

        One Sample t-test

data:  Calibre
t = -3.6942, df = 15, p-value = 0.002165
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 1.517053 1.870447
sample estimates:
mean of x
  1.69375
```

Taula 15

EXERCICI 6

En un experiment per a estudiar les relacions entre paràsits i portadors, es van exposar 242 larves d'arna *Ephestia* a la parasitació de la mosca *ichneumon*. A més **d'estimar el nombre mitjà d'ous per larva**, es desitjava **comprovar si l'afirmació** trobada en nombrosos estudis "el nombre mitjà d'ous per larva és major que 2,5", tenia suport en l'estudi. En la taula següent es mostren el nombre d'ous *ichneumon* trobats en cada larva *Ephestia* (per exemple, 41 larves tenen 3 ous):

Nombre d'ous	0	1	2	3	4	5	6	7
Nombre de larves	21	77	52	41	23	13	9	1

Nombre d'ous	8	9	10	11	12	13	14	15
Nombre de larves	2	0	2	0	0	0	0	1

Resol el problema.

```
> with(Problema6, (t.test(Ous, alternative='two.sided', mu=2.5, conf.level=.95)))

        One Sample t-test

data:  Ous
t = -1.055, df = 241, p-value = 0.2925
alternative hypothesis: true mean is not equal to 2.5
95 percent confidence interval:
 2.120864 2.614674
sample estimates:
mean of x
 2.367769
```

Taula 16

```
> with(Problema6, (t.test(Ous, alternative='greater', mu=2.5, conf.level=.95)))

        One Sample t-test

data:  Ous
t = -1.055, df = 241, p-value = 0.8538
alternative hypothesis: true mean is greater than 2.5
95 percent confidence interval:
 2.160804      Inf
sample estimates:
mean of x
 2.367769
```

Taula 17

EXERCICI 7

Cada espècie de lluernes té una manera peculiar de centelleig. Per a una determinada espècie, consisteix en un centelleig curt de llum seguit per un període de repòs que es pensa que té una duració mitjana de menys de 3,6 segons. Es van obtenir les dades següents sobre el període de repòs entre centellejos per a una mostra de 16 lluernes d'aquesta espècie.

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

a) Proporciona estimacions puntuals per als paràmetres mitjana i variància del període de repòs d'aquesta espècie.
b) Construeix un interval de confiança al 99% per al temps mitjà de repòs entre centellejos de l'espècie de lluernes estudiada.
c) Planteja i resol el problema estadístic corresponent a la teoria exposada.
d) Explica les conclusions que es desprenen de l'anàlisi.

```
> numSummary(Problema7[,"Temps", drop=FALSE], statistics=c("mean", "sd"))
    mean        sd  n
 3.75625 0.3161619 16
```

Taula 18

```
> normalityTest(~Temps, test="shapiro.test", data=Problema7)

        Shapiro-Wilk normality test

data:  Temps
W = 0.97339, p-value = 0.8901
```

Taula 19

```
> with(Problema7, (t.test(Temps, alternative='two.sided', mu=0.0, conf.level=.99)))

        One Sample t-test

data:  Temps
t = 47.523, df = 15, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 3.52334 3.98916
sample estimates:
mean of x
  3.75625
```

Taula 20

```
> with(Problema7, (t.test(Temps, alternative='less', mu=3.6, conf.level=.95)))

        One Sample t-test

data:  Temps
t = 1.9768, df = 15, p-value = 0.9666
alternative hypothesis: true mean is less than 3.6
95 percent confidence interval:
     -Inf 3.894812
sample estimates:
mean of x
  3.75625
```

Taula 21

EXERCICI 8

Diversos investigadors (Link i Panichi, 1979; Gomes, 1982; Gassen, 2001) han mostrat que les xinxes poden produir danys greus en les collites de dacsa. Depenent de les regions i el tipus de xinxa, s'han determinat valors llindars de control. A les regions d'Europa Oriental en què la xinxa s'anomena *Eurygaster integriceps*, aquest llindar se situa en 5 xinxes per m^2. Per tal de prendre mesures, en cas necessari, s'han triat a l'atzar 17 parcel·les de blat d'aquesta zona, cadascuna d'un metre quadrat d'àrea, s'ha observat el corresponent nombre de xinxes per m^2, i ha resultat:

5	6	10	3	2	7	3	3	
3	3	3	4	4	5	8	4	2

A partir dels resultats següents, fes l'anàlisi estadística completa d'aquestes dades per tal de donar alguna conclusió sobre el perill de la plaga.

```
> normalityTest(~Xinxes, test="shapiro.test", data=Problema8)

        Shapiro-Wilk normality test

data:  Xinxes
W = 0.85775, p-value = 0.01411
```

Taula 22

```
> with(Problema8, wilcox.test(Xinxes, alternative='greater', mu=5))

        Wilcoxon signed rank test with continuity correction

data:  Xinxes
V = 38.5, p-value = 0.8975
alternative hypothesis: true location is greater than 5
```

Taula 23

```
> with(Problema8, (t.test(Xinxes, alternative='greater', mu=5, conf.level=.95)))

        One Sample t-test

data:  Xinxes
t = -1.0976, df = 16, p-value = 0.8557
alternative hypothesis: true mean is greater than 5
95 percent confidence interval:
 3.476132      Inf
sample estimates:
mean of x
 4.411765
```

Taula 24

PROBLEMES TEMA 3

EXERCICI 1

En un estudi de Cei & Solaro (1980) es va mesurar, entre altres, la llargària de morro a cloaca en iguanes mascles i femelles.

a) És possible afirmar, a nivell α = 0,01, que la llargària mitjana de morro a cloaca dels mascles és superior a la de les femelles?
b) Proporciona un interval de confiança al 99% per a la diferència de les llargàries mitjanes de morro a cloaca de mascles i femelles $(\mu_{mascles} - \mu_{femelles})$.
c) Altres estudis semblaven demostrar que la diferència de les llargàries mitjanes de morro a cloaca de mascles i femelles és de 7,5 mm. Hi ha evidència, a nivell α = 0,01, que aquesta teoria és incorrecta?

```
> numSummary(problema1[, "LHC", drop = FALSE],
+   groups = problema1$Sexe, statistics = c("mean", "sd"))
             mean       sd LHC:n
Femella 67.51569 6.350413    51
Mascle  71.12791 6.670879    43
```

Taula 1

```
> normalityTest(LHC ~ Sexe, test = "shapiro.test", data = problema1)

 --------
 Sexe = Femella

        Shapiro-Wilk normality test

data:  LHC
W = 0.97725, p-value = 0.4295

 --------
 Sexe = Mascle

        Shapiro-Wilk normality test

data:  LHC
W = 0.98003, p-value = 0.6488
```

Taula 2

```
> leveneTest(LHC ~ Sexe, data = Problema1, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  0.0107 0.9179
      92
```

Taula 3

```
> t.test(LHC~Sexe, alternative='less', conf.level=.95, var.equal=TRUE, data=problema1)

        Two Sample t-test

data:  LHC by Sexe
t = -2.6848, df = 92, p-value = 0.004305
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.376631
sample estimates:
mean in group Femella  mean in group Mascle
             67.51569              71.12791
```

Taula 4

```
> t.test(LHC~Sexe, alternative='two.sided', conf.level=.99, var.equal=TRUE, data=problema1)

        Two Sample t-test

data:  LHC by Sexe
t = -2.6848, df = 92, p-value = 0.008609
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -7.15121220 -0.07322921
sample estimates:
mean in group Femella  mean in group Mascle
             67.51569              71.12791
```

Taula 5

EXERCICI 2

Un equip d'investigació mèdica va estudiar l'efectivitat d'un programa d'entrenament de realimentació biològica dissenyat per a reduir la pressió sanguínia (ps). Els voluntaris van ser assignats aleatòriament a un grup de **biorealimentació** i a un grup de **control**. Tots els voluntaris van rebre informació sobre educació per a la salut. A més, el grup de biorealimentació va rebre vuit setmanes de preparació en relaxació, biorealimentació i exercicis de respiració.

a) Hi havia diferències significatives entre les pressions sanguínies dels dos grups abans de l'experiment?
b) És possible afirmar, a nivell 0,01, que el grup **biorealimentació** va experimentar un descens en la pressió sanguínia?
c) Per al mateix nivell de significació es pot donar idèntica conclusió per al grup de control?
d) Es pot considerar que la biorealimentació és eficaç si la disminució de la pressió sanguínia en el grup **biorealimentació** és superior a la disminució experimentada pel grup de **control**. Hi ha evidència, a nivell 0,01, que la biorealimentació és eficaç?

```
> numSummary(problema2[, c("abans_menys_després", "ps_abans", "ps_després"), drop = FALSE],
+  groups = problema2$Tractament,  statistics = c("mean", "sd"))

Variable: abans_menys_després
          mean       sd  n
Bio     9.9995 1.1658856 20
Control 4.9040 0.9942858 20

Variable: ps_abans
            mean       sd  n
Bio     143.3695 9.093388 20
Control 146.2320 8.206400 20

Variable: ps_després
           mean       sd  n
Bio     133.370 8.533545 20
Control 141.328 8.711413 20
```

Taula 6

```
> normalityTest(ps_abans ~ Tractament, test = "shapiro.test", data = problema2)

 --------
 Tractament = Bio

        Shapiro-Wilk normality test

data:  ps_abans
W = 0.95158, p-value = 0.3918

 --------
 Tractament = Control

        Shapiro-Wilk normality test

data:  ps_abans
W = 0.91647, p-value = 0.08474
```

Taula 7

```
> normalityTest(abans_menys_després ~ Tractament, test = "shapiro.test", data = problema2)

 --------
 Tractament = Bio

        Shapiro-Wilk normality test

data:  abans_menys_després
W = 0.9537, p-value = 0.4267

 --------
 Tractament = Control

        Shapiro-Wilk normality test

data:  abans_menys_després
W = 0.94098, p-value = 0.2502
```

Taula 8

```
> leveneTest(ps_abans ~ Tractament, data = problema2, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  0.5239 0.4736
      38
```

Taula 9

```
> leveneTest(abans_menys_després ~ Tractament, data = problema2, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  1.0672 0.3081
      38
```

Taula 10

```
> t.test(ps_abans ~ Tractament, alternative = "two.sided",
+ conf.level = 0.95, var.equal = TRUE,  data = problema2)

        Two Sample t-test

data:  ps_abans by Tractament
t = -1.0451, df = 38, p-value = 0.3026
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.407173  2.682173
sample estimates:
    mean in group Bio mean in group Control
             143.3695              146.2320
```

Taula 11

```
> with(Bio, (t.test(abans_menys_després, alternative = "greater", mu = 0, conf.level = 0.95)))

        One Sample t-test

data:  abans_menys_després
t = 38.356, df = 19, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 9.548715      Inf
sample estimates:
mean of x
   9.9995
```

Taula 12

```
> with(Control, (t.test(abans_menys_després, alternative = "greater", mu = 0, conf.level = 0.99)))

        One Sample t-test

data:  abans_menys_després
t = 22.057, df = 19, p-value = 2.666e-15
alternative hypothesis: true mean is greater than 0
99 percent confidence interval:
 4.339399      Inf
sample estimates:
mean of x
    4.904
```

Taula 13

```
> t.test(abans_menys_despres~Tractament, alternative='greater', conf.level=.95,
+  var.equal=TRUE, data=problema2)

        Two Sample t-test

data:  abans_menys_despres by Tractament
t = 14.872, df = 38, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.517843      Inf
sample estimates:
    mean in group Bio mean in group Control
               9.9995                4.9040
```

Taula 14

EXERCICI 3

L'adenosina monofosfat cíclica (cAMP) és una substància que pot mediatitzar la resposta cel·lular a les hormones. En un estudi, sobre la maduresa de les cèl·lules ou de la granota *Xenopus laevis*, es van classificar dos oòcits de cadascuna de les quatre femelles en dos grups; a un oòcit es va aplicar progesterona i a l'altre no. Transcorreguts dos minuts, es va determinar el contingut de cAMP en cadascuna de les parelles d'oòcits i es va obtenir:

Granota	cAMP Control	cAMP Progesterona
1	6,01	2,23
2	1,28	1,21
3	3	1,4
4	2,12	1,38

Analitza l'efecte de la progesterona sobre la cAMP, és a dir, la progesterona disminueix la quantitat de cAMP? Resol el problema amb nivell de significativitat α = 0,05 i amb un nivell α = 0,1.

```
> normalityTest(~Proges_menys_Control, test="shapiro.test", data=problema3)

        Shapiro-Wilk normality test

data:  Proges_menys_Control
W = 0.78193, p-value = 0.07357
```

Taula 15

```
> with(problema3, (t.test(Proges_menys_Control, alternative='less', mu=0.0, conf.level=.95)))

        One Sample t-test

data:  Proges_menys_Control
t = -1.5447, df = 3, p-value = 0.1101
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
      -Inf 0.6792873
sample estimates:
mean of x
  -1.2975
```

Taula 16

```
> with(problema3, wilcox.test(Progres_menys_Control, alternative='less', mu=0.0))

        Wilcoxon signed rank test

data:  Progres_menys_Control
V = 0, p-value = 0.0625
alternative hypothesis: true location is less than 0
```

Taula 17

EXERCICI 4

Com a part d'un experiment sobre el metabolisme de les arrels, un botànic va conrear 8 llavors de bedoll en un hivernacle. Va inundar quatre d'aigua durant un dia i va retenir-ne altres quatre com a grup de control. A continuació, va analitzar el contingut d'ATP a les arrels amb l'objectiu de veure si, com que estaven inundades, el contingut era menor. Els resultats (en nmol por mg de teixit) són els que apareixen en la taula adjunta:

Inundades	Control
1,45	1,7
1,19	2,04
1,05	1,49
1,07	1,91

Analitza l'efecte de la inundació, i dona conclusions a nivell $\alpha = 0{,}05$.

```
> normalityTest(ATP ~ Grup, test="shapiro.test", data=problema4)

 --------
 Grup = Control

        Shapiro-Wilk normality test

data:  ATP
W = 0.97489, p-value = 0.8715

 --------
 Grup = Inundades

        Shapiro-Wilk normality test

data:  ATP
W = 0.85577, p-value = 0.2454
```

Taula 18

```
> leveneTest(ATP ~ Grup, data=problema4, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  0.6698 0.4444
       6
```

Taula 19

```
> t.test(ATP~Grup, alternative='greater', conf.level=.95, var.equal=TRUE, data=problema4)

        Two Sample t-test

data:  ATP by Grup
t = 3.9198, df = 6, p-value = 0.003902
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3000353       Inf
sample estimates:
  mean in group Control mean in group Inundades
                  1.785                   1.190
```

Taula 20

EXERCICI 5

Els estudis sobre ocells se solen fer mitjançant anellatge i posada en llibertat, per seguir-ne els moviments. Una de les variables que s'estudia més sovint és la distància de vol des del punt en què s'allibera l'ocell acabat d'anellar fins a la primera posta. Les dades següents corresponen a aquesta distància en dos tipus d'ocells, el pit-roig i el colom de Carolina (la distància es dona en peus).

Pit-roig	128,8	57,2	48,2	160	65,2	69,2	192,1	68,9
	117,3	162,4	24,7	36,5	186,4	37,4	140,8	156,2
	99,7	59,3	70	265	71,3	10	78,8	105,3
Colom	40	381,7	80	120	266,8	13,9	313,9	162,7
	165,5	175,7	76	317,2	55,5	22,1	300,6	44,7
	170	197,7	166,7	263,7	288,1	83,4	369,7	102

Analitza les possibles diferències entre les distàncies recorregudes pels dos tipus d'ocells.

```
> numSummary(Problema5[,"Distancia", drop=FALSE], groups=Problema5$Ocell, statistics=c("mean", "sd", "IQR", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
           mean        sd     IQR   0%    25%   50%     75%  100% Distancia:n
Colom   174.0667 114.25082 193.125 13.9 79.000 166.1 272.125 381.7          24
PitRoig 100.4833  62.54155  85.875 10.0 58.775  75.0 144.650 265.0          24
```

Taula 21

```
> normalityTest(Distancia ~ Ocell, test="shapiro.test", data=Problema5)

--------
Ocell = Colom

        Shapiro-Wilk normality test

data:  Distancia
W = 0.93528, p-value = 0.1279

--------
Ocell = PitRoig

        Shapiro-Wilk normality test

data:  Distancia
W = 0.93523, p-value = 0.1277
```

Taula 22

```
> leveneTest(Distancia ~ Ocell, data=Problema5, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value   Pr(>F)
group  1  9.2652 0.003854 **
      46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula 23

```
> t.test(Distancia~Ocell, alternative='two.sided', conf.level=.95, var.equal=FALSE, data=Problema5)

        Welch Two Sample t-test

data:  Distancia by Ocell
t = 2.7677, df = 35.648, p-value = 0.008896
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  19.64416 127.52251
sample estimates:
  mean in group Colom mean in group PitRoig
             174.0667              100.4833
```

Taula 24

EXERCICI 6

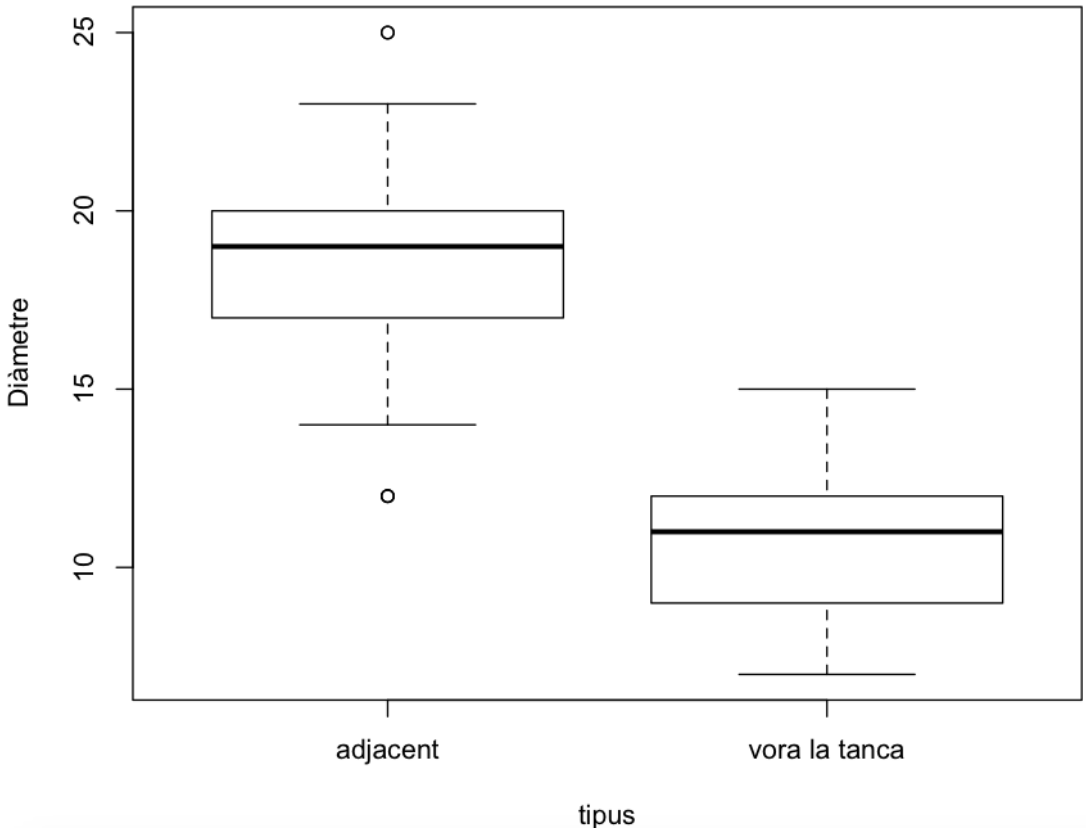
Certes plantes de *Ranunculus acris,* que creixen prop d'una tanca, apareixen a la vista amb botons d'or clarament més menuts que els habituals de l'espècie. Per obtenir una confirmació quantitativa d'aquesta impressió visual, es van mesurar els diàmetres dels botons d'or de les plantes que creixen vora la tanca i d'altres plantes adjacents amb botons d'or normasl. Els diàmetres, en mm, van ser aquests:

Vora la tanca	15	9	13	8	13	9	7	10
	11	9	10	11	11	12	12	11
Adjacents	15	19	14	20	17	12	21	17
	14	16	12	19	12	17	19	19
	19	19	17	19	19	19	20	25
	18	20	21	22	16	19	19	23
	21	20	19	18	18	15	20	15
	19	20	18	21	25	19		

a) Pots donar una estimació per interval al 95% per a la diferència dels diàmetres de les plantes de les dues localitzacions? En cas afirmatiu, proporciona l'interval i comenta el seu significat.
b) Confirmen les dades la impressió visual quant a la diferència de grandària? Contesta a aquesta pregunta amb un nivell 0,05 i amb un nivell 0,001.

```
> numSummary(Problema6[, "Diàmetre", drop = FALSE], groups = Problema6$tipus,
+  statistics = c("mean", "sd", "IQR", "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))
                mean       sd IQR 0% 25% 50% 75% 100% Diàmetre:n
adjacent      18.3913 2.924823   3 12  17  19  20   25         46
vora la tanca 10.6875 2.056494   3  7   9  11  12   15         16
```

Taula 25




```
> normalityTest(Diàmetre ~ tipus, test="shapiro.test", data=Problema6)

 --------
 tipus = adjacent

        Shapiro-Wilk normality test

data:  Diàmetre
W = 0.94264, p-value = 0.02459

 --------
 tipus = vora la tanca

        Shapiro-Wilk normality test

data:  Diàmetre
W = 0.97733, p-value = 0.9388
```

Taula 26

```
> leveneTest(Diàmetre ~ tipus, data=Problema6, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  1  1.2389 0.2701
      60
```

Taula 27

```
> t.test(Diàmetre~tipus, alternative='two.sided', conf.level=.95, var.equal=TRUE, data=Problema6)

        Two Sample t-test

data:  Diàmetre by tipus
t = 9.7094, df = 60, p-value = 6.449e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 6.116699 9.290910
sample estimates:
     mean in group adjacent mean in group vora la tanca
                    18.3913                     10.6875
```

Taula 28

EXERCICI 7

En un estudi parcial de la fisiologia de la maduració del blat, es va tractar d'indagar si la quantitat d'humitat de les llavors de la zona central era superior a la corresponent a les de la zona superior. Es van seleccionar aleatòriament sis plantes de blat d'una parcel·la. Per a cadascuna de les plantes seleccionades es va registrar la quantitat d'humitat de les llavors a la zona central de l'espiga i de les de la zona superior. Les dades obtingudes es presenten a continuació:

Planta	Humitat central	Humitat superior
1	62,7	59,7
2	63,6	61,6
3	60,9	58,2
4	63,0	60,5
5	62,7	60,6
6	63,7	60,8

És major que 2 la diferència d'humitat entre les zones central i superior?

```
> normalityTest(~Central_menys_Superior, test="shapiro.test", data=Problema7)

        Shapiro-Wilk normality test

data:  Central_menys_Superior
W = 0.91824, p-value = 0.4928
```

Taula 29

```
> with(Problema7, (t.test(Central_menys_Superior, alternative='greater', mu=2, conf.level=.95)))

        One Sample t-test

data:  Central_menys_Superior
t = 3.1623, df = 5, p-value = 0.01252
alternative hypothesis: true mean is greater than 2
95 percent confidence interval:
 2.193486      Inf
sample estimates:
mean of x
 2.533333
```

Taula 30

EXERCICI 8

En una investigació sobre el mecanisme de cicatrització de les ferides, es va comparar aquest procés a les extremitats posteriors, dreta i esquerra, de la salamandra *Notophthalmus viridescens.* Després d'amputar cada membre, la biòloga va efectuar una petita cura a la pell i va col·locar un membre en una solució amb benzamil mentre que l'altre es mantingué en una solució de control durant quatre hores. La investigadora pensa que el benzamil pot retardar la cicatrització. Les dades utilitzades en aquesta investigació representen la quantitat de pell cicatritzada, expressada com l'àrea (en mm quadrats) coberta per la nova pell, al cap de quatre hores, als dos membres.

a) Identifica el mètode inferencial (paramètric o no paramètric) que resulta adequat per a analitzar aquesta mostra i justifica la teua elecció.
b) Proporciona, si pots, un interval de confiança al 95% per a la diferència mitjana de quantitat de pell cicatritzada.
c) És efectiu el benzamil amb un nivell de significació $\alpha = 0{,}05$?

```
> numSummary(Problema8[,c("benzamil", "benzamil_menys_control", "control"), drop=FALSE], statistics=c("mean", "sd", "IQR",
+   "quantiles"), quantiles=c(0,.25,.5,.75,1))
                           mean         sd  IQR    0%   25%   50%  75% 100%  n
benzamil                0.11823529 0.09369067 0.09  0.00  0.05  0.11 0.14 0.37 17
benzamil_menys_control -0.09705882 0.14768060 0.16 -0.41 -0.16 -0.08 0.00 0.19 17
control                 0.21529412 0.16424515 0.24  0.00  0.08  0.18 0.32 0.55 17
```

Taula 31

```
> normalityTest(~benzamil_menys_control, test="shapiro.test", data=Problema8)

        Shapiro-Wilk normality test

data:  benzamil_menys_control
W = 0.93663, p-value = 0.2803
```

Taula 32

```
> with(Problema8, (t.test(benzamil, control, alternative='less', conf.level=.95, paired=TRUE)))

        Paired t-test

data:  benzamil and control
t = -2.7098, df = 16, p-value = 0.007729
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.0345251
sample estimates:
mean of the differences
            -0.09705882
```

Taula 33

```
> with(Problema8, (t.test(benzamil_menys_control, alternative='two.sided', mu=0.0, conf.level=.95)))

        One Sample t-test

data:  benzamil_menys_control
t = -2.7098, df = 16, p-value = 0.01546
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.17298918 -0.02112847
sample estimates:
  mean of x
-0.09705882
```

Taula 34

EXERCICI 9

Amb l'objectiu d'investigar l'efecte d'un suplement de calci en pastilles efervescents sobre la relació entre el calci intracel·lular i la pressió sanguínia, un grup d'investigadores va mesurar la concentració de calci lliure a les plaquetes (nM) en 38 persones amb pressions sanguínies normals i en 45 persones amb pressió alta. *Fixa't bé que poden haver-hi taules que no servisquen.*

a) Planteja i resol un contrast d'hipòtesis adequat per a comparar la concentració mitjana de calci lliure a les plaquetes de persones amb pressió sanguínia normal amb la de les persones amb pressió alta.
b) Obté un interval de confiança al 99% per a la diferència de les mitjanes. Interpreta l'interval.

```
> numSummary(Problema9[,"Calci", drop=FALSE], groups=Problema9$Pressió, statistics=c("mean",
+   "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
          mean       sd   IQR    0%      25%    50%      75%   100% Calci:n
Alta   162.7011 35.94429 52.11 91.44 140.3100 157.83 192.4200 243.09      45
Normal 104.8226 15.32051 21.11 78.30  93.9525 103.24 115.0625 134.14      38
```

Taula 35

```
> normalityTest(Calci ~ Pressió, test="shapiro.test", data=Problema9)

--------
Pressió = Alta

        Shapiro-Wilk normality test

data:  Calci
W = 0.98114, p-value = 0.6663

--------
Pressió = Normal

        Shapiro-Wilk normality test

data:  Calci
W = 0.95217, p-value = 0.105
```

Taula 36

```
> leveneTest(Calci ~ Pressió, data=Problema9, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value      Pr(>F)
group  1  24.343 0.000004233 ***
      81
```

Taula 37

```
> t.test(Calci~Pressió, alternative='two.sided', conf.level=.99,
  var.equal=FALSE, data=Problema9)

        Welch Two Sample t-test

data:  Calci by Pressió
t = 9.799, df = 61.579, p-value = 3.491e-14
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 42.17847 73.57849
sample estimates:
  mean in group Alta mean in group Normal
            162.7011              104.8226
```

Taula 38

```
> t.test(Calci~Pressió, alternative='two.sided', conf.level=.95,
  var.equal=TRUE, data=Problema9)

        Two Sample t-test

data:  Calci by Pressió
t = 9.2362, df = 81, p-value = 2.712e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 45.41009 70.34687
sample estimates:
  mean in group Alta mean in group Normal
            162.7011              104.8226
```

Taula 39

EXERCICI 10

La pell dels cadàvers pot utilitzar-se per a proporcionar empelts temporals de pell en persones amb cremades greus. La millora que experimenten els malalts amb aquest tipus d'empelts està en relació directa amb el temps de supervivència de l'empelt, que finalment serà rebutjat pel sistema immunològic del malalt. Un laboratori farmacèutic està investigant l'eficàcia d'un nou tractament contra el rebuig, respecte a l'actual tractament. A cada malalt se li practiquen dos empelts, un amb el nou tractament i l'altre amb l'actual, i es mesura el temps de supervivència en dies.

En els resultats que es mostren a continuació apareix la variable **Dif_Superv**, que es defineix com la diferència "**Superv_Nuevo-Superv_Actual"'**.

a) Proporciona estimacions puntuals del temps mitjà de supervivència dels empelts d'ambdós tractaments. Obté una estimació per intervals per al temps mitjà (o la mediana) de supervivència en cada tractament.
b) Proporciona una estimació puntual de la diferència del temps mitjà de supervivència dels empelts d'ambdós tractaments. Seria vàlida una estimació per intervals?
c) Corroboren aquestes dades que la supervivència dels empelts amb el nou tractament contra el rebuig és superior a l'actual?

```
> Empelts$Superv_Dif <- with(Empelts, Superv_Nou- Superv_Actual)

> numSummary(Empelts[,c("Superv_Actual", "Superv_Dif", "Superv_Nou"), drop=FALSE],
+   statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
                mean       sd IQR 0%  25% 50%  75% 100%  n
Superv_Actual 23.63636 10.88368  11 11 16.5  19 27.5   43 11
Superv_Dif    15.90909 21.21535  14 -6  5.0   8 19.0   67 11
Superv_Nou    39.54545 25.18874  39 16 19.5  29 58.5   93 11
```

Taula 40

```
        Shapiro-Wilk normality test

data:  Superv_Actual
W = 0.88637, p-value = 0.1252


        Shapiro-Wilk normality test

data:  Superv_Dif
W = 0.81964, p-value = 0.01706


        Shapiro-Wilk normality test

data:  Superv_Nou
W = 0.85341, p-value = 0.04728
```

Taula 41

```
> with(Empelts, wilcox.test(Superv_Nou, alternative='two.sided',
 mu=0.0, conf.int=TRUE))

        Wilcoxon signed rank exact test

data:  Superv_Nou
V = 66, p-value = 0.0009766
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 20.0 58.5
sample estimates:
(pseudo)median
         38.75
```

Taula 42

```
> with(Empelts, (t.test(Superv_Actual, alternative='two.sided',
  mu=0.0, conf.level=.95)))

        One Sample t-test

data:  Superv_Actual
t = 7.2028, df = 10, p-value = 0.00002916
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 16.32461 30.94812
sample estimates:
mean of x
 23.63636
```

Taula 43

```
> with(Empelts, wilcox.test(Superv_Nou, Superv_Actual, alternative='greater', paired=TRUE))

        Wilcoxon signed rank test with continuity correction

data:  Superv_Nou and Superv_Actual
V = 60.5, p-value = 0.008131
alternative hypothesis: true location shift is greater than 0
```

Taula 44

PROBLEMES TEMA 4

EXERCICI 1

L'any 1936, sir Ronald Aylmer Fisher va elaborar una base de dades on va fer una classificació de flors de la família *Iris* a partir de les dimensions dels pètals i sèpals. El conjunt de dades consta de les mesures de quatre variables (longituds i amplàries de pètals i sèpals) de 150 exemplars de tres espècies (*setosa*, *versicolor* i *virginica*).

Es desitja comparar les amplàries dels sèpals (**SepalWidth**) de les tres espècies. Fes l'anàlisi escaient a partir de les taules i gràfiques que trobaràs a continuació.

```
> numSummary(problema1[, "SepalWidthCm", drop = FALSE], groups = problema1$Species,
+  statistics = c("mean", "sd", "IQR", "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))
                 mean        sd   IQR  0%   25% 50%   75% 100% SepalWidthCm:n
Iris-setosa     3.418 0.3810244 0.550 2.3 3.125 3.4 3.675  4.4             50
Iris-versicolor 2.770 0.3137983 0.475 2.0 2.525 2.8 3.000  3.4             50
Iris-virginica  2.974 0.3224966 0.375 2.2 2.800 3.0 3.175  3.8             50
```

Taula 1

```
> leveneTest(SepalWidthCm ~ Species, data = problema1, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
       Df F value Pr(>F)
group   2  0.6422 0.5276
      147
```

Taula 2

```
> AnovaModel.1 <- aov(SepalWidthCm ~ Species, data = problema1)

> summary(AnovaModel.1)
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  10.98   5.489   47.36 <2e-16 ***
Residuals   147  17.04   0.116
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula 3

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = SepalWidthCm ~ Species, data = problema1)

Linear Hypotheses:
                                   Estimate Std. Error t value Pr(>|t|)
Iris-versicolor - Iris-setosa == 0  -0.64800    0.06808  -9.518 < 0.0001 ***
Iris-virginica - Iris-setosa == 0   -0.44400    0.06808  -6.521 < 0.0001 ***
Iris-virginica - Iris-versicolor == 0  0.20400    0.06808   2.996  0.00891 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Taula 4

EXERCICI 2

Es va fer un experiment per a determinar l'efectivitat de sis diferents tipus de filtres de membranes, que suporten el creixement de colònies bacterianes. Els recomptes de colònies coliformes fecals, de l'experiència realitzada amb cinc mostres d'aigua del riu, escollides aleatòriament, per a cada filtre, es troben recollits en la taula següent.

G1	G2	G3	G4	G5	G6
33	22	25	33	40	26
37	23	18	25	42	32
34	19	24	31	43	25
38	20	22	28	39	38
34	22	20	29	36	27

a) Identifica les variables.
b) Selecciona el procediment de comparació de mitjanes que consideres adequat (justifica l'elecció). En cas que el contrast de la comparació dels k grups siga significatiu, analitza les diferències entre parelles i construeix els grups homogenis. *Els grups homogenis són els subgrups amb mitjanes similars.*

```
> numSummary(problema2[, "Count", drop = FALSE], groups = problema2$Group,
+ statistics = c("mean", "sd", "IQR", "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))
   mean       sd IQR 0% 25% 50% 75% 100% Count:n
G1 35.2 2.167948   3 33  34  34  37   38       5
G2 21.2 1.643168   2 19  20  22  22   23       5
G3 21.8 2.863564   4 18  20  22  24   25       5
G4 29.2 3.033150   3 25  28  29  31   33       5
G5 40.0 2.738613   3 36  39  40  42   43       5
G6 29.6 5.412947   6 25  26  27  32   38       5
```

Taula 5

```
 Group = G1

        Shapiro-Wilk normality test

data:  Count
W = 0.8713, p-value = 0.2717

 --------
 Group = G2

        Shapiro-Wilk normality test

data:  Count
W = 0.91367, p-value = 0.4899

 --------
 Group = G3

        Shapiro-Wilk normality test

data:  Count
W = 0.96222, p-value = 0.8234
```

Taula 6

```
 Group = G4

        Shapiro-Wilk normality test

data:  Count
W = 0.99174, p-value = 0.9854

 --------
 Group = G5

        Shapiro-Wilk normality test

data:  Count
W = 0.96358, p-value = 0.8327

 --------
 Group = G6

        Shapiro-Wilk normality test

data:  Count
W = 0.86841, p-value = 0.26
```

Taula 7

```
> leveneTest(Count ~ Group, data = problema2, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  5  2.2276 0.0845 .
      24
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula 8

```
> AnovaModel.1 <- aov(Count ~ Group, data = problema2)

> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value        Pr(>F)
Group        5 1355.1  271.02    26.4 0.00000000507 ***
Residuals   24  246.4   10.27
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula 9

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Count ~ Group, data = problema2)

Linear Hypotheses:
             Estimate Std. Error t value Pr(>|t|)
G2 - G1 == 0  -14.000      2.026  -6.908  < 0.001 ***
G3 - G1 == 0  -13.400      2.026  -6.612  < 0.001 ***
G4 - G1 == 0   -6.000      2.026  -2.961  0.06595 .
G5 - G1 == 0    4.800      2.026   2.369  0.20687
G6 - G1 == 0   -5.600      2.026  -2.763  0.09883 .
G3 - G2 == 0    0.600      2.026   0.296  0.99965
G4 - G2 == 0    8.000      2.026   3.948  0.00706 **
G5 - G2 == 0   18.800      2.026   9.277  < 0.001 ***
G6 - G2 == 0    8.400      2.026   4.145  0.00437 **
G4 - G3 == 0    7.400      2.026   3.652  0.01419 *
G5 - G3 == 0   18.200      2.026   8.981  < 0.001 ***
G6 - G3 == 0    7.800      2.026   3.849  0.00889 **
G5 - G4 == 0   10.800      2.026   5.329  < 0.001 ***
G6 - G4 == 0    0.400      2.026   0.197  0.99995
G6 - G5 == 0  -10.400      2.026  -5.132  < 0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Taula 10

EXERCICI 3

Bolek i Coggins (2003) van agafar, durant els mesos d'abril a octubre de 1996, diversos exemplars d'amfibis, incloent-hi 6 salamandres de taques blaves (*Ambystoma laterale*). 7 gripaus americans (*Bufo americanus*) i 6 granotes lleopard (*Rana pipiens*). Després de dissecar-los, van recomptar el nombre de cucs paràsits en cada individu i van obtenir els resultats següents.

Salamandres (**Salamander**)	5	2	3	4	1	3	
Gripaus (**Toad**)	3	2	7	4	3	1	1
Granotes (**Frog**)	6	4	2	7	5	5	

Planteja i resol un test adequat per a comparar la distribució del nombre de paràsits entre aquestes tres espècies. És necessari construir els grups homogenis? Justifica la resposta.

```
 Group = Frog

        Shapiro-Wilk normality test

data:  Count
W = 0.96137, p-value = 0.8302

 --------
 Group = Salamander

        Shapiro-Wilk normality test

data:  Count
W = 0.98176, p-value = 0.96

 --------
 Group = Toad

        Shapiro-Wilk normality test

data:  Count
W = 0.88203, p-value = 0.2356
```

Taula 11

```
> leveneTest(Count ~ Group, data = problema3, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  2    0.22 0.8049
      16
```

Taula 12

```
> AnovaModel.3 <- aov(Count ~ Group, data = problema3)

> summary(AnovaModel.3)
            Df Sum Sq Mean Sq F value Pr(>F)
Group        2  13.80   6.899   2.172  0.146
Residuals   16  50.83   3.177
```

Taula 13

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Count ~ Group, data = problema3)

Linear Hypotheses:
                         Estimate Std. Error t value Pr(>|t|)
Salamander - Frog == 0 -1.833e+00  1.029e+00  -1.782    0.207
Toad - Frog == 0       -1.833e+00  9.917e-01  -1.849    0.186
Toad - Salamander == 0  1.110e-15  9.917e-01   0.000    1.000
(Adjusted p values reported -- single-step method)
```

Taula 14

EXERCICI 4

Considerem una investigació sobre l'amplària de l'escut (closca dorsal) de la larva de la caparra (AEP). Per a indagar si l'hoste influeix en el valor de l'AEP, es tria a l'atzar quatre conills d'angora i es mesura l'AEP de les caparres que es troben en cadascun dels quatre hostes. Les dades obtingudes són:

G1	380	376	360	368	372	366	374	382					
G2	350	356	358	376	338	342	366	350	344	364			
G3	354	360	362	352	366	372	362	344	342	358	351	348	348
G4	320	314	310	360	315	365							

Podem dir que hi ha evidència estadística que l'hoste influeix en el resultat?

```
 Group = G1

        Shapiro-Wilk normality test

data:  Count
W = 0.9768, p-value = 0.9454

 --------
 Group = G2

        Shapiro-Wilk normality test

data:  Count
W = 0.96956, p-value = 0.8867
```

Taula 15

```
 Group = G3

        Shapiro-Wilk normality test

data:  Count
W = 0.97254, p-value = 0.9226

 --------
 Group = G4

        Shapiro-Wilk normality test

data:  Count
W = 0.76479, p-value = 0.02769
```

Taula 16

```
> with(problema4, tapply(Count, Group, median, na.rm = TRUE))
   G1    G2    G3    G4
373.0 353.0 354.0 317.5

> kruskal.test(Count ~ Group, data = problema4)

        Kruskal-Wallis rank sum test

data:  Count by Group
Kruskal-Wallis chi-squared = 15.784, df = 3, p-value = 0.001256
```

Taula 17

```
> pairwise.wilcox.test(problema4$Count,problema4$Group,p.adjust="bonf")

        Pairwise comparisons using Wilcoxon rank sum test

data:  problema4$Count and problema4$Group

   G1     G2     G3
G2 0.0304 -      -
G3 0.0085 1.0000 -
G4 0.0268 0.6931 0.5712

P value adjustment method: bonferroni
```

Taula 18

EXERCICI 5

Les dades de la taula següent recullen el rendiment (en percentatge) de 4 procediments distints d'extracció de penicil·lina a partir de la mateixa matèria primera (extracte de dacsa). Els procediments B, C i D són variants, suposadament millorades, del procediment clàssic A.

a) Els investigadors pensen que aquestes dades s'ajusten perfectament per a fer un test ANOVA. Indica i explica quines condiciones es compleixen.
b) Del valor de l'estadístic de la taula ANOVA, que en podem concloure?
c) Milloren els nous procediments al tradicional?

A	B	C	D
82	87	94	91
85	85	92	92
79	90	90	96
79	88	88	93
84	90	90	90

```
 Procedure = A

        Shapiro-Wilk normality test

data:  Percentage
W = 0.87611, p-value = 0.2921

 --------
 Procedure = B

        Shapiro-Wilk normality test

data:  Percentage
W = 0.91002, p-value = 0.4677
```

Taula 19

```
 Procedure = C

        Shapiro-Wilk normality test

data:  Percentage
W = 0.96086, p-value = 0.814

 --------
 Procedure = D

        Shapiro-Wilk normality test

data:  Percentage
W = 0.94273, p-value = 0.6853
```

Taula 20

```
> leveneTest(Percentage ~ Procedure, data = problema5, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  3  0.2837 0.8364
      16
```

Taula 21

```
> AnovaModel <- aov(Percentage ~ Procedure, data=problema5)

> summary(AnovaModel)
            Df Sum Sq Mean Sq F value    Pr(>F)
Procedure    3  326.9  108.98    19.2 0.0000149 ***
Residuals   16   90.8    5.68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Taula 22

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Percentage ~ Procedure, data = problema5)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
B - A == 0    6.200      1.507   4.115   0.0038 **
C - A == 0    9.000      1.507   5.974   <0.001 ***
D - A == 0   10.600      1.507   7.035   <0.001 ***
C - B == 0    2.800      1.507   1.858   0.2840
D - B == 0    4.400      1.507   2.920   0.0446 *
D - C == 0    1.600      1.507   1.062   0.7166
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Taula 23

```
         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Percentage ~ Procedure, data = problema5)

Quantile = 2.8606
95% family-wise confidence level


Linear Hypotheses:
           Estimate lwr      upr
B - A == 0  6.20000  1.89005 10.50995
C - A == 0  9.00000  4.69005 13.30995
D - A == 0 10.60000  6.29005 14.90995
C - B == 0  2.80000 -1.50995  7.10995
D - B == 0  4.40000  0.09005  8.70995
D - C == 0  1.60000 -2.70995  5.90995
```

Taula 24

EXERCICI 6

En una estació agrícola es va dur a terme un experiment per a estudiar els efectes d'incrementar el nivell d'aplicació de potassi sobre les propietats del cotó resultant. El nivell estàndard (control) que sempre s'utilitzava era de 36 lb K_2O per acre i el nou nivell (experimental) que es va introduir va ser de 72 lb K_2O per acre. Per tal d'analitzar els efectes d'incrementar el potassi, es van triar deu camps a l'atzar i es van dividir els deu camps aleatòriament en dos grups. Es van assignar, aleatòriament, els dos nivells d'aplicació de potassi, un a cada grup. Els resultats de la resistència de les fibres es presenten en la taula següent.

Camp	Control	Experimental
1	7,05	7,95
2	6,75	7,97
3	7,55	8,46
4	7,62	8
5	7,45	7,97

a) Identifica les variables que intervenen en l'estudi.
b) Hi ha evidència estadística, a nivell 0,05, que l'augment del nivell d'aplicació de potassi augmenta la resistència de les fibres?

Un altre equip d'investigadors, convençuts dels beneficis d'augmentar el nivell de potassi, va fer un estudi amb altres nivells d'aplicació (108 i 144). Van triar deu camps diferents i els van repartir aleatòriament en dos grups, aplicant als camps del primer grup el nivell 108 i als cinc del segon grup el nivell 144. Els resultats obtinguts van ser aquests:

Camp	Experimental (108)	Experimental (144)
1	8,29	7,77
2	7,9	7,57
3	7,87	8,14
4	8,12	8,19
5	7,7	7,54

Aquest equip vol analitzar els resultats del seu experiment, però incloent-hi els de l'estació agrícola. Poden dir que hi ha diferències entre els distints nivells d'aplicació de potassi?

```
Grupo = Control

        Shapiro-Wilk normality test

data:  Resistencia
W = 0.88635, p-value = 0.3391

--------
Grupo = Experimental(108)

        Shapiro-Wilk normality test

data:  Resistencia
W = 0.96337, p-value = 0.8312
```

Taula 25

```
Grupo = Experimental(144)

        Shapiro-Wilk normality test

data:  Resistencia
W = 0.85314, p-value = 0.2047

--------
Grupo = Experimental(72)

        Shapiro-Wilk normality test

data:  Resistencia
W = 0.62622, p-value = 0.001368
```

Taula 26

```
> with(problema6, tapply(Resistencia, Grupo.ini, median, na.rm = TRUE))
        Control Experimental(72)
           7.45             7.97

> wilcox.test(Resistencia ~ Grupo.ini, alternative = "less", data = problema6)

        Wilcoxon rank sum test with continuity correction

data:  Resistencia by Grupo.ini
W = 0, p-value = 0.005963
alternative hypothesis: true location shift is less than 0
```

Taula 27

```
> with(problema6, tapply(Resistencia, Grupo, median, na.rm = TRUE))
          Control Experimental(108) Experimental(144)  Experimental(72)
             7.45              7.90              7.77              7.97

> kruskal.test(Resistencia ~ Grupo, data = problema6)

        Kruskal-Wallis rank sum test

data:  Resistencia by Grupo
Kruskal-Wallis chi-squared = 10.219, df = 3, p-value = 0.01679
```

Taula 28

```
> pairwise.wilcox.test(problema6$Resistencia, problema6$Grupo,p.adjust="bonf")

        Pairwise comparisons using Wilcoxon rank sum test

data:  problema6$Resistencia and problema6$Grupo

                  Control Experimental(108) Experimental(144)
Experimental(108) 0.048   -                 -
Experimental(144) 0.333   1.000             -
Experimental(72)  0.072   1.000             1.000

P value adjustment method: bonferroni
```

Taula 29

EXERCICI 7

El potassi està involucrat en el manteniment de l'equilibri normal de l'aigua, l'equilibri osmòtic entre les cèl·lules i el fluid intersticial i l'equilibri àcid-base, determinat pel pH de l'organisme. També està involucrat en la contracció muscular i la regulació de l'activitat neuromuscular, ja que participa en la transmissió de l'impuls nerviós a través dels potencials d'acció de l'organisme.

En un estudi sobre dietes amb un alt contingut en potassi vol determinar-se si la varietat de blat afecta l'aportació de potassi en la dieta. Es va mesurar l'aportació de potassi de les tres varietats de blat més cultivades (compacte, dur i fariner).

a) Planteja i resol el contrast adequat per a comparar el contingut de potassi en les tres varietats de blat.
b) Si creus que hi ha diferències en les aportacions mitjanes de potassi de les tres varietats, determina els grups amb les aportacions que poden considerar-se similars. Si no és el cas, explica la raó per a no fer-ho.

```
> numSummary(Blat[,"Potassi", drop=FALSE], groups=Blat$Varietat,
statistics=c("mean", "sd", "IQR", "quantiles"), quantiles=c(0,.25, .5,.75,1))
            mean        sd    IQR  0%    25%   50%    75% 100% Potassi:n
Compacte 773.4167  46.06213  59.75 702 745.50 774.5 805.25  843        12
Dur      788.8333 151.87425 137.25 497 733.25 810.0 870.50 1000        12
Fariner  830.3333  92.50094 109.00 663 787.75 833.5 896.75  954        12
```

Taula 30

```
> normalityTest(Potassi ~ Varietat, test="shapiro.test", data=Blat)

 --------
 Varietat = Compacte

        Shapiro-Wilk normality test

data:  Potassi
W = 0.94812, p-value = 0.6097

 --------
 Varietat = Dur

        Shapiro-Wilk normality test

data:  Potassi
W = 0.95309, p-value = 0.6825

 Varietat = Fariner

        Shapiro-Wilk normality test

data:  Potassi
W = 0.94873, p-value = 0.6185
```

Taula 31

```
> leveneTest(Potassi ~ Varietat, data=Blat, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value  Pr(>F)
group  2  4.1428 0.02482 *
      33
```

Taula 32

```
> oneway.test(Potassi ~ Varietat, data=Blat) # Welch test

        One-way analysis of means (not assuming equal variances)

data:  Potassi and Varietat
F = 1.7605, num df = 2.00, denom df = 18.38, p-value = 0.1997
```

Taula 33

```
> pairwise.t.test(Blat$Potassi, Blat$Varietat, pool.sd=FALSE)

        Pairwise comparisons using t tests with non-pooled SD

data:  Blat$Potassi and Blat$Varietat

        Compacte Dur
Dur     0.86     -
Fariner 0.22     0.86

P value adjustment method: holm
```

Taula 34

EXERCICI 8

Els triglicèrids són un tipus de lípids que formen part dels greixos. L'excés de concentració sèrica de triglicèrids, superior a 200 mg/dL en sang, s'anomena hipertrigliceridèmia. Aquesta afecció no ha d'anar associada necessàriament a un augment significatiu en els nivells de colesterol. Un nivell alt de triglicèrids pot provocar ateroesclerosi, la qual cosa incrementa el risc de problemes cardiovasculars. Si la hipertrigliceridèmia és molt alta i es té un risc cardiovascular fatal, aleshores els fàrmacs recomanats són niacina, fibrats i estatines.

En un estudi en què participaren 28 malalts amb hipertrigliceridèmia, es provaren els tres fàrmacs, en dosis normal i doble, amb 4 malalts per a cadascuna de les combinacions de cada fàrmac, mentre que als altres 4 se'ls administrà un placebo.

a) Analitza les condicions del disseny i de les distribucions poblacionals, i tria raonadament la tècnica estadística adequada per a comparar els fàrmacs.
b) Amb la tècnica estadística que hages seleccionat en l'apartat anterior, investiga si hi ha diferències entre els resultats dels diferents fàrmacs i, si n' hi ha, estableix els grups de fàrmacs homogenis.
c) Segons les conclusions anteriors, tracta de contestar raonadament a les preguntes següents:
   - Et sembla útil administrar algun fàrmac?
   - Creus que mereix la pena duplicar la dosi del fàrmac?
   - Recomanaries algun fàrmac en particular?

```
> normalityTest(Trigliceridos ~ Tratamiento, test="shapiro.test", data=Trigliceridos)
 --------
 Tratamiento = Estatinas_doble

        Shapiro-Wilk normality test

data:  Trigliceridos
W = 0.9069, p-value = 0.4661
 --------
 Tratamiento = Estatinas_normal

        Shapiro-Wilk normality test

data:  Trigliceridos
W = 0.86964, p-value = 0.2963
 --------
 Tratamiento = Fibratos_doble

        Shapiro-Wilk normality test

data:  Trigliceridos
W = 0.88775, p-value = 0.3728
```

Taula 35

```
 --------
 Tratamiento = Fibratos_normal

        Shapiro-Wilk normality test

data:  Trigliceridos
W = 0.97493, p-value = 0.8717
```

Taula 36

```
 Tratamiento = Niacina_doble

        Shapiro-Wilk normality test

data:  Trigliceridos
W = 0.96278, p-value = 0.7964
 --------
 Tratamiento = Niacina_normal

        Shapiro-Wilk normality test

data:  Trigliceridos
W = 0.85722, p-value = 0.2504
 --------
 Tratamiento = Placebo

        Shapiro-Wilk normality test

data:  Trigliceridos
W = 0.85724, p-value = 0.2505
```

Taula 37

```
> leveneTest(Trigliceridos ~ Tratamiento, data=Trigliceridos, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  6  1.1668 0.3607
      21
```

Taula 38

```
> Anova_Trigliceridos <- aov(Trigliceridos ~ Tratamiento, data=Trigliceridos)
> summary(Anova_Trigliceridos)
            Df Sum Sq Mean Sq F value       Pr(>F)
Tratamiento  6 392447   65408   22.47 0.0000000401 ***
Residuals   21  61121    2911
```

Taula 39

```
> with(Trigliceridos, numSummary(Trigliceridos, groups=Tratamiento, statistics=c("mean", "sd")))
                   mean       sd data:n
Estatinas_doble  113.50 71.28581      4
Estatinas_normal 224.00 56.96198      4
Fibratos_doble   132.25 44.34993      4
Fibrato
Niacina
Niacina
Placebo
```

a 40

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = Trigliceridos ~ Tratamiento, data = Trigliceridos)
Linear Hypotheses:
                                          Estimate Std. Error t value Pr(>|t|)
Estatinas_normal - Estatinas_doble == 0   110.50      38.15   2.897  0.10173
Fibratos_doble - Estatinas_doble == 0      18.75      38.15   0.492  0.99871
Fibratos_normal - Estatinas_doble == 0    131.25      38.15   3.441  0.03350 *
Niacina_doble - Estatinas_doble == 0      -10.00      38.15  -0.262  0.99997
Niacina_normal - Estatinas_doble == 0     154.25      38.15   4.043  0.00887 **
Placebo - Estatinas_doble == 0            356.75      38.15   9.352  < 0.001 ***
Fibratos_doble - Estatinas_normal == 0    -91.75      38.15  -2.405  0.24503
Fibratos_normal - Estatinas_normal == 0    20.75      38.15   0.544  0.99773
Niacina_doble - Estatinas_normal == 0    -120.50      38.15  -3.159  0.06034 .
Niacina_normal - Estatinas_normal == 0     43.75      38.15   1.147  0.90586
Placebo - Estatinas_normal == 0           246.25      38.15   6.455  < 0.001 ***
Fibratos_normal - Fibratos_doble == 0     112.50      38.15   2.949  0.09231 .
Niacina_doble - Fibratos_doble == 0       -28.75      38.15  -0.754  0.98701
Niacina_normal - Fibratos_doble           135.50      38.15   3.552  0.02614 *
Placebo - Fibratos_doble == 0             338.00      38.15   8.860  < 0.001 ***
Niacina_doble - Fibratos_normal == 0     -141.25      38.15  -3.703  0.01902 *
Niacina_normal - Fibratos_normal == 0      23.00      38.15   0.603  0.99600
Placebo - Fibratos_normal == 0            225.50      38.15   5.911  < 0.001 ***
Niacina_normal - Niacina_doble == 0       164.25      38.15   4.306  0.00483 **
Placebo - Niacina_doble == 0              366.75      38.15   9.614  < 0.001 ***
Placebo - Niacina_normal == 0             202.50      38.15   5.308  < 0.001 ***
```

Taula 41

EXERCICI 9

La deficiència de vitamina A és un problema de salut pública. S'ha demostrat que afegint vegetals de fulla verda a la dieta, s'obté un augment de les concentracions en sèrum sanguini de vitamina A.

Es realitza un estudi per a determinar si s'obté algun benefici per afegir greix a la dieta. Un grup de 28 infants, amb concentracions similars de vitamina A en sèrum sanguini, s'assigna aleatòriament a tres subgrups. Cada subgrup rep diàriament 40 g d'espinacs, però el contingut en greix varia. Al final de l'experiment s'obtenen les següents dades de concentració de vitamina A en el sèrum:

Dieta 1	18,1	16,5	21	18,7	7,4	12,4	16,1	17,9		
Dieta 2	29,1	15,8	20,4	23,5	18,5	21,3	23,1	23,8	20,1	11,9
Dieta 3	26,6	16,1	18,8	25	21,8	15,4	19,9	15,5	21,1	25,5

a) Selecciona raonadament la tècnica estadística adequada per a comparar les dietes.

b) Amb la tècnica estadística que hages seleccionat en l'apartat anterior, investiga si el fet d'afegir greix a la dieta afecta la concentració de vitamina A. Utilitza el nivell de significació $\alpha = 0{,}1$.

c) Quins subconjunts homogenis s'obtenen amb el mètode de comparacions múltiples per al nivell $\alpha = 0{,}1$?

```
> normalityTest(vitam_A ~ dieta, test="shapiro.test", data=Vitamina_A)
 --------
 dieta = Dieta 1

        Shapiro-Wilk normality test

data:  vitam_A
W = 0.88981, p-value = 0.2331
 --------
 dieta = Dieta 2

        Shapiro-Wilk normality test

data:  vitam_A
W = 0.97251, p-value = 0.9131
 --------
 dieta = Dieta 3

        Shapiro-Wilk normality test

data:  vitam_A
W = 0.91693, p-value = 0.3321
```

Taula 42

```
> leveneTest(vitam_A ~ dieta, data=Vitamina_A, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group  2  0.0533 0.9482
      25
```

Taula 43

```
> Anova_Vitamina_A <- aov(vitam_A ~ dieta, data=Vitamina_A)
> summary(Anova_Vitamina_A)
             Df Sum Sq Mean Sq F value Pr(>F)
dieta         2  123.6   61.79   3.176  0.059 .
Residuals    25  486.4   19.46
```

Taula 44

```
> with(Vitamina_A, numSummary(vitam_A, groups=dieta, statistics=c("mean", "sd")))
           mean       sd data:n
Dieta 1 16.0125 4.267631      8
Dieta 2 20.7500 4.723993     10
Dieta 3 20.5700 4.191009     10
```

Taula 45

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = vitam_A ~ dieta, data = Vitamina_A)

Linear Hypotheses:
                       Es                d. Error t value Pr(>|t|)
Dieta 2 - Dieta 1 == 0                     2.092   2.264   0.0797 .
Dieta 3 - Dieta 1 == 0    4.558            2.092   2.178   0.0946 .
Dieta 3 - Dieta 2 == 0   -0.180            1.973  -0.091   0.9954
```

Taula 46

EXERCICI 10

El Servei d'Al·lergologia de l'Hospital Universitari La Fe ha fet un estudi sobre al·lèrgies al pol·len. Dins d'aquest estudi es mesurà el contingut mitjà de lípids totals (%) en el pol·len de tres zones pròximes a la ciutat de València.

Zona 1	0,94	0,7	5,75	3,34	2,03	1,35	1,42	0,88	10,11	3,89	2,38	0,74
Zona 2	4,53	1,5	4,38	0,83	18,45	3,25	7,9	8,2	8,31	1,4	2,39	1,2
Zona 3	22,96	2,89	1,88	2.39	1,46	3,03	8,69	3,51	5,95	23,57		

a) Selecciona raonadament la tècnica estadística adequada per a comparar les dietes.
b) Amb la tècnica estadística que hages seleccionat en l'apartat anterior, investiga si el contingut de lípids totals és similar a les tres zones. Utilitza el nivell de significació $\alpha = 0{,}05$.
c) Quins subconjunts homogenis s'obtenen amb el mètode de comparacions múltiples? A què creus que és degut aquest resultat?

```
> normalityTest(Lípidos ~ Zona, test="shapiro.test", data=Polen)
 --------
 Zona = Zona 1

        Shapiro-Wilk normality test

data:  Lípidos
W = 0.76367, p-value = 0.003726
 --------
 Zona = Zona 2

        Shapiro-Wilk normality test

data:  Lípidos
W = 0.79491, p-value = 0.008204
 --------
 Zona = Zona 3

        Shapiro-Wilk normality test

data:  Lípidos
W = 0.69553, p-value = 0.0007778
```

Taula 47

```
> with(Polen, tapply(Lípidos, Zona, median, na.rm=TRUE))
Zona 1 Zona 2 Zona 3
 1.725  3.815  3.270

> kruskal.test(Lípidos ~ Zona, data=Polen)

        Kruskal-Wallis rank sum test

data:  Lípidos by Zona
Kruskal-Wallis chi-squared = 4.4919, df = 2, p-value = 0.1058
```

Taula 48

```
> pairwise.wilcox.test(Polen$Lípidos,Polen$Zona,p.adjust="bonf")

        Pairwise comparisons using Wilcoxon rank sum test

data:  Polen$Lípidos and Polen$Zona

       Zona 1 Zona 2
Zona 2 0.43   -
Zona 3 0.15   1.00

P value adjustment method: bonferroni
```

Taula 49

PROBLEMES TEMA 5

EXERCICI 1

Mendel va crear pèsols heterozigòtics que donaven una qualitat llisa/rugosa. La proporció esperada en la progènie és de 3 qualitat llisa: 1 qualitat rugosa. Ell en va observar 423 de qualitat llisa i 133 de rugosa.

Són els resultats consistents amb la progènie esperada?

- Completa la taula de valors esperats.
- Planteja i resol el contrast d'hipòtesis adient.

```
> chisq.test(c(423,133),p=c(3/4,1/4))

        Chi-squared test for given probabilities

data:  c(423, 133)
X-squared = 0.34532, df = 1, p-value = 0.5568
```

EXERCICI 2

Manano i Meslow (1984) van estudiar com es comportaven en l'alimentació algunes aus en un bosc d'Oregon. En aquell bosc, el 54% del volum d'arbres era *Pseudotsuga menziesii* (PsMen), el 40% era *Pinus ponderosa* (PiPond), el 5% *Abies Grandis* (AG) i l'1% *Larix Occidentalis* (LOcc). Ells van fer 156 observacions en l'alimentació dels enfiladors de pit-roig, 70 observacions (45% del total) en *Pseudotsuga menziesii*, 79 (51%) en *Pinus ponderosa*, 3 (2%) en *Abies Grandis* i 4 (3%) en *Larix Occidentalis*. La hipòtesi biològica és que les aus s'alimenten a l'atzar, sense tindre en compte en quina espècie d'arbre es troben.

Planteja la hipòtesi estadística corresponent. Resol el contrast per a decidir si les dades són consistents amb la hipòtesi biològica.

```
> chisq.test(c(70,79,3,4),p=c(0.54,0.4,0.05,0.01))

        Chi-squared test for given probabilities

data:  c(70, 79, 3, 4)
X-squared = 13.593, df = 3, p-value = 0.003514


> chisq.test(c(70,79,3,4),p=c(0.54,0.4,0.05,0.01))$expected
[1] 84.24 62.40  7.80  1.56
```

EXERCICI 3

Falk i Ayala (1971) van arreplegar les dades de 1.187 persones, van registrar de cadascuna si aplaudien amb el polze dret (D) o el polze esquerre (E) en la part superior. Hi van trobar 535 persones E i 652 persones D. Hi ha diferències d'una proporció 1:1 entre (D) i (E)?

```
> binom.test(535,1187,alternative="two.sided",p=1/2,conf.level=.95)

        Exact binomial test

data:  535 and 1187
number of successes = 535, number of trials = 1187, p-value = 0.0007534
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4221413 0.4795356
sample estimates:
probability of success
             0.4507161
```

EXERCICI 4

McDonald (1989) va arreplegar amfípodes (*Platorchestia platensis*) en una platja de Nova York i va determinar-ne el genotip en la *isomerasa mannosa-6-fosfat (MPI) locus*. Els totals observats en diferents dates van ser: 1002 $MPI^{100/100}$, 1715 $MPI^{100/90}$, i 761 $MPI^{90/90}$, femelles; hi havia 676 $MPI^{100/100}$, 1204 $MPI^{100/90}$, i 442 $MPI^{90/90}$ mascles. Podem dir que hi ha diferències significatives en les proporcions genotípiques entre mascles i femelles?

```
> .Table  # Counts
        Genotip
Sexe      MPI(100/100) MPI(100/90) MPI(90/90)
  Femella         1002        1715        761
  Mascle           676        1204        442
```

```
> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 7.2658, df = 2, p-value = 0.02644
```

```
> .Test$expected # Expected Counts
        Genotip
Sexe      MPI(100/100) MPI(100/90) MPI(90/90)
  Femella    1006.2214    1750.393   721.3852
  Mascle      671.7786    1168.607   481.6148
```

EXERCICI 5

Roberts (1993) mostrejà quadrats en una planura d'inundació i els va classificar d'acord amb les variables: presència/absència d'arbres morts (Si=Mort, No=No mort) i la posició al llarg de transsectes (superior=dunes, mig=intermedi, fons=llac).

a) Planteja el contrast d'hipòtesis adient per analitzar el problema.
b) A partir de la taula de contingència, calcula els valors esperats. Es compleixen les condicions d'aplicabilitat del test khi quadrat?
c) Estudia si la presència d'arbres morts depèn de la posició.

```
> .Table  # Counts
           Mort
Posicio     SI NO
  Llac      15 13
  Intermedi  4  8
  Dunes      0 17
```

```
> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 13.661, df = 2, p-value = 0.00108
```

EXERCICI 6

S'ha demostrat que la dieta mediterrània, rica en vegetals, fruites i llegums, pot disminuir el risc de patir malalties coronàries. Es va dur a terme un estudi per a comprovar si la dieta mediterrània era també recomanable enfront de dietes hipocalòriques, recomanades per l'AHA (American Heart Association) i per a analitzar si, a més, podia tindre un efecte positiu per a reduir el risc de càncer. Utilitzem els resultats obtinguts pels autors; de Lorgeril *et al.* (1998): «Mediterranean dietary pattern in a randomized trial: prolonged survival and possible reduced cancer rate» publicat en Arch. Intern. Med., 158(11), p. 1181-1187.

A un grup de 605 malalts que havien sobreviscut a un atac al cor se'ls va dividir aleatòriament en dos subgrups. Als malalts del grup 1 (grup AHA) se'ls va donar una dieta hipocalòrica basada en les normes de l'AHA (el seguiment dels malalts el feia un metge especialista en malalties coronàries), mentre que als malalts del grup 2 (grup Mediterrània) se'ls va indicar que havien de seguir una dieta mediterrània recomanada per un dietista, que en realitzava un seguiment i control anuals. Al cap de quatre anys es comptabilitzaren els malalts que seguien sans, els que havien patit càncer o alguna altra malaltia i els que havien faltat. S'obtingueren els resultats següents:

	Càncer	Defunció	Altra malaltia	Sans	Total
AHA	15	24	25	239	303
Mediterrània	7	14	8	273	302
Total	22	38	33	512	605

a) Calcula les proporcions de les persones sanes i de les defuncions en cada grup i presenta els valors en una taula.
b) Són compatibles les dades amb la hipòtesi que l'estat de salut no es distribueix de la mateixa manera en els malalts que han seguit aquestes dues dietes diferents? Planteja la hipòtesi nul·la amb paraules i expressa la conclusió del test per $\alpha = 0{,}01$.
c) Proporcionen les dades evidència suficient per a concloure que els malalts que han seguit la dieta mediterrània tenen una major probabilitat d'estar sans? Planteja i resol el contrast d'hipòtesi adequat ($\alpha = 0{,}01$). Calcula i interpreta l'ODDS ràtio.
d) Proporcionen les dades evidència suficient per a concloure que els malalts que han seguit la dieta mediterrània estarien més protegits enfront del càncer? Planteja i resol el contrast d'hipòtesi adequat ($\alpha = 0{,}01$). Calcula i interpreta l'ODDS ràtio.

```
> .Table  # Counts
             Salut
Dieta          Cancer Defuncio Altra Malaltia Sans
  AHA              15       24              25  239
  Mediterranea      7       14               8  273

> rowPercents(.Table) # Row Percentages
             Salut
Dieta          Cancer Defuncio Altra Malaltia Sans Total Count
  AHA             5.0      7.9             8.3 78.9 100.1   303
  Mediterranea    2.3      4.6             2.6 90.4  99.9   302
```

```
> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 16.554, df = 3, p-value = 0.0008726
```

```
> .Test$expected # Expected Counts
              Salut
Dieta            Cancer Defuncion Altra Malaltia     Sans
  AHA          11.01818   19.0314       16.52727 256.4231
  Meidterranea 10.98182   18.9686       16.47273 255.5769
```

```
> .Table1  # Counts
         Dieta
Salut     Mediterranea AHA
  Sano             273 239
  No Sano           29  64

> colPercents(.Table1) # Column Percentages
         Dieta
Salut     Mediterranea   AHA
  Sano            90.4  78.9
  No Sano          9.6  21.1
  Total          100.0 100.0
  Count          302.0 303.0
```

```
> fisher.test(.Table1)

        Fisher's Exact Test for Count Data

data:  .Table1
p-value = 0.0001099
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.540598 4.193340
sample estimates:
odds ratio
  2.517074
```

```
> .Table2  # Counts
           Dieta
Salut       Mediterranea AHA
  Cancer              10  19
  No Cancer          292 284

> colPercents(.Table2) # Column Percentages
           Dieta
Salut       Mediterranea   AHA
  Cancer             3.3   6.3
  No Cancer         96.7  93.7
  Total            100.0 100.0
  Count            302.0 303.0
```

```
> fisher.test(.Table2)

        Fisher's Exact Test for Count Data

data:  .Table2
p-value = 0.1266
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2089918 1.1820407
sample estimates:
odds ratio
 0.5124518
```

EXERCICI 7

L'agressió d'un organisme a una planta pot induir-hi resistència a futures agressions d'organismes diferents? En un estudi sobre aquest problema, es van assignar plantes de cotó a dos grups de forma aleatòria. En un dels grups es van infectar les plantes amb

aranya tacada; les plantes de l'altre grup es van considerar com a grup de control. Transcorregudes dues setmanes, es van netejar les plantes d'aranyes i, posteriorment, es va inocular, a totes les plantes dels dos grups, un fang que causa la malaltia de la tristesa. En la taula següent es mostra el nombre de plantes que van desenvolupar els símptomes de la malaltia esmentada.

		Tractament	
		Aranyes	No aranyes
Resposta al fang	Malaltia	11	17
	No Malaltia	15	4

a) Planteja el contrast d'hipòtesis adient per analitzar el problema.
b) Calcula els valors esperats.
c) Proporcionen les dades prou evidència per a concloure que la infecció amb aranyes indueix resistència a la malaltia de la tristesa? Per a contestar a aquesta pregunta has de:
    - Resoldre el contrast.
    - Calcular els percentatges que consideres necessàries per a contestar a la pregunta: indueix resistència?
d) Contesta la pregunta anterior aplicant el test de Fisher. Calcula i interpreta l'ODDS ràtio.

```
> .Table  # Counts
        Aranyes
Malaltia SI NO
      SI 11 17
      NO 15  4

> colPercents(.Table) # Column Percentages
        Aranyes
Malaltia    SI  NO
   SI     42.3  81
   NO     57.7  19
   Total 100.0 100
   Count  26.0  21

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 7.2037, df = 1, p-value = 0.007275
```

```
> fisher.test(.Table)

        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.008964
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.03404878 0.76046642
sample estimates:
odds ratio
 0.1796308
```

EXERCICI 8

Els medicaments genèrics són aquells que, encara que no tenen patent, mantenen les mateixes qualitats que els de marca coneguda i registrada. Darrerament ha augmentat el consum de medicaments genèrics en nombre d'unitats i ara constitueixen quasi el 50% del total consumit en 2017, segons les dades publicades per l'Asociación Española de Medicamentos Genéricos (AESEG).

En un estudi realitzat a farmàcies de València, de 2.296 unitats venudes de certs medicaments, 1.348 no eren genèrics. Contradiuen aquestes dades la informació del 50% d'unitats de genèrics consumits publicades per l'AESEG?

```
> binom.test(948,2296,alternative="two.sided",p=0.5,conf.level=.95)

        Exact binomial test

data:  948 and 2296
number of successes = 948, number of trials = 2296, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3926572 0.4333494
sample estimates:
probability of success
              0.412892
```

EXERCICI 9

La ciàtica és una síndrome d'inflamació dolorosa del nervi ciàtic. El seu símptoma principal són els formiguejos localitzats a la cama i que, de vegades, poden arribar a l'atròfia muscular. La causa més freqüent de ciàtica és l'hèrnia del disc lumbar, que produeix una lesió permanent. En un estudi sobre l'efecte d'un nou medicament, s'administrà el dit medicament a 128 malalts amb ciàtica. A 82 malalts els van desaparèixer els símptomes en menys d'una setmana.

a) Quina diferència hi ha entre l'ordre executada (color roig) de les dues primeres taules? Afecta el càlcul de l'interval de confiança?
b) Calcula un interval de confiança del 99% per a la proporció de malalts per als quals desapareixeran els símptomes en menys d'una setmana.
c) Si els fabricants del medicament esperaven que els símptomes desaparegueren almenys en un 65% dels malalts, confirmen aquestes dades les seues expectatives?

d) Dona una estimació puntual de la proporció de malalts per als quals desapareixeran els símptomes en menys d'una setmana.

```
> binom.test(82,128,alternative="two.sided",p=0.8,conf.level=.99)

        Exact binomial test

data:  82 and 128
number of successes = 82, number of trials = 128, p-value = 0.00003196
alternative hypothesis: true probability of success is not equal to 0.8
99 percent confidence interval:
 0.5237830 0.7467129
sample estimates:
probability of success
              0.640625
```

```
> binom.test(82,128,alternative="two.sided",p=0.5,conf.level=.99)

        Exact binomial test

data:  82 and 128
number of successes = 82, number of trials = 128, p-value = 0.001862
alternative hypothesis: true probability of success is not equal to 0.5
99 percent confidence interval:
 0.5237830 0.7467129
sample estimates:
probability of success
              0.640625
```

```
> binom.test(82,128,alternative="greater",p=0.65)

        Exact binomial test

data:  82 and 128
number of successes = 82, number of trials = 128, p-value = 0.6267
alternative hypothesis: true probability of success is greater than 0.65
95 percent confidence interval:
 0.5650152 1.0000000
sample estimates:
probability of success
              0.640625
```

EXERCICI 10

Es realitzà un assaig per veure l'eficàcia d'una vacuna contra la COVID-19 en persones majors. Se'n van triar 350 a l'atzar i es van dividir en dos grups de la mateixa grandària; un grup va rebre la vacuna i l'altre un placebo. Després d'un cert període de temps, 28 ancians vacunats i 42 de no vacunats van passar la grip.

a) Quin percentatge d'ancians vacunats i no vacunats passaren la grip?
b) Completa la taula de valors esperats.
c) Planteja el contrast d'hipòtesis adequat per a aquest problema. Comprova les condicions d'aplicabilitat del procediment i resol el problema. Utilitza $\alpha = 0{,}1$.
d) Calcula l'ODDS ràtio i interpreta els resultats obtinguts.

```
> .Table  # Counts
     Tractament
Gripe Vacuna Placebo
   SI     28      42
   NO    147     133

> colPercents(.Table) # Column Percentages
       Tractament
Gripe   Vacuna Placebo
  SI        16      24
  NO        84      76
  Total    100     100
  Count    175     175
```

```
> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 3.5, df = 1, p-value = 0.06137
```

```
> fisher.test(.Table)

        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.08187
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3400202 1.0603844
sample estimates:
odds ratio
 0.6040555
```

EXERCICI 11

En un estudi clínic, uns malalts amb osteoartritis dolorosa de genoll s'assignaren aleatòriament a un dels cinc possibles tractaments: glucosamina, condroïtina, ambdós (glucosamina + condroïtina), placebo o Celebrex, que és la teràpia estàndard. Es prengué nota de si els malalts experimentaven o no una millora substancial en el dolor o presentaven recuperació funcional (resultat satisfactori). Les dades foren:

	Glucosamina	Condroïtina	Ambdós	Placebo	Celebrex
SI	192	202	208	178	214
NO	125	116	109	135	104

```
> .Table  # Counts
            Tractament
Satisfactori Glucosamina Condroitina Ambdos Placebo Celebrex
          SI         192         202    208     178      214
          NO         125         116    109     135      104

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 9.2856, df = 4, p-value = 0.05434


> .Test$expected # Expected Counts
            Tractament
Satisfactori Glucosamina Condroitina   Ambdos  Placebo Celebrex
          SI    199.0512    199.6791 199.0512 196.5395 199.6791
          NO    117.9488    118.3209 117.9488 116.4605 118.3209
```

Planteja i resol el contrast d'hipòtesis adequat per analitzar si les probabilitats d'èxit són les mateixes en tots els tractaments, comprova les condicions d'aplicabilitat del procediment. Resol el problema amb els nivells de significació $\alpha = 0{,}1$ i $\alpha = 0{,}05$.

EXERCICI 12

En un estudi es va examinar l'eficiència relativa de la morfina i la meperidina, medicaments utilitzats per a controlar el dolor en els pacients. En l'estudi, hi van participar 320 pacients entre 20 i 65 anys. La meitat va rebre morfina i l'altra meitat meperidina. Al final de l'estudi, van descriure la seua apreciació del dolor mitjançant una escala amb quatre nivells: agut, moderat, lleuger i sense dolor. Els resultats van ser els següents:

	Agut	Moderat	Lleuger	Sense dolor
Meperidina	16	42	65	37
Morfina	20	30	94	16

a) Proporcionen les dades evidència suficient per a concloure que la probabilitat de dolor agut, entre els pacients que han pres morfina, és menor de 0,15?
b) Quin seria el percentatge de participants amb dolor agut entre els que han pres morfina? I el percentatge dels que van prendre meperidina?
c) Podem concloure que les proporcions de les 4 escales de dolor són compatibles amb les ràtios 2:3:3:2 per als pacients que van prendre meperidina? Quins serien els valors esperats?
d) Proporcionen les dades evidència suficient per a concloure que el nivell de dolor és igual amb tots dos medicaments?

```
> binom.test(20,160,alternative='less',p=0.15)

        Exact binomial test

data:  20 and 160
number of successes = 20, number of trials = 160, p-value = 0.2223
alternative hypothesis: true probability of success is less than 0.15
95 percent confidence interval:
 0.0000000 0.1764363
sample estimates:
probability of success
                 0.125
```

```
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 16.055, df = 3, p-value = 0.001105

> .Test$expected # Expected Counts
             Dolor
Tractament    Agut Moderat Lleuguer Sense
  Meperidina    18      36     79.5  26.5
  Morfina       18      36     79.5  26.5
```

```
> chisq.test(c(16,42,65,37),p=c(0.2,0.3,0.3,0.2))

        Chi-squared test for given probabilities

data:  c(16, 42, 65, 37)
X-squared = 15.552, df = 3, p-value = 0.001401
```

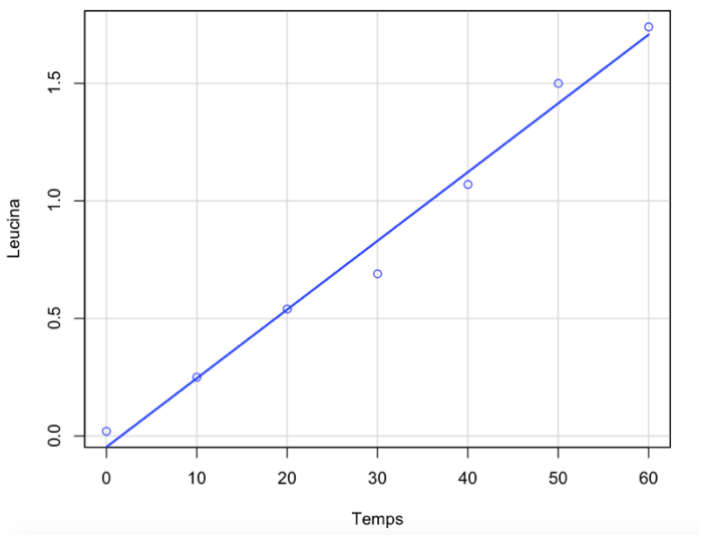
PROBLEMES TEMA 6

EXERCICI 1

En un estudi sobre la síntesi de les proteïnes en l'ovòcit de la granota *Xenopus laevis*, un biòleg va injectar leucina etiquetada com a radioactiva en ovòcits individuals. En diferents moments posteriors a la injecció es van prendre mesures de la radioactivitat i es va calcular la quantitat de leucina que s'havia incorporat a la proteïna. Els resultats es presenten en la taula següent. Cada valor de la leucina és el contingut d'aquesta observat en dos ovòcits (tots els ovòcits són de la mateixa femella).

Temps (min)	0	10	20	30	40	50	60
Leucina (ng)	0,02	0,25	0,54	0,69	1,07	1,5	1,74

a) A la vista del diagrama de dispersió, creus que hi ha relació lineal entre les variables temps i leucina? Calcula el coeficient de correlació lineal i valora la força de la relació lineal.
b) Podem afirmar que en el nivell $\alpha = 0{,}05$, hi ha relació lineal entre el temps i la quantitat de Leucina?
c) Obtín la recta de regressió que explica la relació entre les variables. Interpreta els coeficients de la recta.
d) Quin nivell de leucina correspondria a un temps de 15 minuts? Podries obtenir aquesta predicció per a un temps de dues hores?
e) Quin percentatge de la variabilitat de la leucina està explicat per la regressió lineal?




```
> rcorr.adjust(Problema1[,c("Leucina","Temps")],
  type="pearson", use="complete")

 Pearson correlations:
        Leucina  Temps
Leucina  1.0000 0.9927
Temps    0.9927 1.0000

 Number of observations: 7

 Pairwise two-sided p-values:
        Leucina Temps
Leucina         <.0001
Temps   <.0001

 Adjusted p-values (Holm's method)
        Leucina Temps
Leucina         <.0001
Temps   <.0001
```

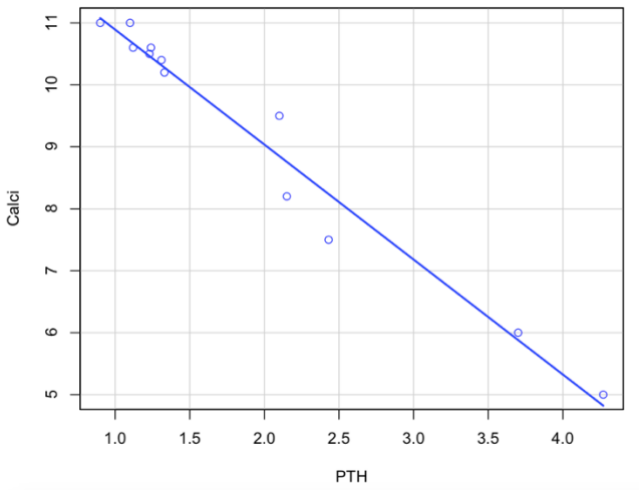
```
> RegModel <- lm(Leucina~Temps, data=Problema1)

> summary(RegModel)

Call:
lm(formula = Leucina ~ Temps, data = Problema1)

Residuals:
      1       2       3       4       5       6       7
 0.0675  0.0050  0.0025 -0.1400 -0.0525  0.0850  0.0325

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) -0.047500   0.057192  -0.831      0.444
Temps        0.029250   0.001586  18.440 0.00000863 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08393 on 5 degrees of freedom
Multiple R-squared:  0.9855, Adjusted R-squared:  0.9826
F-statistic:   340 on 1 and 5 DF,  p-value: 0.000008628
```

EXERCICI 2

La parathormona, també denominada hormona paratiroïdal o PTH, és una hormona segregada per la glàndula paratiroide que intervé en la regulació del metabolisme del calci. En un estudi sobre un nou fàrmac que inhibeix la producció d'hormones per la tiroide es vol estudiar si la concentració de calci en sang depèn de la concentració de PTH. Les dades següents són les mesures de les concentracions de calci (en mg/100ml) i de PTH (en mg/ml) en el plasma sanguini de dotze individus sans:

Calci	11,0	11,0	10,6	10,5	10,6	10,4	10,2	9,5	8,2	7,5	6,0	5,0
PTH	1,10	0,90	1,12	1,23	1,24	1,31	1,33	2,10	2,15	2,43	3,70	4,27

a) A la vista del diagrama de dispersió, creus que hi ha relació lineal entre les dues variables? Calcula el coeficient de correlació lineal i valora la força de la relació lineal.
b) Obtín la recta de regressió adequada per a l'estudi sobre el nou fàrmac.
c) Quin efecte té en el nivell de calci augmentar una unitat el valor de la PTH?
d) Quin nivell de calci correspondria a una concentració de la PTH d'1,5 unitats? Pots obtenir aquesta predicció per a una concentració de PTH de cinc unitats?
e) Quin percentatge de la variabilitat de la concentració del calci no és explicat per la regressió lineal?

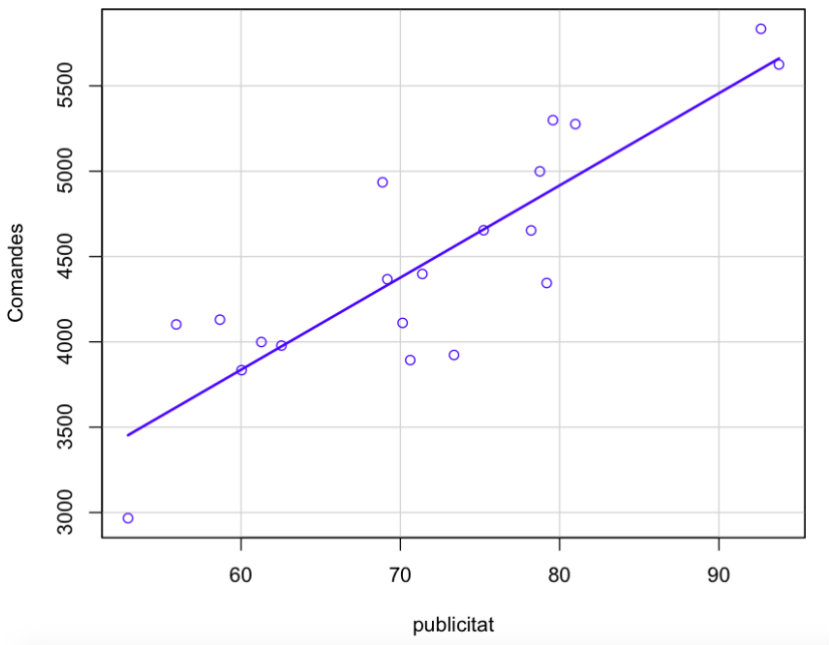


```
> RegModel <- lm(Calci~PTH, data=Problema2)

> summary(RegModel)

Call:
lm(formula = Calci ~ PTH, data = Problema2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73711 -0.07704  0.06012  0.16024  0.65046

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.7468     0.2297   55.48 8.77e-14 ***
PTH          -1.8558     0.1057  -17.55 7.65e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3818 on 10 degrees of freedom
Multiple R-squared:  0.9686, Adjusted R-squared:  0.9654
F-statistic: 308.2 on 1 and 10 DF,  p-value: 0.00000000765
```

EXERCICI 3

A un director general d'una gran companyia li agradaria determinar si s'ha d'assignar més diners al pressupost en publicitat televisiva del pròxim any d'un nou producte. Es pregunta si hi ha una forta relació entre la quantitat de diners gastats en publicitat televisiva d'aquest nou producte i el nombre de comandes rebudes. Les dades, recollides en els últims vint mesos, corresponen a la quantitat de diners mensuals (en milers d'euros) que es gasta en la publicitat televisiva i al nombre de comandes rebudes.

a) Calcula el coeficient de correlació. Valora la força de la relació lineal entre la quantitat de diners gastats en publicitat i el nombre de comandes rebudes.
b) Planteja i resol el contrast de linealitat. Què podem concloure?
c) Prediu el nombre de comandes quan el cost mensual en publicitat és de 65.000 euros.
d) Calcula el coeficient de determinació i interpreta el seu significat.



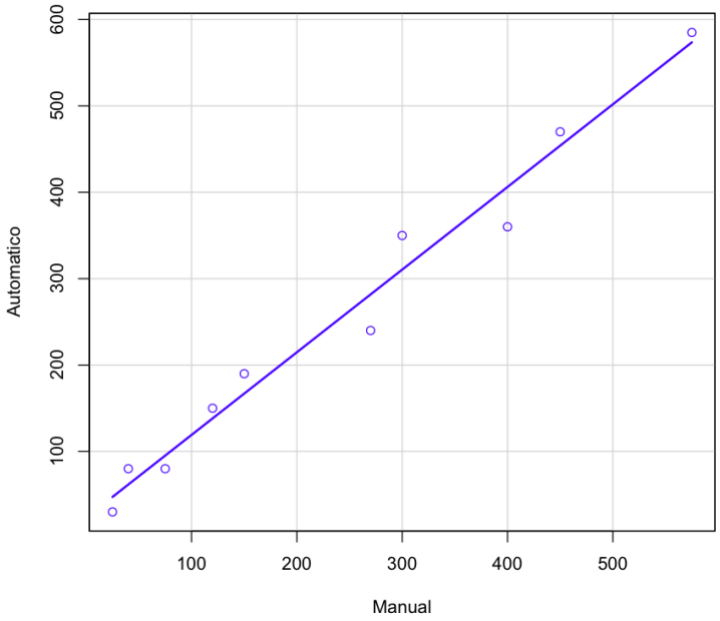
```
> RegModel <- lm(Comandes~publicitat, data=Problema3)

> summary(RegModel)

Call:
lm(formula = Comandes ~ publicitat, data = Problema3)

Residuals:
    Min      1Q  Median      3Q     Max
-635.28 -193.24    0.77  252.67  619.25

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  595.168    544.058   1.094      0.288
publicitat    54.015      7.506   7.196 0.00000107 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 364.4 on 18 degrees of freedom
Multiple R-squared:  0.7421, Adjusted R-squared:  0.7277
F-statistic: 51.78 on 1 and 18 DF,  p-value: 0.00000107
```

EXERCICI 4

La concentració de nitrat en l'aigua és una de les variables estudiades en els treballs que es dediquen a analitzar la influència de les aigües residuals dels pous de visita en la contaminació de les aigües dels llacs. Per a monitorar aquesta variable s'utilitza habitualment un antic mètode manual que, encara que és costós, proporciona una informació pràcticament correcta sobre aquesta concentració. Es proposa un nou mètode de lectura automàtica molt menys costós que el manual. Si aquest nou mètode fóra fiable, es rebutjaria l'antic mètode manual i s'utilitzaria de manera habitual el mètode automàtic. Així, es realitza un experiment que consisteix a determinar la concentració de nitrat (en micrograms per litre d'aigua) en l'aigua de deu mostres utilitzant ambdós mètodes. Els resultats obtinguts són els següents:

Manual	25	40	120	75	150	300	270	400	450	575
Automàtic	30	80	150	80	190	350	240	360	470	585

a) Calcula el coeficient de correlació. Valora la força de la relació lineal entre els dos mètodes, manual i automàtic.
b) Obtín la recta de regressió que explica la relació entre les variables. Interpreta els coeficients.
c) Calcula el coeficient de determinació i interpreta el seu significat.
d) Podem afirmar que el nou mètode és fiable en un nivell de significació $\alpha = 0{,}025$?




```
> rcorr.adjust(aguas[,c("Automatico","Manual")],
type="pearson", use="complete")

 Pearson correlations:
           Automatico Manual
Automatico     1.0000 0.9878
Manual         0.9878 1.0000

 Number of observations: 10

 Pairwise two-sided p-values:
           Automatico Manual
Automatico            <.0001
Manual     <.0001
```

```
> RegModel.3 <- lm(Automatico~Manual, data=aguas)

> summary(RegModel.3)

Call:
lm(formula = Automatico ~ Manual, data = aguas)

Residuals:
   Min     1Q Median     3Q    Max
-46.12 -16.76  11.62  17.77  39.57

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept) 23.38103   16.03262   1.458        0.183
Manual       0.95684    0.05342  17.912 0.0000000967 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.33 on 8 degrees of freedom
Multiple R-squared:  0.9757, Adjusted R-squared:  0.9726
F-statistic: 320.8 on 1 and 8 DF,  p-value: 0.00000009673
```

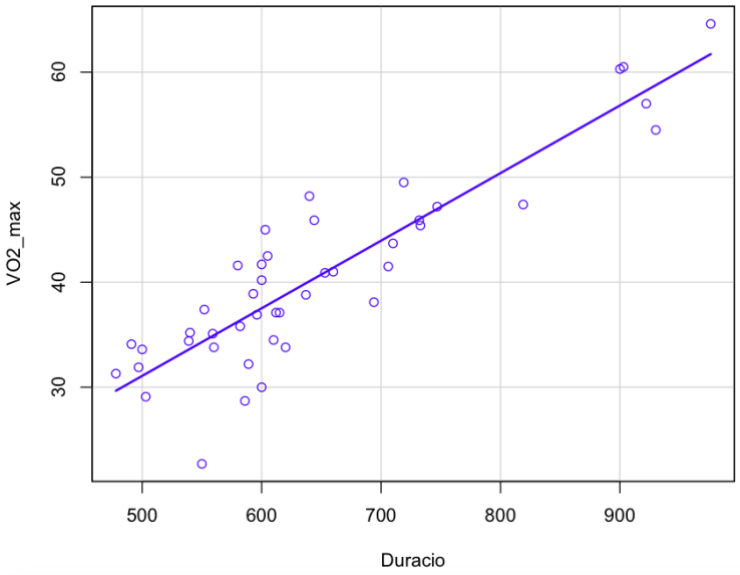
EXERCICI 5

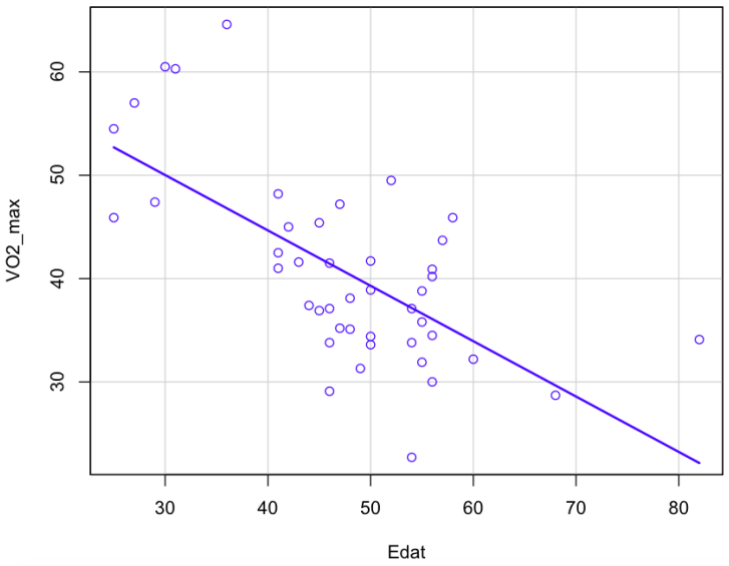
Per determinar l'estat de forma d'un individu sa és útil saber quanta energia és capaç de consumir. Com el consum d'energia requereix oxigen, una manera d'avaluar això és observar la taxa de consum d'oxigen quan es realitza una activitat física forta. Fent ús d'una cinta de córrer s'han desenvolupat diferents mètodes per anar cansant els individus que estan sent avaluats i portar-los fins al màxim de la seua capacitat física. Aquest màxim el determina cada individu: aquest para quan és incapaç de continuar.

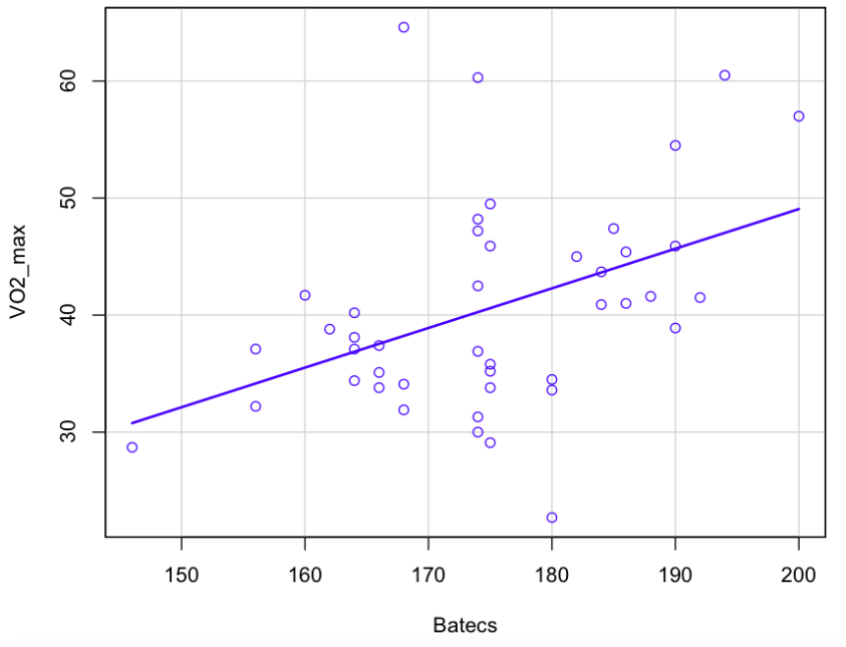
Es disposa de dades referents a 44 individus, tots ells homes sans i actius. En concret, la variable **VO2_màx** (mm./min. x kg.) es va calcular mesurant el volum d'oxigen consumit per minut (mesurament complicat d'obtenir) i dividint pel pes (és a dir, VO2_max és la taxa de consum d'oxigen en mm./min., normalitzada pel pes de cada individu); la variable

duració correspon al temps (en segons) que l'individu roman corrent; la variable **edat** és l'edat en anys de l'individu, i la variable **batecs**, el nombre de batecs per minut.

a) Com l'esforç realitzat depèn del temps que l'individu roman corrent, sembla raonable plantejar-se si hi ha relació entre les variables VO2_màx i la duració. Valora aquesta suposició.
b) Si és així, troba la recta de regressió que permet estimar la taxa normalitzada de consum d'oxigen en funció del temps que un individu roman corrent.
c) Analitza la relació entre la taxa VO2_màx i l'edat. Troba la recta de regressió i interpreta els seus coeficients. Creus que el nou model explica millor la taxa normalitzada de consum d'oxigen (VO2_màx)?
d) Analitza la relació entre la taxa VO2_màx i el nombre de batecs per minut. Troba la recta de regressió i interpreta els seus coeficients.
e) Compara els tres models i justifica quin explica millor la variable VO2_màx.







```
> rcorr.adjust(Problema5[,c("Batecs","Duracio","Edat","VO2_max")],
 type="pearson",  use="complete")

 Pearson correlations:
         Batecs Duracio    Edat VO2_max
Batecs   1.0000  0.4223 -0.5630  0.4277
Duracio  0.4223  1.0000 -0.6634  0.8921
Edat    -0.5630 -0.6634  1.0000 -0.6587
VO2_max  0.4277  0.8921 -0.6587  1.0000

 Number of observations: 44

 Pairwise two-sided p-values:
        Batecs Duracio Edat   VO2_max
Batecs         0.0043  <.0001 0.0038
Duracio 0.0043         <.0001 <.0001
Edat    <.0001 <.0001         <.0001
VO2_max 0.0038 <.0001  <.0001
```

```
> RegModel <- lm(VO2_max~Duracio, data=Problema5)

> summary(RegModel)

Call:
lm(formula = VO2_max ~ Duracio, data = Problema5)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5995  -1.5811  -0.0721   2.9023   8.1073

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.10367    3.31523  -0.333    0.741
Duracio      0.06437    0.00503  12.797 4.39e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.128 on 42 degrees of freedom
Multiple R-squared:  0.7959, Adjusted R-squared:  0.791
F-statistic: 163.8 on 1 and 42 DF,  p-value: 4.385e-16
```

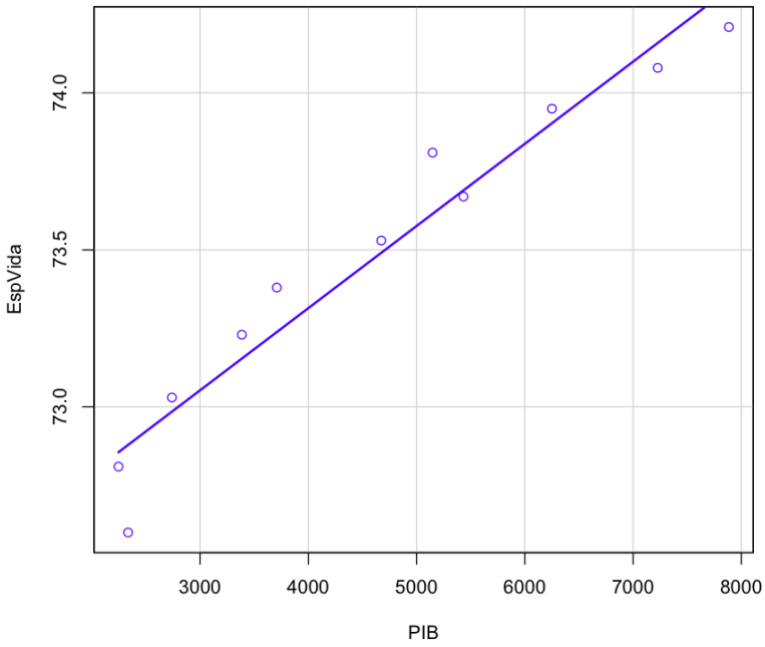
```
> RegModel <- lm(VO2_max~Edat, data=Problema5)

> summary(RegModel)

Call:
lm(formula = VO2_max ~ Edat, data = Problema5)

Residuals:
    Min      1Q  Median      3Q     Max
-14.458  -4.946  -1.204   4.085  17.796

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 66.09728    4.61797  14.313    < 2e-16 ***
Edat        -0.53592    0.09447  -5.673 0.00000117 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.875 on 42 degrees of freedom
Multiple R-squared:  0.4338, Adjusted R-squared:  0.4203
F-statistic: 32.18 on 1 and 42 DF,  p-value: 0.000001172
```

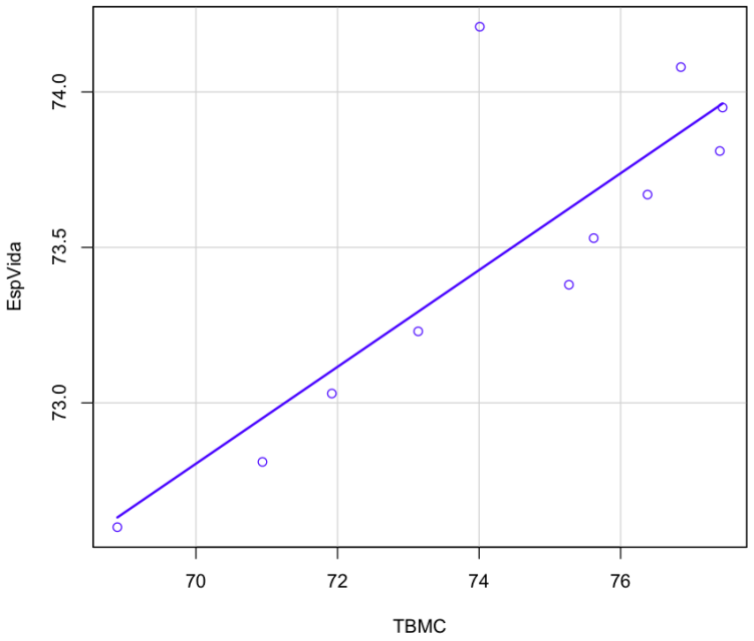
```
> RegModel <- lm(VO2_max~Batecs, data=Problema5)

> summary(RegModel)

Call:
lm(formula = VO2_max ~ Batecs, data = Problema5)

Residuals:
    Min      1Q  Median      3Q     Max
-19.586  -4.807  -1.046   3.357  26.381

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.7199    19.3723  -0.966  0.33941
Batecs        0.3389     0.1105   3.067  0.00378 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.259 on 42 degrees of freedom
Multiple R-squared:  0.183, Adjusted R-squared:  0.1635
F-statistic: 9.405 on 1 and 42 DF,  p-value: 0.003777
```

EXERCICI 6

L'article "*La educación y el ingreso como determinantes de la esperanza de vida en Colombia – 2002-2012*," exposa l'estudi dut a terme per analitzar la influència de factors socioeconòmics, com el nivell econòmic i l'educatiu, en l'esperança de vida a Colòmbia. Disposem de dades sobre les variables següents: el PIB per càpita, la taxa bruta de matriculació combinada (TBMC) calculada com el quocient entre el nombre de matriculats en preescolar, en primària i en secundària i la població total entre 5 i 19 anys, i l'esperança de vida.

a) Hi ha relació lineal entre les variables esperança de vida i producte interior brut? I entre les variables esperança de vida i taxa bruta de matriculació combinada?
b) En les eixides de *R* es proporciona informació sobre dos models de regressió ajustats. El primer d'ells explica la variable esperança de vida en funció del producte interior brut i el segon explica l'esperança de vida en funció de la taxa bruta de matriculació combinada. Quin model proporciona millors resultats? Justifica la resposta.
c) Troba la recta de regressió que permet predir l'esperança de vida a partir de la variable explicativa que hages seleccionat en l'apartat anterior.
d) Calcula l'estimació puntual i per intervals de confiança dels paràmetres del model de regressió lineal.






```
> rcorr.adjust(espvida[,c("EspVida","PIB","TBMC")],
type="pearson", use="complete")

 Pearson correlations:
        EspVida    PIB   TBMC
EspVida  1.0000 0.9685 0.8417
PIB      0.9685 1.0000 0.7169
TBMC     0.8417 0.7169 1.0000

 Number of observations: 11

 Pairwise two-sided p-values:
        EspVida PIB    TBMC
EspVida         <.0001 0.0012
PIB     <.0001         0.0130
TBMC    0.0012  0.0130
```

```
> RegModel.7 <- lm(EspVida~PIB, data=espvida)

> summary(RegModel.7)

Call:
lm(formula = EspVida ~ PIB, data = espvida)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27886 -0.06228  0.03916  0.06144  0.19507

Coefficients:
              Estimate  Std. Error t value    Pr(>|t|)
(Intercept) 72.26760695  0.11196986  645.42     < 2e-16 ***
PIB          0.00026170  0.00002241   11.68 0.000000971 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1378 on 9 degrees of freedom
Multiple R-squared:  0.9381, Adjusted R-squared:  0.9312
F-statistic: 136.4 on 1 and 9 DF,  p-value: 0.0000009707

> Confint(RegModel.7, level=0.95)
                 Estimate         2.5 %        97.5 %
(Intercept) 72.2676069479 72.0143135295 72.5209003662
PIB          0.0002616978  0.0002110021  0.0003123935
```

```
> RegModel.8 <- lm(EspVida~TBMC, data=espvida)

> summary(RegModel.8)

Call:
lm(formula = EspVida ~ TBMC, data = espvida)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24496 -0.14364 -0.07323 -0.02213  0.78127

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  61.9025     2.4777  24.984 0.00000000127 ***
TBMC          0.1557     0.0333   4.677       0.00116 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.299 on 9 degrees of freedom
Multiple R-squared:  0.7085, Adjusted R-squared:  0.6761
F-statistic: 21.87 on 1 and 9 DF,  p-value: 0.001158


> Confint(RegModel.8, level=0.95)
              Estimate       2.5 %     97.5 %
(Intercept) 61.9025506 56.29764958 67.5074516
TBMC         0.1557381  0.08040356  0.2310726
```


```
> RegModel.8 <- lm(EspVida~TBMC, data=espvida)

> summary(RegModel.8)

Call:
lm(formula = EspVida ~ TBMC, data = espvida)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24496 -0.14364 -0.07323 -0.02213  0.78127

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  61.9025     2.4777  24.984 0.00000000127 ***
TBMC          0.1557     0.0333   4.677       0.00116 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.299 on 9 degrees of freedom
Multiple R-squared:  0.7085, Adjusted R-squared:  0.6761
F-statistic: 21.87 on 1 and 9 DF,  p-value: 0.001158


> Confint(RegModel.8, level=0.95)
              Estimate       2.5 %     97.5 %
(Intercept) 61.9025506 56.29764958 67.5074516
TBMC         0.1557381  0.08040356  0.2310726
```

PRÀCTICA 1: Anàlisi exploratòria de dades. Part I

Els objectius d'aquesta pràctica són els següents:

- Identificació dels elements de l'estudi.
- Introducció de les dades.
  - Creació d'un arxiu amb les dades del problema.
  - Lectura d'arxius existents.
- Transformació i categorització de variables.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

IDENTIFICACIÓ DELS ELEMENTS DE L'ESTUDI

La primera cosa que hem de fer per a realitzar qualsevol estudi estadístic és el disseny de l'experiment. Per a treballar amb una mostra amb *n* individus d'una població, hem d'identificar les variables objecte d'estudi i classificar-les, tot indicant, si cal, l'escala de mesura utilitzada.

Per tant, en primer lloc hem de definir la població, la mostra i les variables a estudiar. Una vegada tinguem aquesta informació, hem de classificar les variables per tal d'analitzar-les correctament.

Exercici 1

Volem analitzar alguns aspectes dels estudiants de primer de biològiques de tot Espanya. Per a fer l'estudi, prenem una mostra aleatòria de 10 individus. Les característiques a investigar d'aquesta població són: "Percentatge de xiques (o xics)", "Alçada mitjana", "Pes mitjà" i "Mitjana d'edat".

a) Identifica la població i la mostra.
b) Defineix les variables estadístiques que cal mesurar, indica'n les unitats i l'escala.
c) Classifica les variables de l'apartat b).

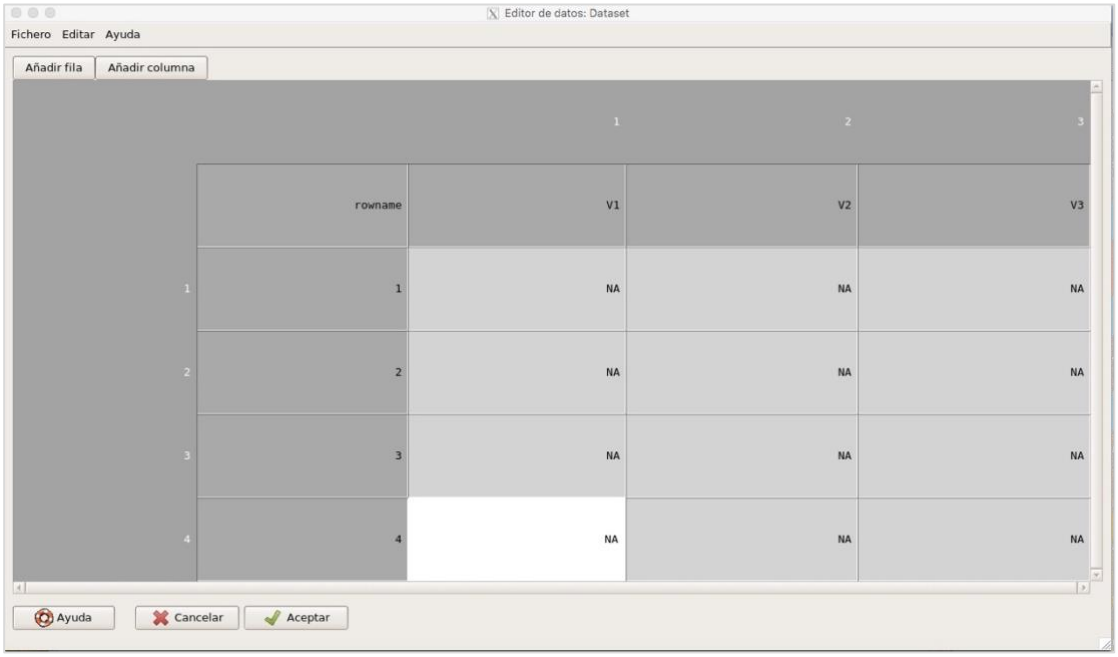
INTRODUCCIÓ DE LES DADES

Una vegada realitzat l'experiment i obtingudes les dades que hem observat, abans de començar l'estudi estadístic hem d'introduir les dades en un arxiu o una base de dades. La majoria de programes estadístics (inclòs *R-Commander*) necessiten les dades en format "taula" (o matriu). En les columnes de la taula apareixen les variables que hem observat en l'experiment i en cada fila tenim els valors de les variables observades per a

cada individu. Per tant, tot allò que sabem d'un individu es trobarà en la fila corresponent a aquest individu.

Dades/Nou conjunt de dades

R-Commander proporciona diverses maneres d'introduir les dades. Podem introduir les

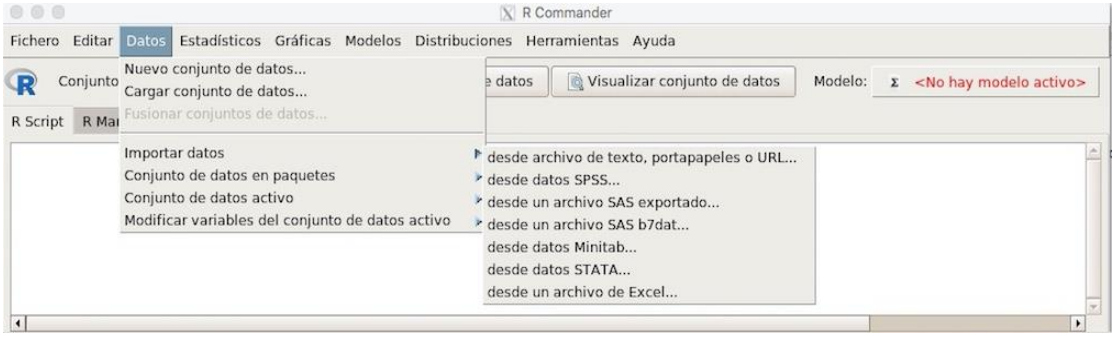


dades directament mitjançant el menú:

Aquest menú, després d'indicar el nom que volem donar al conjunt de dades, obri l'editor de dades, on posarem el nom de les variables i introduirem les dades.

Dades/Importar dades

També podem importar dades des de fitxers de text, dels fitxers creats per altres paquets estadístics (SPSS, Minitab o STATA) o de fulls de càlcul o bases de dades (Excel, Access o dBase) mitjançant l'opció:



Com a exemple il·lustratiu, importarem el fitxer de dades **Picea.xlsx**, en què ens trobem amb les dades recollides per Rudemo (1979) sobre les longituds, en mil·límetres, de 20

cons de coníferes.

En aquest cas, observem que l'arxiu té dues columnes, una per identificar els individus i un altra per a la variable considerada, i 20 files (una per a cada individu considerat).

Picea

	individu	Longitud
1	1	125.1
2	2	114.6
3	3	99.3
4	4	119.1
5	5	109.6
6	6	102.0
7	7	104.9
8	8	109.6
9	9	134.0
10	10	108.6
11	11	120.3
12	12	98.7
13	13	104.2
14	14	91.4
15	15	115.3
16	16	107.7
17	17	97.8
18	18	126.4
19	19	104.8
20	20	118.8

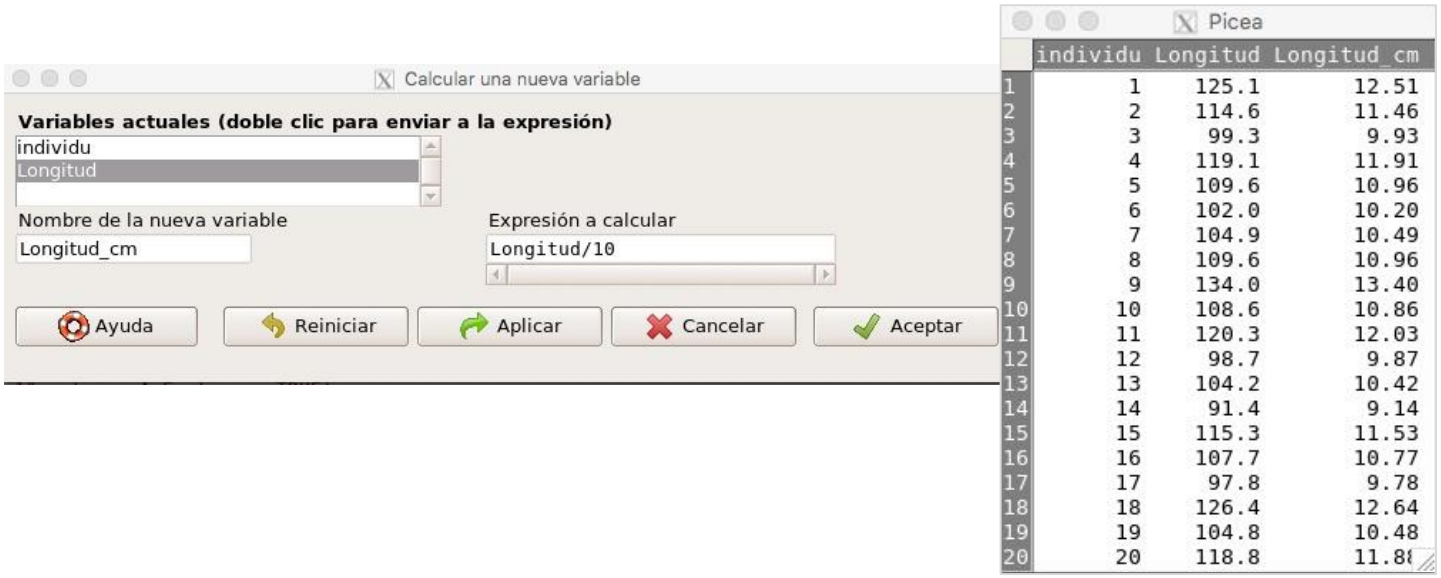
TRANSFORMACIÓ I CATEGORITZACIÓ DE VARIABLES

Una vegada hem elaborat l'arxiu de dades, i prèviament a l'anàlisi estadística, és possible modificar les dades originals que hi ha a l'arxiu. Pot interessar-nos crear noves variables mitjançant **transformacions** de les que ja hi són. Per exemple, si desitgem conèixer el valor de la longitud en centímetres, a partir de les dades importades anteriorment, podem generar una nova variable que continga aquesta informació mitjançant el menú:

> Dades/Modificar les variables del conjunt de dades actiu/Calcular una nova variable

En aquest menú, elegim la variable que volem transformar, el nom de la nova variable i introduïm l'expressió a calcular.

Nou conjunt de dades *Picea*



	individu	Longitud	Longitud_cm
1	1	125.1	12.51
2	2	114.6	11.46
3	3	99.3	9.93
4	4	119.1	11.91
5	5	109.6	10.96
6	6	102.0	10.20
7	7	104.9	10.49
8	8	109.6	10.96
9	9	134.0	13.40
10	10	108.6	10.86
11	11	120.3	12.03
12	12	98.7	9.87
13	13	104.2	10.42
14	14	91.4	9.14
15	15	115.3	11.53
16	16	107.7	10.77
17	17	97.8	9.78
18	18	126.4	12.64
19	19	104.8	10.48
20	20	118.8	11.8

Quan treballem amb variables contínues, si no necessitem el valor exacte d'una observació sinó la seua classificació en categories (creades segons un determinat criteri), podem **categoritzar-la** creant una nova variable de manera que els nous valors siguen una recodificació dels de la variable original. En aquests casos hem de fer el següent:

Dades/Modificar les variables del conjunt de dades actiu/Recodificar variables

En aquest menú hem d'elegir la variable a recodificar i introduir les directrius.

Nou conjunt de dades *Picea*

Picea

	individu	Longitud	Longitud_cm	Longitud_recodificada
1	1	125.1	12.51	Més de 120
2	2	114.6	11.46	106 a 120
3	3	99.3	9.93	105 o menys
4	4	119.1	11.91	106 a 120
5	5	109.6	10.96	106 a 120
6	6	102.0	10.20	105 o menys
7	7	104.9	10.49	105 o menys
8	8	109.6	10.96	106 a 120
9	9	134.0	13.40	Més de 120
10	10	108.6	10.86	106 a 120
11	11	120.3	12.03	Més de 120
12	12	98.7	9.87	105 o menys
13	13	104.2	10.42	105 o menys
14	14	91.4	9.14	105 o menys
15	15	115.3	11.53	106 a 120
16	16	107.7	10.77	106 a 120
17	17	97.8	9.78	105 o menys
18	18	126.4	12.64	Més de 120
19	19	104.8	10.48	105 o menys
20	20	118.8	11.88	106 a 12(

Exercici 2

Farem servir l'article “Dimorfismo sexual de *Liolaemus cuyanus* en una población de San Juan, Argentina” de Cei i Scolaro (1980) per a practicar la introducció de dades i la seua descripció amb *R-Commander*.

a) Llegeix el resum per a identificar els elements d’estudi.
   - Objectiu de l’estudi.
   - Mètode utilitzat pels autors del treball per a demostrar l’objectiu.
   - Població objecte d’estudi.
   - Variables quantitatives a mesurar (indica’n almenys tres).
   - Variables categòriques (indica les categories corresponents).
   - Grandària mostral.

Rev. peru. biol. **14(1):** 047- 050 (Agosto 2007)
© Facultad de Ciencias Biológicas UNMSM

Versión *Online* ISSN 1727-9933

NOTA CIENTÍFICA

Dimorfismo sexual de *Liolaemus cuyanus* Cei & Scolaro, 1980 (Iguania: Liolaemidae) en una población de San Juan, Argentina

Sexual dimorphism of *Liolaemus cuyanus* Cei & Scolaro, 1980 (Iguania: Liolaemidae) in a population of San Juan, Argentina

Alejandro Laspiur y Juan Carlos Acosta

Departamento de Biología e Instituto y Museo de Ciencias Naturales, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan. Av. España 400 (N), CP: 5400, San Juan, Argentina.

Email: Alejandro Laspiur
laspiursaurus@gmail.com

Resumen

El estudio del dimorfismo sexual nos puede ofrecer respuestas sobre el significado biológico que implica la diferenciación morfológica entre machos y hembras. Sin embargo, estas diferencias pueden explicarse consistentemente sabiendo que existen presiones selectivas que influyen en el grado de dimorfismo sexual en las especies. Por estas razones, estudiamos el dimorfismo sexual en una población de *Liolaemus cuyanus* del Monte de San Juan. Medimos 12 variables morfométricas en 51 hembras y 43 machos. De éstas, ocho fueron significativamente más grandes en los machos: el largo hocico-cloaca, el ancho, el largo y el alto de la cabeza, la distancia interocular, longitud del fémur, el largo tibio-fíbula y el largo de la cola. La distancia de separación entre los miembros anteriores y posteriores fue significativamente mayor en las hembras. Se explican y comparan los resultados obtenidos con dimorfismos hallados en otras especies del género.

Palabras claves: Dimorfismo sexual, *Liolaemus cuyanus*, Monte, San Juan, Argentina.

b) Siga la “mini mostra” de dades simulades següent:

Mascles		Femelles	
LHC Longitud morro-cloaca	LCA Longitud cap	LHC Longitud morro-cloaca	LCA Longitud cap
69,5	13,8	62,3	13,4
76,1	16,5	64,0	13,3
72,2	15,4	58,7	12,2
76,0	15,9	65,1	13,8
		73,2	14,3

- Quantes files i columnes necessitem per a descriure les dades de la “mini mostra"?
- Explica el contingut corresponent a cada fila i columna. Indica si alguna variable requereix **etiquetes de valor** (aquest procés és especialment útil si l'arxiu de dades utilitza codis numèrics per a representar categories què no són numèriques. Per exemple, codis 0 i 1 per a home i dona).
- Crea el banc de dades.

PROBLEMES

Per a cadascun dels problemes següents, recordant els conceptes introduïts anteriorment, com aquells corresponents a las sessions de teoria (vegeu tema 1), has de:

a) Classificar la(les) variable(s) observada(es), tot indicant, si cal, l'escala de mesura utilitzada.
b) Identificar la població, la mostra i la grandària de la mostra.
c) Crear o importar l'arxiu de dades, segons corresponga.
d) En el cas que s'indique, realitzar la transformació o categorització corresponent.

PROBLEMA 1

(Font: *The Cotton Rat in Biomedical Research*) Un grup de rates blanques, infectades amb el cuc filarial *Litomosoides carinii*, es van parasitar amb àcars. Posteriorment, es van dissecar i es van comptabilitzar NM="Ne de microfilàries en cada àcar". Els resultats de la dissecció de 40 àcars es recullen en l'arxiu **Pràctica1.xlsx.**

Crea una nova variable que classifique el nombre de microfilàries de cada àcar en 'pocs' si són menys d'11, 'regular' si n'hi ha entre 11 i 20 i 'molts' si n'hi ha més de 20.

PROBLEMA 2

En un experiment dissenyat per a estudiar l'efecte d'un fertilitzant sobre el creixement dels raves, es compara el creixement en dos grups de raves: el grup de control, format per 20 plantes de rave que no han estat tractades amb fertilitzant, i el grup experimental, format per 24 plantes tractades amb aquest fertilitzant. Les dades corresponents a la longitud (en mm), d'un cotilèdon de cadascuna de les plantes considerades es recullen en l'arxiu **Pràctica1.xlsx**.

Crea una nova variable canviant a centímetres la unitat de la longitud en les dades.

PROBLEMA 3

Una de les variables d'interès en l'estudi del cranc *Xantido* (petit cranc que habita prop de Gloucester Point, Virginia) és el nombre d'ous que pon cada individu. Les observacions següents corresponen al nombre d'ous obtinguts per a 25 crancs:

1959	4534	7020	6725	6964	7428	9359	9166	2802	246
4000	3378	7343	4189	8973	4327	2412	7624	1548	480
737	5321	849	5749	6837					

PROBLEMA 4

Un arbre dendrític és una estructura de la cèl·lula que és essencial per a la transmissió dels impulsos nerviosos. En un estudi sobre el desenvolupament del cervell es van observar 36 cèl·lules nervioses procedents dels cervells de cries de conills porquins. Les investigadores van registrar el nombre de branques dendrítiques procedents de cada cèl·lula nerviosa i vam obtenir els resultats recollits en l'arxiu **Pràctica1.xlsx**.

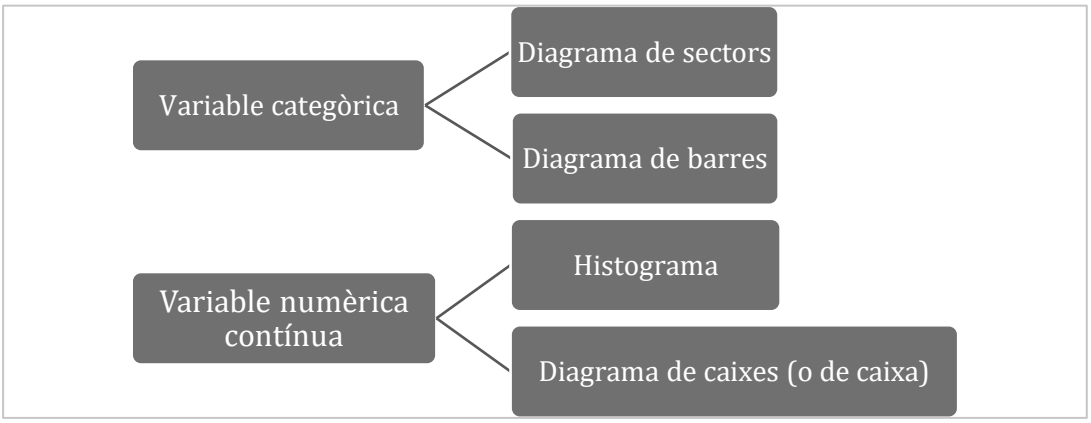
PRÀCTICA 2: Anàlisi exploratòria de dades. Part II

El primer pas de tota anàlisi estadística és l'anàlisi exploratòria de les dades que constitueixen la mostra. Així, doncs, en les pràctiques següents veurem com s'han de calcular els principals estadístics descriptius i construir i interpretar diferents gràfics segons el tipus de variable.

En particular, l'objectiu d'aquesta pràctica és la **descripció gràfica de variables**. Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

REPRESENTACIÓ GRÀFICA DE VARIABLES

La primera cosa que hem de fer a l'hora de representar una variable és classificar-la. Així, serem capaços de triar quin tipus de gràfica és millor. Recordem a continuació la classificació vista en teoria per a la descripció gràfica de variables:



Variable numèrica discreta. Segons que siga la grandària de la mostra, la considerarem categòrica o numèrica contínua.

a) Si és petita, la podem tractar com una variable categòrica, i la representarem amb un diagrama de barres.
b) Si és gran, la tractarem com una variable numèrica contínua, i la representarem amb un histograma o amb un diagrama de caixa.

PROBLEMA

Se sap que l'exposició materna a tòxics mediambientals pot repercutir en la salut i el desenvolupament del fetus. D'altra banda, els nounats resultants d'embarassos gemel·lars monozigòtics (del mateix sexe) suposen una oportunitat única per explorar, de manera

eficient, el paper relatiu de factors genètics en l'associació entre exposició mediambiental i la salut fetal. Es pretén estudiar els efectes adversos de l'exposició a contaminants ambientals durant l'embaràs sobre la salut i el desenvolupament del fetus i del nadó en embarassos gemel·lars monozigòtics. L'arxiu **datwinsdef_lab.xlsx** consta de 108 registres ficticis, cadascun dels quals representa un embaràs gemel·lar. Les variables recollides en la base es detallen en l'arxiu d'Excel que porta per títol **codebooktwins.xlsx**.

A continuació, respon a les preguntes següents:

a) Quantes files apareixen en el fitxer **datwinsdef_lab.xlsx**? A què és degut?
b) Quantes columnes apareixen en el fitxer **datwinsdef_lab.xlsx**? A què és degut?
c) Per començar a treballar amb *R-Commander* necessitem importar les dades del fitxer **datwinsdef_lab.xlsx**. Importa les dades amb el nom **Twins**.

Dades/Importar dades/ Des d'un arxiu Excel

DESCRIPCIÓ GRÀFICA D'UNA VARIABLE CATEGÒRICA

Les variables categòriques poden ser **descrites numèricament** mitjançant taules de freqüències, que indiquen el nombre (o percentatge) de vegades que s'observa cada categoria en la mostra. D'altra banda, les freqüències poden ser representades mitjançant els diagrames de sectors i els diagrames de barres. Treballarem amb la variable **tanalget1** que indica el tipus d'analgèsic consumit durant el primer trimestre de l'embaràs.

A continuació, respon a les preguntes següents:

d) A partir de la informació proporcionada en l'arxiu **codebooktwins.xlsx**, quantes categories té la variable **tanalget1**? Explica el significat de cadascuna d'aquestes categories.
e) Realitza un diagrama de sectors i un diagrama de barres per a representar gràficament la variable **tanalget1**.

Gràfiques/Gràfica de sectors

Gràfiques/Gràfica de barres

f) Canvia el títol del diagrama de sectors a **Analgèsics T1**. Per fer-ho, executa directament la instrucció:

```
with(Twins, pie(table(tanalget1), labels=levels(tanalget1), xlab=" ", ylab=" ",
main="Analgèsics T1", col=rainbow hcl(4)))
```

Nota. La instrucció anterior és la mateixa que es genera quan fas el diagrama de sectors, però modificant el **main**.

g) Representa el diagrama de barres amb les barres en roig, canviant el nom dels eixos x i y a **Tipus d'analgèsic** i **Freqüència**, respectivament.

```
with(Twins, Barplot(tanalget1, xlab="Tipus d'analgèsic", ylab="Freqüència", col="red"))
```

Nota. La instrucció anterior és la mateixa que es genera quan fas el diagrama de barres, però modificant **xlab** i **ylab** i afegint el color amb **col="red".**

DESCRIPCIÓ GRÀFICA D'UNA VARIABLE NUMÈRICA

Els gràfics més adequats per a representar una variable numèrica discreta **amb poques** dades són els diagrames de barres. En el cas de variables numèriques contínues utilitzarem l'histograma (representació gràfica de la distribució de freqüències agrupades) o el diagrama de caixa (representació gràfica de la informació obtinguda en el resum numèric que veurem amb detall en la pràctica següent). Els diagrames de caixa, a més de representar el mínim, màxim i quartils, també ens permeten fer comparacions entre grups definits per una variable categòrica.

A continuació, respon a les preguntes següent:

h) Representa gràficament la variable **pes2** i identifica els possibles valors extrems.

```
Gràfiques/Histograma
```

```
Gràfiques/Diagrama de caixa
```

i) Personalitza l'amplitud dels intervals en l'histograma associat a la variable **pes2** de manera que l'extrem inferior siga 400, el superior 3.700 i l'amplitud 100. Per fer-ho, canvia l'opció **breaks="Sturges"** per **breaks=seq(400,3700,100)**.

j) Descriu gràficament la variable **pes2** en cadascun dels grups definits en la variable **sexe**. Hi ha alguna diferència entre tots dos grups? Els pesos de quin sexe presenten major dispersió?

```
Gràfiques/Diagrama de caixa (seleccionar gràfica per grups)
```

EXERCICI

Per a cada conjunt de dades dels problemes de la pràctica 1, proporciona almenys una representació gràfica. S'adjunten de nou les dades en l'arxiu **Pràctica 2.xlsx**.

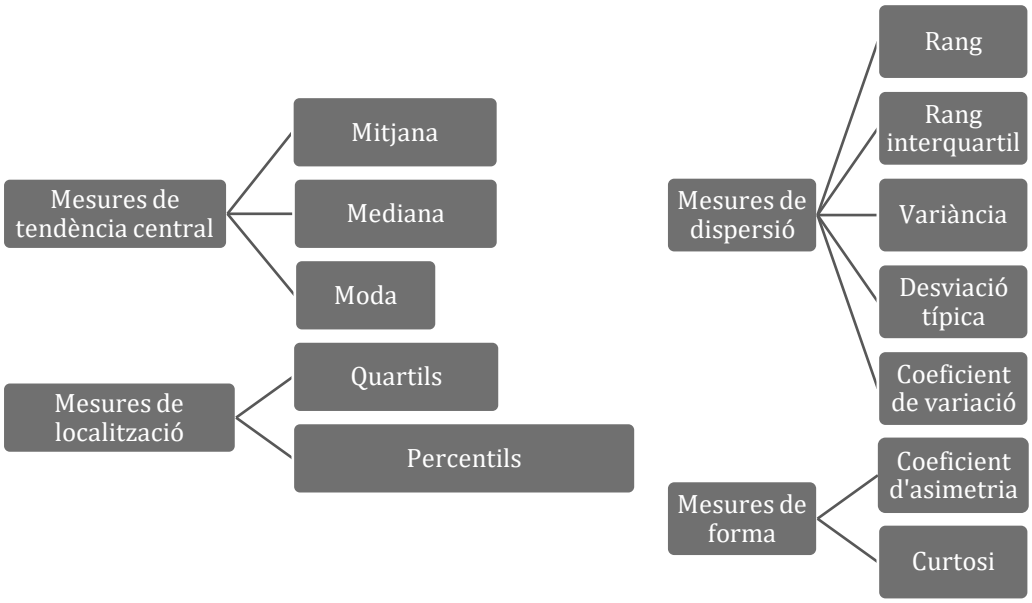
PRÀCTICA 3: Anàlisi exploratòria de dades. Part III

El primer pas de tota anàlisi estadística és l'anàlisi exploratòria de dades que constitueixen la mostra. Així, doncs, en l'anterior pràctica hem vist com construir i interpretar diferents gràfics segons el tipus de variable.

En aquesta pràctica veurem com calcular els **principals estadístics descriptius i la seua interpretació**. A més, completarem la pràctica anterior explicant amb diferents exemples en què consisteix el diagrama de caixa. Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

REPRESENTACIÓ NUMÈRICA DE VARIABLES

Recordem les mesures vistes en teoria per a la descripció numèrica d'una **variable numèrica contínua.**



En el cas de **variables categòriques** (o **numèriques discretes** amb pocs valors possibles), la representació numèrica es fa amb la taula de freqüències (també coneguda com a distribució de freqüències).

PROBLEMA

En aquesta pràctica continuem amb el mateix problema que s'ha enunciat en la pràctica 2.

Se sap que l'exposició materna a tòxics mediambientals pot repercutir en la salut i el desenvolupament del fetus. D'altra banda, els nounats resultants d'embarassos gemel·lars monozigòtics (del mateix sexe) suposen una oportunitat única per explorar, de manera eficient, el paper relatiu de factors genètics en l'associació entre exposició mediambiental i la salut fetal. Es pretén estudiar els efectes adversos de l'exposició a contaminants ambientals durant l'embaràs sobre la salut i el desenvolupament del fetus i del nadó en embarassos gemel·lars monozigòtics. L'arxiu **datwinsdef_lab.xlsx** consta de 108 registres ficticis, cadascun dels quals representa un embaràs gemel·lar. Les variables recollides en la base es detallen en l'arxiu d'Excel que porta per títol **codebooktwins.xlsx**.

A continuació, respon a les preguntes següents:

a) Importa les dades del fitxer **datwinsdef_lab.xlsx**. Importa-les amb el nom **Twins**.
b) Construeix la taula de freqüències corresponent a la variable **tanalget1** i identifica-hi les freqüències absolutes i relatives de cadascuna de les categories.

Estadístics/Resums/Distribució de freqüències

c) Dels 108 embarassos gemel·lars, en quants no es va prendre cap medicament durant el primer trimestre?
d) En quin percentatge d'embarassos gemel·lars estudiats es va consumir paracetamol durant el primer trimestre?
e) Quina és la probabilitat de consumir algun medicament durant el primer trimestre?
f) Calcula i interpreta els estadístics indicats en la introducció de la pràctica per descriure numèricament la variable **pes2**, que indica el pes al moment de nàixer (en grams) del bessó amb menor pes.

Estadístics/Resums/Resums numèrics

f.1) Calcula el percentil 20. Per fer-ho, canvia l'opció **quantiles=c(0,.25,.5,.75,1)** per **quantiles=c(.20)**. Una altra opció és tornar a fer el resum numèric indicant en la pestanya *estadístics* el quantil .20.
f.2) Per sota de quin pes se situa el 10% de les dades? Quin pes deixa per damunt un 20% de dades?
f.3) Hi ha algun bessó amb un pes inferior a 400 grams? I que pese més de 3.700 grams?
f.4) Dibuixa, a mà, amb les dades del resum numèric, el diagrama de caixa. IMPORTANT: Calcula primer els valors extrems (atípics) per a dibuixar-ho correctament. Comprova el resultat amb el diagrama obtingut amb l'eina *R-Commander*.

EXERCICI

Per a cada conjunt de dades dels problemes de la pràctica 1, fes la representació numèrica que consideres adequada. S'adjunten de nou les dades en l'arxiu **Pràctica3.xlsx**. A partir de la informació obtinguda amb *R-Commander*, calcula els estadístics (mesures) que falten. Què succeeix amb la moda?

PRÀCTICA 4: INFERÈNCIA EN UNA POBLACIÓ. Part I

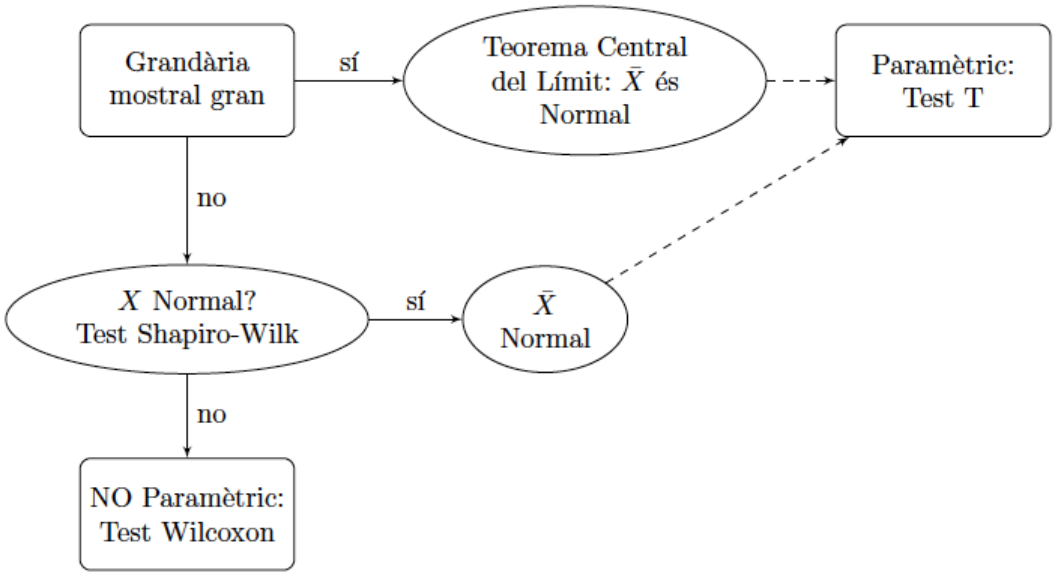
L'anàlisi bàsica d'una mostra, des del punt de vista de la inferència estadística, comporta l'obtenció d'intervals de confiança i/o la resolució de contrastos d'hipòtesis, tots dos referents a la mitjana de la població de què prové la mostra.

Els objectius d'aquesta pràctica són:

- Identificació dels objectius de l'estudi: inferència sobre una mitjana poblacional.
- Comprovació de les condicions d'aplicabilitat del test T (normalitat).
- Elecció de les proves adequades per a analitzar les dades.
- Obtenció d'estimacions puntuals, per interval i/o resolució del contrast d'hipòtesis.
- Interpretació de resultats.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

COMPROVACIÓ DE LES CONDICIONS D'APLICABILITAT DEL TEST T I ELECCIÓ DE LES PROVES ADEQUADES PER ANALITZAR LES DADES



Comprovació de la normalitat. Si la mostra és prou gran (n≥30), pel teorema central del límit, la mitjana mostral es comporta com si procedira d'una distribució normal (i podem aplicar el test T). Si la grandària de la mostra és petita, hem de fer servir proves de normalitat.

Estadístics/Resums/Test de normalitat

En la finestra que apareix hem de triar la variable que volem analitzar i el test que volem usar (en el nostre cas usarem per defecte el **test de Shapiro-Wilk**). Recordem el contrast d'hipòtesis de la normalitat:

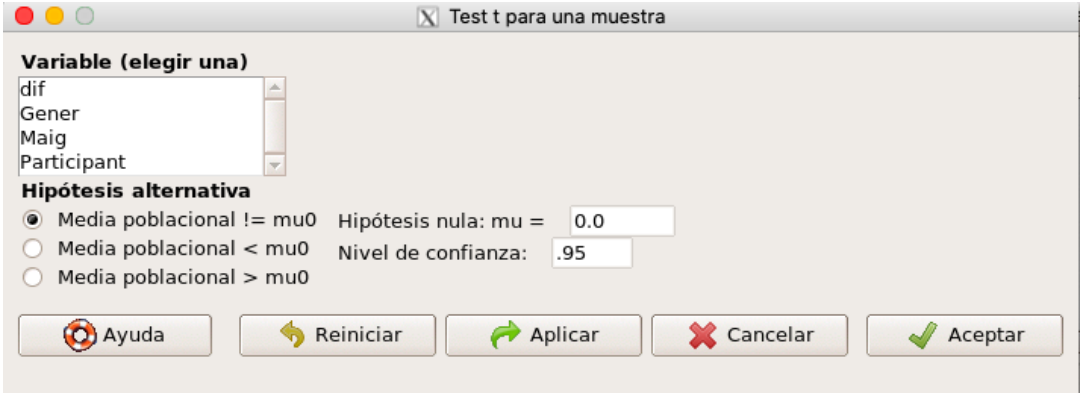
$$\begin{cases} H_0: \text{La distribució de la variable és normal} \\ H_A: \text{La distribució de la variable no és normal} \end{cases}$$

Per tant, fixat un nivell de significació α, resolem el contrast:

- Si p-valor $< \alpha$, aleshores rebutgem la normalitat (H_0).
- Si p-valor $\geq \alpha$, aleshores NO rebutgem la normalitat.

Test paramètric: test T per a una mostra. Obtenció d'estimacions puntuals i per interval de la mitjana poblacional. Resolució de contrastos d'hipòtesis. Apliquem aquest test quan es compleix la condició de normalitat de la mitjana mostral.

Estadístics/Mitjanes/Test t per a una mostra



Expliquem a continuació què significa cada opció en la imatge anterior:

- ***Variable***. Dades que volem analitzar.
- ***Hipòtesi alternativa.*** Condició en la hipòtesi alternativa d'un contrast d'hipòtesis, en què *mu0* és el valor que es desitja contrastar: "Mitjana poblacional != mu0" (mitjana poblacional distinta de mu0), < (menor) i > (major).
- ***Hipòtesi nul·la: mu =.*** Valor que es desitja contrastar.
- ***Nivell de confiança.*** Nivell de confiança triat per a construir l'interval de confiança.

Interpretació dels resultats.

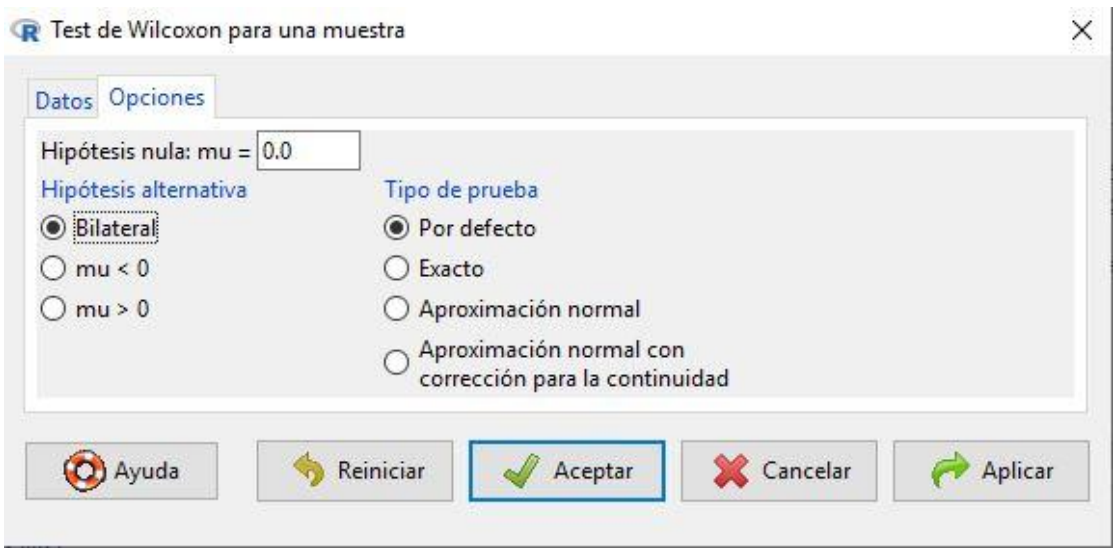
```
> with(Dataset, (t.test(Temps, alternative = "two.sided", mu = 1.3, conf.level = 0.95)))

        One Sample t-test

data:  Temps
t = 3.6283, df = 11, p-value = 0.003969
alternative hypothesis: true mean is not equal to 1.3
95 percent confidence interval:
 1.401623 1.715044
sample estimates:
mean of x
 1.558333
```

En aquest cas, hem aplicat un test t de Student (t.test) en què la variable és Temps, la hipòtesi alternativa és $\mu \neq 1{,}3$ (és a dir, el valor que estem contrastant és 1,3) i el nivell de confiança és 95% (per si demanaren un interval de confiança). Recorda que per a fer un interval de confiança hem de posar-ho sempre. En blau tenim els resultats:

- p-value = 0,003969. Com que el p-valor és menor que 0,05, rebutgem H_0. És a dir, la mitjana poblacional no és igual 1,3, a un nivell de significació 0,05.
- Interval de confiança al 95%: [1,401623; 1,715044].
- Estimació puntual de la mitjana poblacional (és a dir, mitjana de la mostra): 1,55833.

Test no paramètric: test de Wilcoxon. Obtenció d'estimacions puntuals i per interval de la mediana poblacional. Resolució de contrastos d'hipòtesis sobre la mediana. Apliquem aquest test quan NO es compleix la condició de normalitat de la mitjana mostral.

Estadístics/Tests no paramètrics/Test de Wilcoxon per a una mostra



En el menú **opcions** podem triar la hipòtesi nul·la i l'alternativa. També podem elegir el tipus de prova (que en el nostre cas deixarem per defecte).

Interpretació dels resultats

```
> with(Exercici5, wilcox.test(Edat, alternative = "two.sided", mu = 0))

        Wilcoxon signed rank test with continuity correction

data:  Edat
V = 45, p-value = 0.009091
alternative hypothesis: true location is not equal to 0
```

En aquest cas, hem aplicat un test de Wilcoxon (wilcox.test) en què la variable és Edat, la hipòtesi alternativa és $\mu \neq 0$ i el valor que es desitja contrastar és 0. En blau tenim el p-valor=0,009091. Com que és menor que 0,05, rebutgem H_0, és a dir, la mediana poblacional no és igual a 0 a un nivell de significació 0,05.

EXERCICI 1

Durant la reintroducció de les tortugues de terra *Testudo Hermanni* al delta de l'Ebre, es va tractar de determinar els rangs de valors de determinades característiques de les tortugues sanes. Com que, a més d'aquest estudi, es va analitzar la concentració de sodi en sang de 120 tortugues sanes trobades algun temps després d'haver-les deixades anar al delta (les dades es troben en l'arxiu **Practica4.xlsx**).

a) Identifica els elements (població, mostra i variable).
b) Indica el mètode inferencial (paramètric o no paramètric) que resulta adequat per analitzar aquesta mostra. Justifica la teua elecció.
c) Construeix intervals de confiança al 90%, 95% i 99%. Compara'n els límits i interpreta la informació que proporcionen.
d) Per a completar l'estudi, analitza si es pot confirmar la teoria dels investigadors que les tortugues en llibertat tenen una concentració de sodi en sang menor que 130 mEq/L.

EXERCICI 2

S'està provant una nova anestèsia amb la qual es pretén reduir el temps de permanència a la clínica veterinària després de l'operació de gossos de raça petita. És a dir, es pretén reduir el temps que tarden els efectes residuals de l'anestèsia a desaparèixer. Amb les anestèsies utilitzades fins ara, els gossos podien deixar la clínica veterinària en 6 hores com a mínim, amb una permanència mitjana de 6,3 hores. Es considera que la nova anestèsia és millor si aconsegueix que els gossos deixen la clínica en menys de 5 hores. La nova anestèsia s'ha provat en 17 operacions i s'han observaren els temps següents (en hores):

5	6	10	3	2	7	3	3	3
3	3	4	4	5	8	4	2	

Amb el propòsit de donar alguna conclusió sobre l'eficàcia de la nova anestèsia, és necessari fer l'anàlisi estadística completa d'aquestes dades.

a) Dibuixa l'histograma de les dades. Què pot afirmar-se sobre la simetria de la distribució de les dades?
b) Podem suposar que les dades segueixen una distribució normal? Et sembla adequat utilitzar mètodes estadístics basats en la t de Student?
c) Planteja i resol (per α =0,05) un contrast d'hipòtesis adequat per a valorar l'efectivitat de la nova anestèsia.

EXERCICI 3

La ß-endorfina humana (BEH) és una hormona segregada per la glàndula pituïtària sota condicions d'estrès. Un investigador va realitzar un estudi per a investigar si es podia establir un límit per al nivell de BEH en la sang d'una persona sense estrès. Va mesurar els nivells de BEH en sang de 10 participants al gener. Els resultats es mostren a continuació:

42	47	37	9	33	70	54	27	41	18

L'arxiu **Practica4.xlsx** conté les dades de nivell de BEH al gener. Respon a les qüestions següents:

a) Indica el mètode inferencial que resulta adequat per analitzar aquesta mostra. Justifica la teua elecció.
b) Construeix un interval de confiança al 95% per al nivell mitjà poblacional de BEH al gener.
c) Planteja i resol el contrast d'hipòtesi corresponent per a contestar la pregunta següent: hi ha evidència que els nivells de BEH són menors que 60?

EXERCICI 4

L'arxiu **Practica4.xlsx** conté les dades dels nivells de glucosa en sang per a 100 xiquets en dejú. Indica el mètode inferencial que resulta adequat per analitzar aquesta mostra. Justifica la teua elecció.

a) Podem assumir que les dades segueixen una distribució normal a un nivell de significació 0,05? I a un nivell 0,01? Et sembla adequat utilitzar els mètodes estadístics basats en la t de Student?
b) Calcula una estimació puntual i un interval de confiança al 99% per al nivell mitjà de glucosa, només si està justificat fer-ho. En cas contrari, calcula l'interval de confiança per a la mediana del nivell de glucosa.
c) Planteja i resol (per α =0,05) un contrast d'hipòtesis adequat per a valorar si el nivell mitjà de glucosa és significativament diferent de 65. Canviaria la conclusió si α =0,01?

EXERCICI 5

Les dades següents corresponen als nivells de fosfoquinasa creatina en sang de 36 homes sans, ordenades de menor a major:

25	42	48	57	58	60	62	64	67	68	70	78
82	83	84	92	93	94	95	95	100	101	104	110
110	113	118	119	121	123	139	145	151	163	201	203

a) Podem suposar que les dades segueixen una distribució normal? Et sembla adequat utilitzar-hi els mètodes estadístics basats en la t de Student?
b) Obté una estimació puntual i un interval de confiança al 95% per al nivell mitjà de fosfoquinasa creatina, només si està justificat fer-ho. En cas contrari, calcula l'interval de confiança per a la mediana de fosfoquinasa creatina.
c) Planteja i resol (per α =0,05) un contrast d'hipòtesis adequat per a valorar si el nivell mitjà de fosfoquinasa creatina és inferior a 100.

EXERCICI 6

La cataracta nuclear, com indica el seu nom, és l'opacificació del nucli del cristal·lí que després avança fins a la totalitat del cristal·lí. Però, aquest progrés d'opacificació és lent, generalment comença cap als 55 anys i arriba a la màxima expressió al voltant dels 70 anys. S'ha provat un nou fàrmac en 9 malalts amb cataracta nuclear per comprovar si retarda l'edat a què es produeix la màxima opacitat del cristal·lí. Els resultats són:

69	74	75	70	72	73	71	73

Fes una anàlisi completa del problema estadístic que s'hauria de plantejar i indica la decisió que caldria prendre pel que fa al comportament d'aquest nou fàrmac.

PRÀCTICA 5: INFERÈNCIA EN UNA POBLACIÓ. Part II

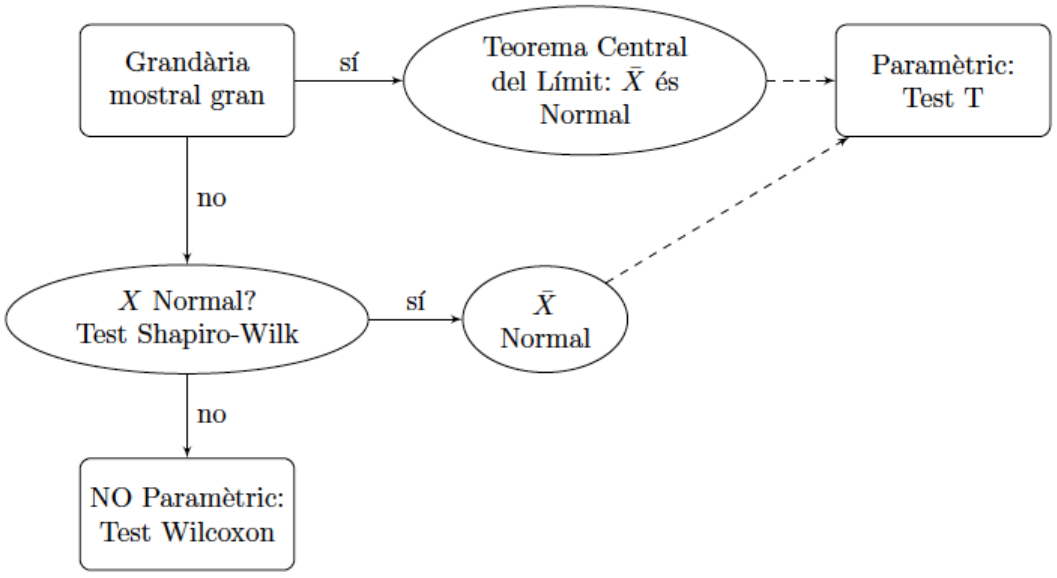
L'anàlisi bàsica d'una mostra, des del punt de vista de la inferència estadística, comporta l'obtenció d'intervals de confiança i/o la resolució de contrastos d'hipòtesis, els dos referents a la mitjana de la població de què prové la mostra.

Els objectius d'aquesta pràctica són:

- Identificació dels objectius de l'estudi: inferència sobre una mitjana poblacional.
- Comprovació de les condicions d'aplicabilitat del test T (normalitat).
- Elecció de les proves adequades per a analitzar les dades.
- Obtenció d'estimacions puntuals, per interval i/o resolució del contrast d'hipòtesis.
- Interpretació de resultats.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

COMPROVACIÓ DE LES CONDICIONS D'APLICABILITAT DEL TEST T I ELECCIÓ DE LES PROVES ADEQUADES PER A ANALITZAR LES DADES



Comprovació de la normalitat. Si la mostra és suficientment gran (n≥30), pel teorema central del límit, la mitjana mostral es comporta com si procedira d'una distribució normal (i podem aplicar-hi el test T). Si la grandària de la mostra és menuda, hem de fer servir proves de normalitat.

Estadístics/Resums/Test de Normalitat

En la finestra que apareix hem de triar la variable a analitzar i el test que volem usar (en el nostre cas usarem per defecte el **test de Shapiro-Wilk**). Recordem el contrast d'hipòtesis de la normalitat:

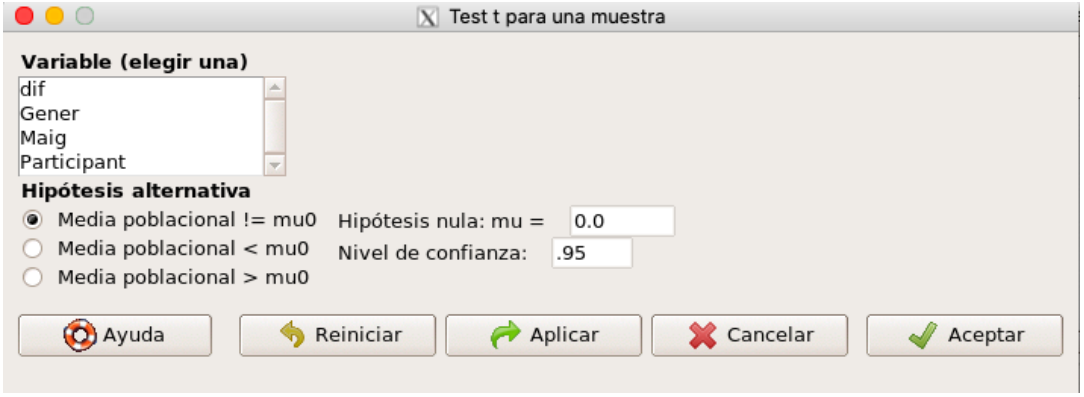
$$\begin{cases} H_0: \text{La distribució de la variable és normal} \\ H_A: \text{La distribució de la variable no és normal} \end{cases}$$

Per tant, fixat un nivell de significació α, resolem el contrast:

- Si el valor p $< \alpha$, aleshores rebutgem la normalitat (H_0).
- Si el valor p $\geq \alpha$, aleshores NO rebutgem la normalitat.

Test paramètric: test T per a una mostra. Obtenció d'estimacions puntuals i per interval de la mitjana poblacional. Resolució de contrastos d'hipòtesis. Apliquem aquest test quan es compleix la condició de normalitat de la mitjana mostral.

Estadístics/Mitjanes/Test t per a una mostra



Expliquem a continuació què significa cada opció en la imatge anterior:

- ***Variable***. Dades que volem analitzar.
- ***Hipòtesi alternativa.*** Condició en la hipòtesi alternativa d'un contrast d'hipòtesis, en què *mu0* és el valor que es desitja contrastar: "Mitjana poblacional != mu0" (mitjana poblacional distinta de mu0), < (Menor) i > (major).
- ***Hipòtesi nul·la: mu =.*** Valor que es desitja contrastar.
- ***Nivell de confiança.*** Nivell de confiança triat per a construir l'interval de confiança.

Interpretació dels resultats.

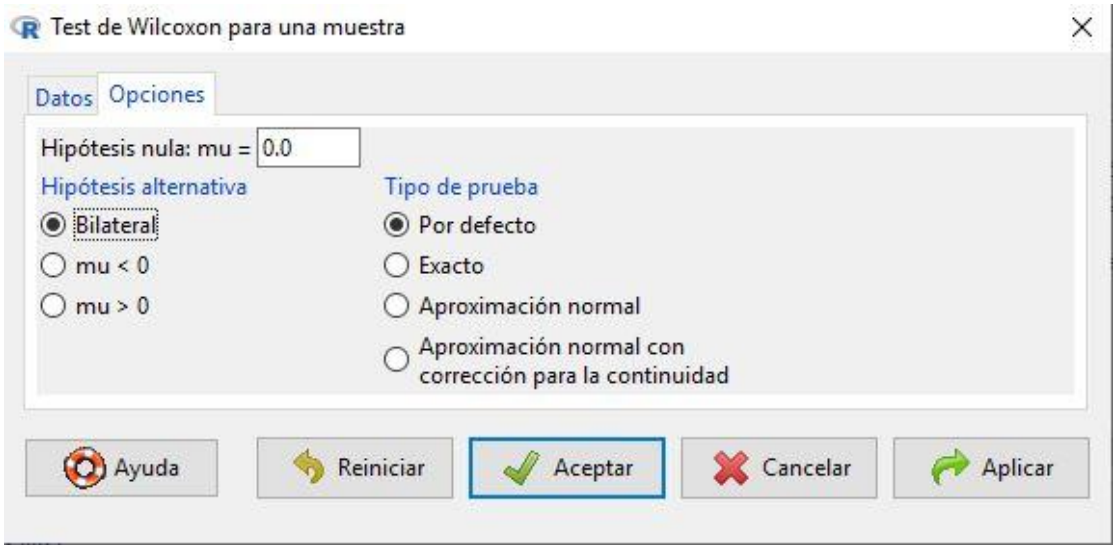
```
> with(Dataset, (t.test(Temps, alternative = "two.sided", mu = 1.3, conf.level = 0.95)))

        One Sample t-test

data:  Temps
t = 3.6283, df = 11, p-value = 0.003969
alternative hypothesis: true mean is not equal to 1.3
95 percent confidence interval:
 1.401623 1.715044
sample estimates:
mean of x
 1.558333
```

En aquest cas hem aplicat un test t de Student (t.test) on la variable és Temps, la hipòtesi alternativa és $\mu \neq 1{,}3$ (és a dir, el valor que estem contrastant és 1,3) i el nivell de confiança 95% (per si demanaren un interval de confiança). Recorda que per a fer un interval de confiança hem de posar-ho sempre. En blau tenim els resultats:

- p-value = 0,003969. Com que el valor p és menor que 0,05, rebutgem H_0. És a dir, la mitjana poblacional no és igual 1,3, a un nivell de significació 0,05.
- Interval de confiança al 95%: [1,401623; 1,715044].
- Estimació puntual de la mitjana poblacional (és a dir, mitjana de la mostra): 1,55833.

Test no paramètric: test de Wilcoxon. Obtenció d'estimacions puntuals i per interval de la mediana poblacional. Resolució de contrastos d'hipòtesis sobre la mediana. Apliquem aquest test quan NO es compleix la condició de normalitat de la mitjana mostral.

Estadístics/Tests no paramètrics/Test de Wilcoxon per a una mostra



En el menú **opcions** podem triar la hipòtesi nul·la i l'alternativa. També podem elegir el tipus de prova (que en el nostre cas deixarem per defecte).

Interpretació dels resultats.

```
> with(Exercici5, wilcox.test(Edat, alternative = "two.sided", mu = 0))

        Wilcoxon signed rank test with continuity correction

data:  Edat
V = 45, p-value = 0.009091
alternative hypothesis: true location is not equal to 0
```

En aquest cas hem aplicat un test de Wilcoxon (wilcox.test) la variable és Edat, la hipòtesi alternativa és $\mu \neq 0$ i el valor que es desitja contrastar és 0. En blau tenim el valor p=0,009091. Com que és menor que 0,05 rebutgem H_0, és a dir, la mediana poblacional no és igual a 0 a un nivell de significació 0,05.

PROBLEMA

Continuem en aquesta pràctica amb l'arxiu **datwinsdef_lab.xlsx**, que consta de 108 registres ficticis, cadascun representant un embaràs gemel·lar.

Dins del projecte "Análisis de los efectos del medio ambiente sobre la salud y el desarrollo del feto en embarazos gemelares" es pretén estudiar la relació pes-sexe dels bessons monozigòtics, així com les diferències de pes existents en els bebès nascuts prematurs segons el sexe.

Parlem de bebès prematurs segons el criteri de la seua edat gestacional. La duració mitjana habitual per a un embaràs de bessons és de 37 setmanes. Per tant, si els bessons naixen abans de la setmana 37 són prematurs.

Les variables que intervenen en aquest estudi són:

- **sexe**, amb dues opcions {xiquet, xiqueta}.
- **pes1,** pes en nàixer del bessó amb major pes.
- **tallam,** alçada de les embarassades.
- **sges,** setmanes de gestació.

A continuació, respon a les preguntes següents.

a) Calcula i interpreta els estadístics mostrals que consideres oportuns per a comparar el pes del bessó amb major pes en xiquetes i xiquets (*recorda que es pot calcular el resum numèric per grups*). Fes-ne una representació gràfica i compara, una altra vegada, els dos grups (xiquetes i xiquets).

b) Hi ha evidència que l'alçada mitjana de totes les embarassades amb bessons monozigòtics és igual a 165 centímetres?

   Per a contestar a la pregunta has de fer el que segueix:

   - És normal la variable **tallam**?
   - Planteja i resol el contrast d'hipòtesis corresponent.

c) Hi ha evidència que, en el cas de les **xiquetes**, el pes mitjà del bessó amb major pes (**pes1**) no està, en general, per davall dels 2100 grams ni per damunt dels 2700 grams?

Ajuda: Noteu que ens demanen treballar només amb les xiquetes, aleshores la primera cosa que heu de fer és filtrar la base de dades i crear un subconjunt de dades XIQUETES, on estarà la informació dels embarassos de les que han nascut xiquetes.

Dades/Conjunt de dades actiu/Filtrar el conjunt de dades actiu (i posar en l'expressió de selecció: **sexe==”xiqueta”**)

Per a contestar a la pregunta has de fer el que segueix:

- És normal la variable pes1 per a les xiquetes?
- Calcula i interpreta l'interval de confiança corresponent.

d) D'acord amb les dades de l'estudi, es pot afirmar que, en embarassos **prematurs de xiquetes**, el pes en nàixer del bessó amb major pes no arriba als dos quilos, amb un nivell de confiança del 95%?
*Ajuda: Noteu que ens demanen treballar només amb les xiquetes prematures. Per tant, treballarem amb la base de dades que hem anomenat XIQUETES. Ara, ens interessa treballar amb les xiquetes prematures. Però no disposem d'una variable categòrica que indique si l'embaràs és prematur o no ho és; el que tenim és una variable numèrica **sges** que indica les setmanes de gestació de l'embaràs. Per tant, de primer heu de crear una nova variable categòrica PREMATUR recodificant la variable contínua sges.*

Dades/Modificar variables del conjunt de dades actiu/Recodificar variables

Una vegada creada la nova variable, haurem de filtrar en la base de dades XIQUETES i crear un subconjunt XIQUETES_PREMATURES.

Per a contestar a la pregunta has de fer el que segueix:

- És normal la variable pes1 per a les xiquetes prematures?
- Planteja i resol el contrast d'hipòtesis corresponent.

e) Per a les **xiquetes no prematures**, els investigadors assumeixen que el pes en nàixer, dels bessons que més pesen, és superior a dos quilos i mig. Tenen raó, els especialistes, amb un nivell de significació 0,01? I a un nivell 0,05?

Ajuda: Ens pregunten per les xiquetes no prematures, per tant, per a filtrar i treballar amb les no prematures cal tornar al conjunt de dades XIQUETES.

Per a contestar a la pregunta has de fer el que segueix:

- És normal la variable pes1 per a les xiquetes no prematures?
- Planteja i resol el contrast d'hipòtesis corresponent.

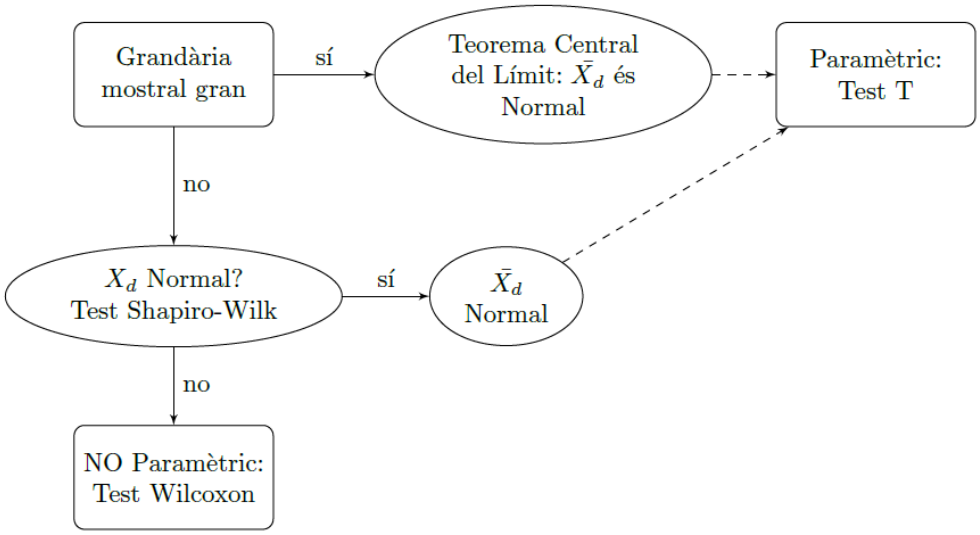
PRÀCTICA 6: ANÀLISI ESTADÍSTICA DE DUES MOSTRES. Part I

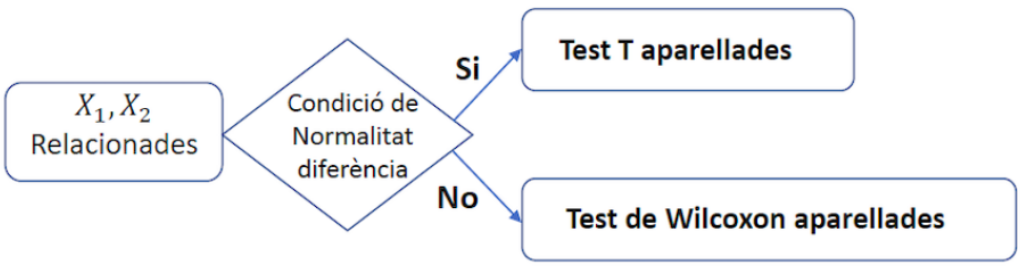
En aquesta pràctica analitzarem i compararem dues mostres, tant relacionades com independents. Els objectius són:

- Identificació dels objectius de l'estudi. Mostres relacionades o independents.
- Comprovació de les condicions d'aplicabilitat del test T (normalitat).
- Elecció de les proves adequades per a analitzar les dades.
- Obtenció d'estimacions puntuals, per interval i/o resolució de contrastos d'hipòtesis sobre la diferència de mitjanes.
- Interpretació de resultats.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

COMPROVACIÓ DE LES CONDICIONS D'APLICABILITAT DEL TEST T I ELECCIÓ DE LES PROVES ADEQUADES PER A ANALITZAR LES DADES. MOSTRES RELACIONADES





Comparació de dues mostres relacionades amb *R-Commander*

Hi ha dues maneres de treballar amb mostres relacionades, en *R-Commander,* **quan fem comparacions amb el valor 0.**

La primera manera consisteix a definir la variable diferència i treballar-hi com si fora una única mostra.

$$\begin{cases} H_0: \mu_d = 0 \\ H_A: \mu_d \neq 0 \end{cases}$$

Per tant, utilitzem les opcions del menú següents:

Estadístics/Resums/Test de normalitat (per a la diferència)

Estadístics/Mitjanes/Test t per a una mostra

Estadístics/Test no paramètrics/Test Wilcoxon per a una mostra

La segona manera consisteix a utilitzar test, definits per a dues mostres, indicant clarament que les mostres estan relacionades.

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_A: \mu_1 \neq \mu_2 \end{cases}$$

Per tant, utilitzem les opcions del menú següents:

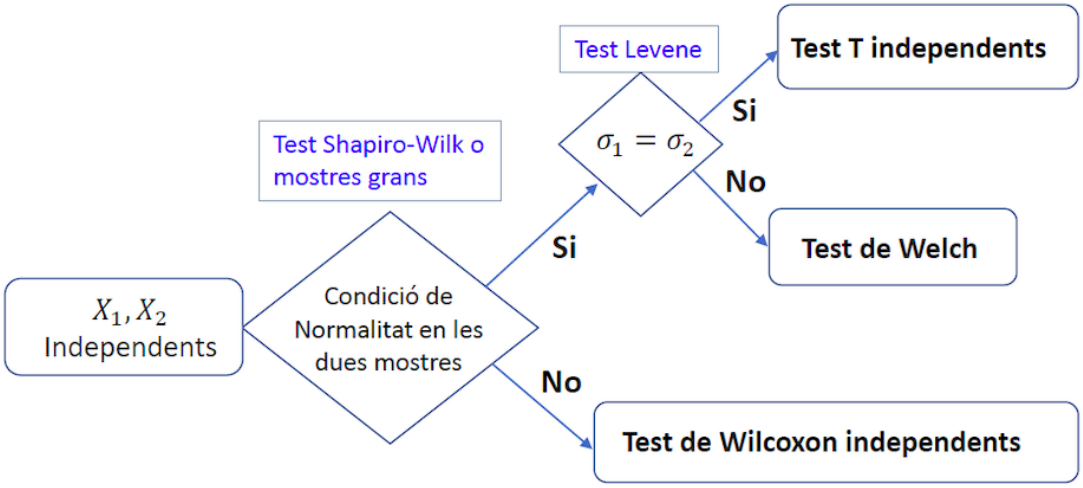
Estadístics/Resums/Test de normalitat (per a la diferència)

Estadístics/Mitjanes/Test t per a dades relacionades

Estadístics/Test no paramètrics/Test Wilcoxon per a mostres aparellades

Nota: La segona manera únicament la utilitzem quan volem fer un contrast de la diferència amb el valor zero; en altres casos sempre utilitzarem directament la variable diferència.

COMPROVACIÓ DE LES CONDICIONS D'APLICABILITAT DEL TEST T I ELECCIÓ DE LES PROVES ADEQUADES PER A ANALITZAR LES DADES. MOSTRES INDEPENDENTS



Quan tenim dues mostres independents, la normalitat hem d'estudiar-la separadament en cadascuna de les mostres. És a dir, per a aplicar els tests paramètrics hem de comprovar la normalitat de la variable mitjana mostral per a cada una de les dues distribucions. Recordem que podem considerar la distribució de la mitjana normal si la grandària de la mostra és suficientment gran o si la distribució de la mostra és normal (test de Shapiro-Wilk).

Test de Levene

Si es compleixen les condicions de normalitat en ambdues mostres, hem de resoldre el test de Levene per a la igualtat de variàncies.

Estadístics/Variàncies/Test de Levene (seleccionant centre: mitjana)

Recordem el contrast d'hipòtesis:

$$\begin{cases} H_0: \sigma_1 = \sigma_2 \\ H_A: \sigma_1 \neq \sigma_2 \end{cases}$$

Per tant, fixat un nivell de significació α, resolem el contrast:

- Si el valor p < α, aleshores rebutgem H_0. I, per tant, assumim que ambdues distribucions tenen variàncies distintes.
- Si el valor p ≥ α, aleshores no rebutgem H_0. I considerem igualtat de variàncies.

Test paramètric: test T per a mostres independents (cas $\sigma_1 = \sigma_2$)

En aquest cas, per al càlcul de l'interval de confiança i del valor p s'utilitza l'error de mostreig combinat.

Estadístics/Mitjanes/Test t per a mostres independents

Dins de la finestra, en el menú opcions, heu de seleccionar igualtat de variàncies.

```
> t.test(LHC ~ Sexe, alternative = "two.sided", conf.level = 0.95,
+  var.equal = TRUE, data = Dataset)

        Two Sample t-test

data:  LHC by Sexe
t = -2.6848, df = 92, p-value = 0.008609
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.2844115 -0.9400299
sample estimates:
mean in group Femella  mean in group Mascle
             67.51569              71.12791
```

Fixeu-vos que apareix **var.equal=TRUE**, la qual cosa significa que les variàncies són iguals. En fer aquest test obtenim la informació següent:

- Valor p
- Interval de confiança per a la diferència de mitjanes (Femella – Mascle).
- Estimacions puntuals de les mitjanes d'ambdues poblacions.

Test paramètric: test T per a mostres independents (cas $\sigma_1 \neq \sigma_2$)

En aquest cas, per al càlcul de l'interval de confiança i del valor p s'utilitza l'error de mostreig no combinat.

Estadístics/Mitjanes/Test t per a mostres independents

Dins de la finestra, en el menú opcions, heu de seleccionar diferència de variàncies.

```
> t.test(LHC ~ Sexe, alternative = "two.sided", conf.level = 0.95,
+  var.equal = FALSE, data = Dataset)

        Welch Two Sample t-test

data:  LHC by Sexe
t = -2.6734, df = 87.696, p-value = 0.008954
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.2975008 -0.9269406
sample estimates:
mean in group Femella  mean in group Mascle
             67.51569              71.12791
```

Fixeu-vos que apareix **var.equal=FALSE**, la qual cosa significa que les variàncies són diferents. També apareix la paraula **Welch**, que indica el test paramètric considerat. En fer aquest test obtenim la informació següent:

- Valor p
- Interval de confiança per a la diferència de mitjanes.
- Estimacions puntuals de les mitjanes d'ambdues poblacions.

Test no paramètric: test de Wilcoxon per a mostres independents

En aquest cas falla la normalitat en algun dels dos casos o en tots dos. Per tant, hi apliquem un test no paramètric.

Estadístics/Test no paramètrics/Test de Wilcoxon per a dues mostres

```
> wilcox.test(LHC ~ Sexe, alternative = "two.sided", data = Dataset)

        Wilcoxon rank sum test with continuity correction

data:  LHC by Sexe
W = 760, p-value = 0.01077
alternative hypothesis: true location shift is not equal to 0
```

A més, amb el test de Wilcoxon podem calcular intervals de confiança per a la **diferència de medianes**. Per això, hem d'afegir **conf.int=TRUE, conf.level=0.95** i tornar a executar.

```
> wilcox.test(LHC ~ Sexe, alternative = "two.sided",conf.int=TRUE,conf.level = 0.98, data = Dataset)

        Wilcoxon rank sum test with continuity correction

data:  LHC by Sexe
W = 760, p-value = 0.01077
alternative hypothesis: true location shift is not equal to 0
98 percent confidence interval:
 -6.9999450 -0.2999777
sample estimates:
difference in location
             -3.799952
```

EXERCICI 1

La β-endorfina humana (BEH) és una hormona secretada per la glàndula pituïtària en condicions d'estrès. Un investigador va realitzar un estudi per a investigar si un programa d'exercici regular podria afectar les concentracions en repòs (sense estrès) de BEH en la sang. Va mesurar els nivells de BEH en sang, al gener i de nou al maig, de 10 participants en un programa d'exercici físic. Els resultats es mostren en la taula següent.

	Nivell de BEH		
Participant	Gener	Maig	Diferència
1	42	22	20
2	47	29	18
3	37	9	28
4	9	9	0
5	33	26	7
6	70	36	34
7	54	38	16
8	27	32	-5
9	41	33	8
10	18	14	4

L'arxiu **Practica6.xlsx** conté les dades de nivell de BEH al gener i al maig. Respon a les qüestions següents:

a) Crea una nova variable **diferència** en el nivell de BEH entre gener i maig.
Ajuda: En la pràctica 1 s'explica com calcular una nova variable.

Dades/Modificar variables del conjunt de dades actiu/Calcular una nova variable

b) Indica el mètode inferencial que resulta adequat per a analitzar aquesta mostra. Justifica l'elecció.
c) Construeix un interval de confiança al 95% per a la diferència de mitjanes poblacionals de nivells de BEH entre gener i maig. Interpreta el resultat.
d) A partir de l'interval de confiança de l'apartat c), hi ha evidència que els nivells de BEH són menors al maig que al gener?
e) Planteja i resol el contrast d'hipòtesis corresponent per a contestar a la pregunta següent: Hi ha evidència del fet que els nivells de BEH són menors al maig que al gener? Fes l'estudi de les dues formes indicades en la introducció, amb un contrast per a la diferència i amb un contrast amb el test de mostres aparellades. Què se'n pot concloure?

EXERCICI 2

En un estudi sobre l'impacte ecològic de la infecció del paràsit de malària *Plasmodium* en fardatxos que no estaven en captivitat, es va investigar, entre altres coses, si la infecció disminuïa la resposta a una prova de resistència. Es va mesurar, per a 15 fardatxos infectats i 15 no infectats, la distància (en metres) que cada animal recorria en dos minuts (les dades es troben en l'arxiu **Practica6.xlsx**):

a) Identifica els elements de l'estudi (població, mostra i variables).
b) Descriu gràficament i numèricament, per grups, les dades de l'estudi. Quines diferències hi veus?
c) Analitza si els resultats obtinguts confirmen, a nivell α = 0,05, la teoria dels investigadors i justifica l'elecció del mètode inferencial emprat. Canvia la conclusió a nivell α = 0,01?

EXERCICI 3

Durant un estudi morfomètric sobre l'àguila calçada comuna (*Hieraaetus pennatus*) a la península Ibèrica, es va mesurar la longitud (mil·límetres) dels dits posterior i anterior mitjans **de cada au** per comprovar si la diferència mitjana en longitud (posterior menys anterior) era superior a 3 mm, valor usual en altres localitzacions. L'arxiu **Practica6.xlsx** conté les longituds anteriors i posteriors de cada individu. Respon a les qüestions següents:

a) Identifica els elements de l'estudi (població, mostra i variables).
b) Descriu gràficament i numèricament les dades que permeten observar millor l'objectiu de l'estudi.
c) Mitjançant les dades i la prova estadística que consideres més adequada, analitza si es pot confirmar, a nivell α = 0,05, la teoria dels investigadors.

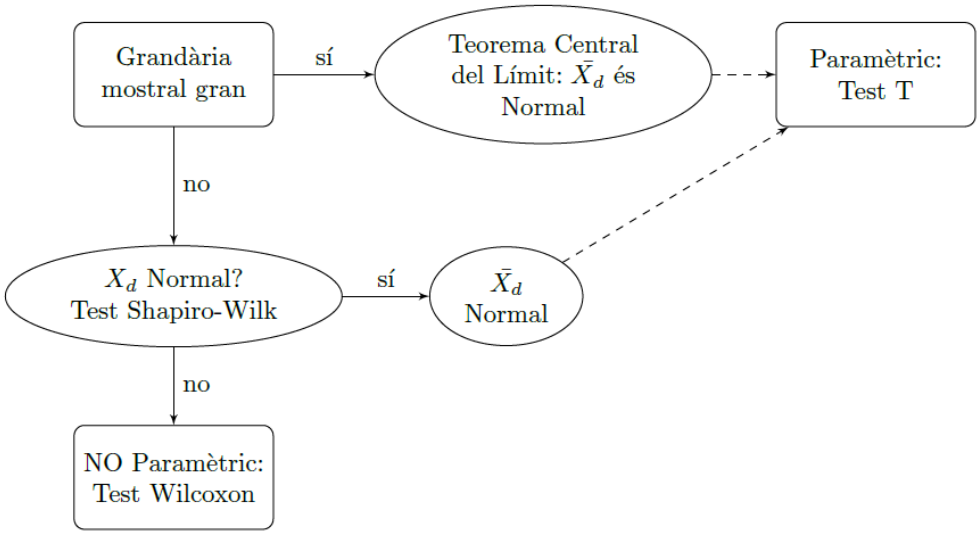
PRÀCTICA 7: ANÀLISI ESTADÍSTICA DE DUES MOSTRES. Part II.

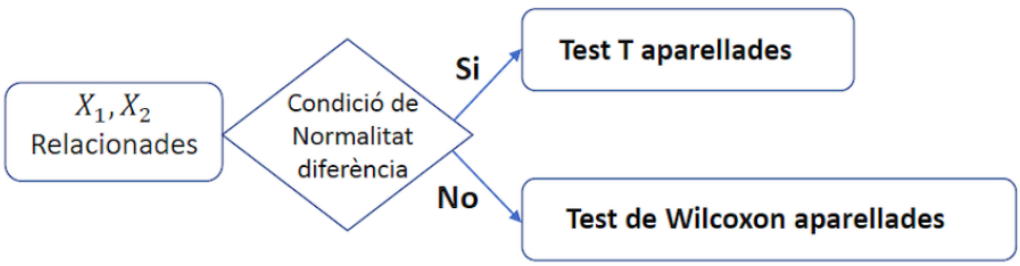
En aquesta pràctica analitzarem i compararem dues mostres, tant relacionades com independents. Els objectius són:

- Identificació dels objectius de l'estudi. Mostres relacionades o independents.
- Comprovació de les condicions d'aplicabilitat del test T (normalitat).
- Elecció de les proves adequades per a analitzar les dades.
- Obtenció d'estimacions puntuals per interval i/o resolució de contrast d'hipòtesis sobre la diferència de mitjanes.
- Interpretació de resultats.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

COMPROVACIÓ DE LES CONDICIONS D'APLICABILITAT DEL TEST T I ELECCIÓ DE LES PROVES ADEQUADES PER A ANALITZAR LES DADES. MOSTRES RELACIONADES





Comparació de dues mostres relacionades amb *R-Commander*

Hi ha dues formes de treballar amb mostres relacionades en *R-Commander* **quan fem comparacions amb el valor 0.**

La primera forma consisteix a definir la variable diferència i treballar amb ella com si fóra una única mostra.

$$\begin{cases} H_0: \mu_d = 0 \\ H_A: \mu_d \neq 0 \end{cases}$$

Per tant, utilitzem les opcions següents del menú:

Estadístics / resums / test de normalitat (per a la diferència)

Estadístics / mitjanes / test t per a una mostra

Estadístics / tests no paramètrics / test Wilcoxon per a una mostra

La segona forma consisteix a utilitzar tests definits per a dues mostres, amb la indicació clara que les mostres estan relacionades.

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_A: \mu_1 \neq \mu_2 \end{cases}$$

Per tant, utilitzem les opcions següents del menú:

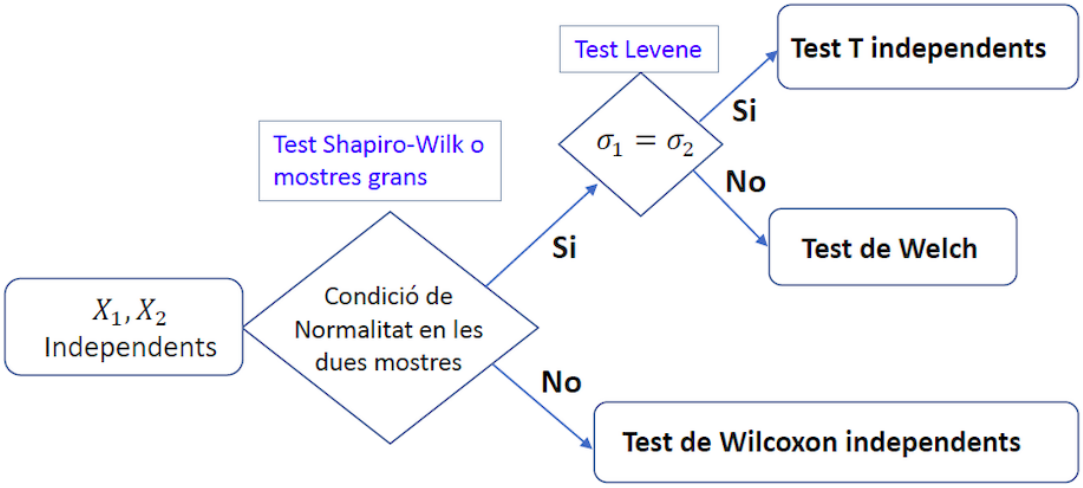
Estadístics / resums / test de normalitat (per a la diferència)

Estadístics / mitjanes / test t per a dades relacionades

Estadístics / test no paramètrics / test Wilcoxon per a mostres aparionades

Nota. Únicament utilitzarem la segona manera quan vulguem fer un contrast de la diferència amb el valor zero, en els altres casos utilitzarem sempre directament la variable diferència.

COMPROVACIÓ DE LES CONDICIONS D'APLICABILITAT DEL TEST T I ELECCIÓ DE LES PROVES ADEQUADES PER A ANALITZAR LES DADES. MOSTRES INDEPENDENTS



Quan tenim dues mostres independents hem d'estudiar la normalitat per separat en cadascuna de les mostres. És a dir, per aplicar els tests paramètrics hem de comprovar la normalitat de la variable mitjana mostral per a cada una de les dues distribucions. Recordem que podem considerar la distribució de la mitjana normal si la grandària de la mostra és suficientment gran o si la distribució de la mostra és normal (test de Shapiro-Wilk).

Test de Levene

Si es compleixen les condicions de normalitat en ambdues mostres, hem de resoldre el test de Levene per a la igualtat de variàncies.

Estadístics / variàncies / test de Levene (seleccionant centre: mitjana)

Recordem el contrast d'hipòtesis:

$$\begin{cases} H_0: \sigma_1 = \sigma_2 \\ H_A: \sigma_1 \neq \sigma_2 \end{cases}$$

Per tant, fixat un nivell de significació α, resolem el contrast:

- Si p-valor < α, aleshores rebutgem H_0. I, per tant, considerem que ambdues distribucions tenen variàncies distintes.
- Si p-valor ≥ α, aleshores no rebutgem H_0. I, considerem que hi ha igualtat de variàncies.

Test paramètric: test T per a mostres independents (cas $\sigma_1 = \sigma_2$).

En aquest cas per al càlcul de l'interval de confiança i del p-valor s'utilitza l'error de mostreig combinat.

Estadístics / mitjanes / test t per a mostres independents

Dins de la finestra, en el menú opcions, hem de seleccionar igualtat de variàncies.

```
> t.test(LHC ~ Sexe, alternative = "two.sided", conf.level = 0.95,
+  var.equal = TRUE, data = Dataset)

        Two Sample t-test

data:  LHC by Sexe
t = -2.6848, df = 92, p-value = 0.008609
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.2844115 -0.9400299
sample estimates:
mean in group Femella  mean in group Mascle
             67.51569              71.12791
```

Fixat que apareix **var.equal=TRUE**, el que significa que les variàncies són iguals. En fer aquest test obtenim la informació següent:

- p-valor
- Interval de confiança per a la diferència de mitjanes (femella - mascle).
- Estimacions puntuals de les mitjanes d'ambdues poblacions.

Test paramètric: test T per a mostres independents (cas $\sigma_1 \neq \sigma_2$).

En aquest cas, per al càlcul de l'interval de confiança i del p-valor s'utilitza l'error de mostreig no combinat.

Estadístics / mitjanes / test t per a mostres independents

Dins de la finestra, en el menú opcions, has de seleccionar la diferència de variàncies.

```
> t.test(LHC ~ Sexe, alternative = "two.sided", conf.level = 0.95,
+  var.equal = FALSE, data = Dataset)

        Welch Two Sample t-test

data:  LHC by Sexe
t = -2.6734, df = 87.696, p-value = 0.008954
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.2975008 -0.9269406
sample estimates:
mean in group Femella  mean in group Mascle
             67.51569              71.12791
```

Fixat que apareix **var.equal=FALSE**, el que significa que les variàncies són diferents. També apareix la paraula **Welch**, que indica el test paramètric considerat. En fer aquest test obtenim la informació següent:

- p-valor
- Interval de confiança per a la diferència de mitjanes.
- Estimacions puntuals de les mitjanes d'ambdues poblacions.

Test no paramètric: test de Wilcoxon per a mostres independents.

En aquest cas falla la normalitat en algun dels dos casos o en els dos. Per tant, apliquem un test no paramètric.

Estadístics / test no paramètrics / test de Wilcoxon per a dues mostres

```
> wilcox.test(LHC ~ Sexe, alternative = "two.sided", data = Dataset)

        Wilcoxon rank sum test with continuity correction

data:  LHC by Sexe
W = 760, p-value = 0.01077
alternative hypothesis: true location shift is not equal to 0
```

A més, amb el test de Wilcoxon, pots calcular intervals de confiança per a la **diferència de medianes**. Per això, has d'afegir **conf.int=TRUE, conf.level=0.95** i tornar a executar.

```
> wilcox.test(LHC ~ Sexe, alternative = "two.sided",conf.int=TRUE,conf.level = 0.98, data = Dataset)

        Wilcoxon rank sum test with continuity correction

data:  LHC by Sexe
W = 760, p-value = 0.01077
alternative hypothesis: true location shift is not equal to 0
98 percent confidence interval:
 -6.9999450 -0.2999777
sample estimates:
difference in location
             -3.799952
```

EXERCICI 1

Un zoòleg decideix realitzar un estudi per comprovar si realment el pes dels faisans mascles és major que el de les femelles. Per fer-lo selecciona vint faisans joves a l'atzar, dels quals deu són mascles i deu són femelles, i anota els pesos corresponents en grams, amb els resultats següents:

Mascles	1.293	1.380	1.614	1.497	1.340	1.643	1.466	1.627	1.383	1.711
Femelles	1.061	1.065	1.092	1.017	1.021	1.138	1.143	1.094	1.270	1.028

L'arxiu **Practica7.xlsx** conté les dades per a fer aquest exercici.

a) Es pot afirmar amb un nivell de significació de $\alpha = 0{,}01$, que el pes mitjà dels faisans mascles és diferent que el de les femelles? Resol l'exercici amb nivells de significació $\alpha = 0{,}1$ i $\alpha = 0{,}05$.

b) Determina un interval de confiança del 99% per a la diferència dels pesos mitjans dels faisans mascles i femelles.

EXERCICI 2

Un ecologista va estudiar l'hàbitat d'un peix d'un escull marí per comprovar si hi havia proves del fet que la densitat de colonització decreixia quan la distància a la cresta de l'escull augmentava (ja que la forma com les ones trenquen sobre la cresta causa que el menjar tendisca a decréixer en allunyar-se d'aquest lloc). Va fixar 48 parcel·les de colònies, totes de la mateixa dimensió. **Per a cada parcel·la** va calcular la densitat de colonització a 250 i 800 metres de la cresta de l'escull.

L'arxiu **Practica7.xlsx** conté la densitat de colonització a 250 i a 800 metres per a cada parcel·la.

a) Indica el mètode d'inferència que resulta adequat per a analitzar aquesta mostra. Justifica l'elecció.
b) Estableix intervals de confiança del 95% per a la diferència mitjana de densitat de colonització i interpreta la informació que dona.
c) Per a completar l'estudi comprova si els resultats obtinguts confirmen en el nivell $\alpha = 0{,}05$ la teoria dels investigadors.

EXERCICI 3

En un estudi sobre l'efectivitat d'un nou fertilitzant en la producció de blat es van seleccionar cinc camps. **Cada camp** es va dividir en dues parcel·les. En una d'elles es va utilitzar el fertilitzant i en l'altra no. Les quantitats de blat collides van ser les següents (les dades es troben en l'arxiu **Practica7.xlsx**):

Camp	**1**	**2**	**3**	**4**	**5**
Amb fertilitzant	5,5	3,2	3,8	1,9	2,8
Sense	5	2	4	1,3	2

Analitza l'eficàcia del fertilitzant (el fertilitzant és eficaç si la producció de blat augmenta).

- Gràficament i numèricament.
- Interval de confiança del 95%.
- Contrast d'hipòtesis.

EXERCICI 4

Es va fer un estudi per a investigar l'efecte produït pel desballestament d'una zona d'aparcament en la densitat de la vegetació circumdant. Es van estudiar dues àrees. Una era objecte del desballestament d'una gran zona d'aparcament; l'altra no estava prop de cap aparcament i es va utilitzar de control. Cada àrea es va subdividir en una sèrie de panells de dos per vint metres. Es va comptar el nombre de plantes trobades en cadascun i es van obtenir aquestes dades (les dades es troben en l'arxiu **Practica7.xlsx**):

Àrea de drenatge de l'aparcament	62	76	58	57	79	82	72	77
Àrea de control	72	77	60	59	61	64	69	65
Àrea de drenatge de l'aparcament	64	74	71	59	54	49	53	

Àrea de control	59	64	62	75	69	64	71	

a) Estima el nombre mitjà de plantes per panell trobades en cada àrea.
b) Estima la diferència del nombre mitjà de plantes trobades per panell.
c) Es creu que els contaminants procedents de l'aparcament faran disminuir el nombre de plantes trobades en l'àrea de drenatge. Confirmaria aquesta idea l'estimació puntual trobada en l'apartat b)? Pots estar molt segur que aquesta idea és correcta a partir d'aquesta estimació puntual? Si volgueres reforçar aquesta idea de manera que es poguera aportar un percentatge d'error què faries? Fes-ho.

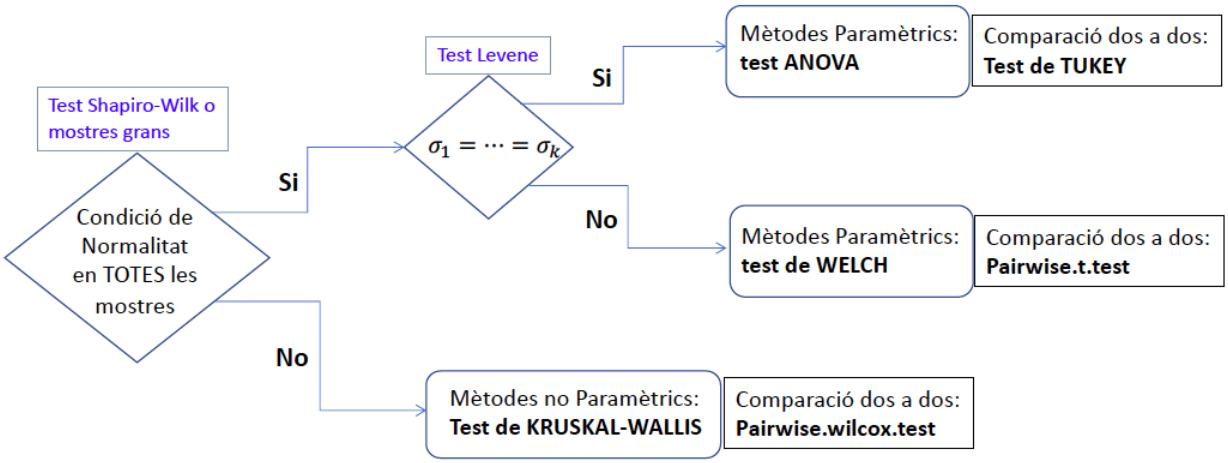
PRÀCTICA 8: ANÀLISI ESTADÍSTICA DE K MOSTRES INDEPENDENTS. Part I

En aquesta pràctica analitzarem i compararem k mostres independents. Els objectius són:

- Identificació dels objectius de l'estudi i del disseny emprat.
- Elecció de les proves adequades per a analitzar les dades.
- Interpretació de resultats.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

ELECCIÓ DE LES PROVES ADEQUADES PER A ANALITZAR LES DADES



Test ANOVA

Estadístics/Mitjanes/ANOVA d'un factor

Recordem el contrast d'hipòtesis:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A: \text{no es compleix } H_0 \end{cases}$$

Per tant, havent fixat un nivell de significació α resolem el contrast:

- Si p-valor < α, aleshores rebutgem H_0. I, per tant, assumim que les mitjanes no són iguals en els k grups. Per a veure entre quins grups hi ha diferències fem el **Test de TUKEY** (comparacions dos a dos).

Estadístics/Mitjanes/ ANOVA d'un factor (Seleccioneu: comparacions dos a dos de les mitjanes).

- Si p-valor≥ α, aleshores no rebutgem H_0. I considerem que les mitjanes en els k grups són iguals. En aquest cas no és necessari fer comparacions dos a dos.

Test de Welch

Com que és també un test paramètric, el contrast d'hipòtesis és el següent:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A: \text{No es compleix } H_0 \end{cases}$$

Estadístics/Mitjanes/ANOVA d'un factor (Seleccioneu Welch F-test)

En aquest cas, per a fer les comparacions dos a dos utilitzarem la instrucció:

```
pairwise.t.test(Dataset$VariableNumérica,Dataset$VariableCategòrica,pool.sd=FALSE)
```

Dataset és el nom del conjunt de dades, *VariableNumérica* és el nom de la variable amb els valors numèrics de la mostra i *VariableCategórica* és el nom de la variable en què s'especifiquen els grups.

Test de Kruskal-Wallis

En aquest cas estem tractant amb l'alternativa no paramètrica; per tant, el contrast d'hipòtesis és:

$$\begin{cases} H_0: \text{Totes les mostres es distribueixen igual (les medianes són iguals)} \\ H_A: \text{no es compleix } H_0 \end{cases}$$

Estadístics/Test no paramètrics/Test de Kruskal-Wallis

En aquest cas, per a fer les comparacions dos a dos hem d'utilitzar la instrucció:

```
pairwise.wilcox.test(Dataset$VariableNumérica,Dataset$VariableCategòrica,p.adjust="bonf")
```

EXERCICI 1

Les dades següents provenen d'un experiment realitzat a l'estació experimental de Rothamsted. L'objectiu era mesurar l'eficàcia de tres insecticides, el *cloronitrobenzè* (CN), el *disulfur de carboni* (CD) i un preparat propi denominat *cymag* (CM). Cada insecticida es va aplicar a dosi normal (1) i doble (2). Finalment, es va comptar amb un grup de control al qual no es va aplicar cap insecticida. Els pesticides es van aplicar abans de la sembra de blat, i les dades arreplegades mostren l'increment del nombre de cucs trobats en cada parcel·la després de la recol·lecció del blat.

	Insecticida						
Increment en el nombre de cucs	Control	1CN	1CD	1CM	2CN	2CD	2CM
	466	222	194	306	92	166	28
	421	219	221	176	114	172	179
	561	332	308	215	80	111	165
	433	298	256	199	128	80	82

a) Tria la tècnica estadística adequada per a comparar els insecticides.
b) Amb la tècnica estadística seleccionada en l'apartat anterior, investiga si hi ha diferències entre els resultats dels distints pesticides. Construeix els grups homogenis.

EXERCICI 2

L'Agència Internacional per a la Investigació del Càncer ha emès un informe, elaborat per un grup de 31 experts de 14 països, en què classifiquen els camps electromagnètics de radiofreqüència com un carcinogen del grup 2B. L'ús de telèfons mòbils s'inclou dins de les categories d'exposició a camps electromagnètics de radiofreqüència. Per comprovar experimentalment aquesta possible relació, es van seleccionar 120 animals de laboratori i se'ls va assignar a l'atzar a sis grups, 1 de control i 5 experimentals, de la mateixa grandària (20 unitats per grup). Es va subjectar un telèfon mòbil al cap de cada animal i es va mantindre en funcionament 0, 1, 2, 3, 4 i 5 hores segons el grup. Acabat l'experiment, es van mesurar els tumors (en mm^3) situats prop del lloc on havien estat les antenes dels telèfons (darrere l'orella). Indica el mètode inferencial que resulta adequat per analitzar aquesta mostra. Justifica la teua elecció.

a) Mitjançant l'anàlisi escaient, indica el mètode estadístic adequat per a comparar les sis mostres de l'estudi.
b) Amb el mètode seleccionat en l'apartat anterior, comprova si les dades proporcionen evidència suficient per a dir que hi ha alguna diferència entre els grups pel que fa a la variable d'interès.
c) Si es compleixen les condicions per a utilitzar els tests de comparacions múltiples, construeix, a partir d'ells, els conjunts homogenis i explica breument les conclusions que proporcionen. Si no es compleixen, explica per què.

EXERCICI 3

El cucut *Cuculus canorus* és un ocell de l'ordre dels cuculiformes. Una característica d'aquesta espècie és el parasitisme a què la femella sotmet els nius d'altres espècies d'ocells mitjançant la substitució dels ous que hi ha als nius pels ous propis. En abril-juny, la femella cerca uns pares adoptius que siguen de la mateixa espècie que la va criar quan era petita i pon un ou en cadascun dels nius que visita, després de retirar-ne un altre amb l'ajut del bec. Aquesta acció pot repetir-se fins a 12-13 vegades, és a dir, la femella pot pondre 12 o 13 ous en diferents nius de l'espècie corresponent. Amb el manteniment del nombre inicial d'ous, i amb el sorprenent mimetisme d'aquests amb els originals, és suficient perquè els pares parasitats confonguen la seua niada.

En efecte, un estudi realitzat per E. B. Chance en 1940 anomenat «La veritat sobre el cucut» va demostrar que els cucuts, després d'anys de retorn al mateix territori, ponen els seus ous en els nius d'una espècie d'acollida en particular. Per tant, es desenvolupen subespècies geogràfiques, cada una amb una espècie dominant de pare adoptiu.

Tippett va ser un dels pioners en el camp del control estadístic de qualitat, presenta dades sobre les longituds dels ous de cucut als nius d'altres aus. Les dades arreplegades per Tippett estan en l'arxiu **pràctica8.xlsx**.

a) Analitza'n les dades per a decidir el mètode estadístic adequat i analitzar-les.
b) Indaga, amb el mètode seleccionat en l'apartat anterior, si hi ha diferències en la grandària dels ous segons l'espècie d'acollida. Estableix els grups homogenis.

EXERCICI 4

En un estudi sobre la clòtxina *Mytilus trossulus* es va mesurar, entre altres variables, la longitud de la cicatriu de múscul adductor anterior, dividida per la longitud total. Les següents dades recullen la longitud en individus de cinc localitats diferents:

Tillamook	Newport	Petersburg	Magadan	Tvasminne
0,0571	0,0873	0,0974	0,1033	0,0703
0,0813	0,0662	0,1352	0,0915	0,1026
0,0831	0,0672	0,0817	0,0781	0,0956
0,0976	0,0819	0,1016	0,0685	0,0973
0,0817	0,0749	0,0968	0,0677	0,1039
0,0859	0,0649	0,1064	0,0697	0,1045
0,0735	0,0835	0,1050	0,0764	
0,0659	0,0725		0,0689	
0,0923				
0,0836				

Analitza aquestes dades amb la tècnica estadística que consideres adequada i proporciona les conclusions oportunes.

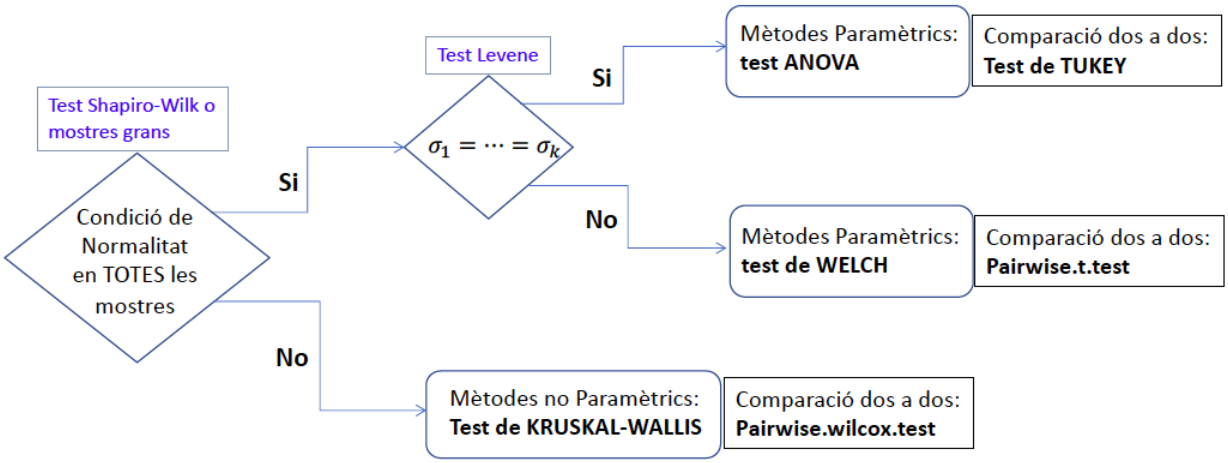
PRÀCTICA 9: ANÀLISI ESTADÍSTICA DE K MOSTRES INDEPENDENTS. Part II

En aquesta pràctica analitzarem i compararem k mostres independents. Els objectius són:

- Identificació dels objectius de l'estudi i del disseny emprat.
- Elecció de les proves adequades per a analitzar les dades.
- Interpretació de resultats.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

ELECCIÓ DE LES PROVES ADEQUADES PER A ANALITZAR LES DADES



Test ANOVA

Estadístics/Mitjanes/ANOVA d'un factor

Recordem el contrast d'hipòtesis:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A: \text{No es compleix } H_0 \end{cases}$$

Per tant, havent fixat un nivell de significació α resolem el contrast:

- Si p-valor < α, aleshores rebutgem H_0. Per tant, assumim que les mitjanes no són iguals en els k grups. Per a veure entre quins grups hi ha diferències fem el **Test de TUKEY** (comparacions dos a dos).

Estadístics/Mitjanes/ ANOVA d'un factor (Seleccioneu: comparacions dos a dos de les mitjanes).

- Si p-valor≥ α, aleshores no rebutgem H_0. I considerem que les mitjanes en els k grups són iguals. En aquest cas no és necessari fer comparacions dos a dos.

Test de Welch

Com que és també un test paramètric, el contrast d'hipòtesis és el següent:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A: \text{No es compleix } H_0 \end{cases}$$

Estadístics/Mitjanes/ANOVA d'un factor (Seleccioneu Welch F-test)

En aquest cas, per a fer les comparacions dos a dos utilitzarem la instrucció:

```
pairwise.t.test(Dataset$VariableNumérica,Dataset$VariableCategòrica,pool.sd=FALSE)
```

Dataset és el nom del conjunt de dades, *VariableNumérica* és el nom de la variable amb els valors numèrics de la mostra i *VariableCategòrica* és el nom de la variable en què s'especifiquen els grups.

Test de Kruskal-Wallis

En aquest cas estem tractant amb l'alternativa no paramètrica; per tant, el contrast d'hipòtesis és:

$$\begin{cases} H_0: \text{Totes les mostres es distribueixen igual (les medianes són iguals)} \\ H_A: \text{no es compleix } H_0 \end{cases}$$

Estadístics/Test no paramètrics/Test de Kruskal-Wallis

En aquest cas, per a fer les comparacions dos a dos utilitzarem la instrucció:

```
pairwise.wilcox.test(Dataset$VariableNumérica,Dataset$VariableCategòrica,p.adjust="bonf")
```

EXERCICI

Medley i Clements (1988) van estudiar la resposta de les comunitats de diatomees als metalls pesants, especialment zinc, als rius de la regió de les Muntanyes Rocoses de Colorado, EUA. Com a part del seu estudi, van mostrejar una sèrie d'estacions (entre quatre i set) en sis rierols que, se sap, estan contaminats per metalls. Es va registrar la concentració de zinc (ZN) en cada estació, així com la riquesa de les espècies (DIVERSITAT).

En una de les seues anàlisis, va dividir les 34 estacions en quatre grups segons els nivells de zinc: molt baix ($\leq$ 20, 8 estacions), baix (21-50, 8 estacions), mitjà (51-200, 9 estacions) i alt ($\geq$ 200, 9 estacions).

a) Representa el diagrama de caixes de la DIVERSITAT segons els NIVELLS_ZN. Hi ha algun signe d'asimetria? S'hi detecten observacions anòmales? La dispersió de les dades de DIVERSITAT resulta homogènia entre els diferents nivells de zinc?
b) Analitza, amb la tècnica estadística adequada, la influència del nivell de ZN en la diversitat.
    1. Identifica els grups homogenis si n'hi haguera.
    2. Si és possible, dona l'interval de confiança al 95% per a la diferència de mitjanes entre el grup baix i alt.

PROBLEMA

Continuem aquesta pràctica amb l'arxiu **datwinsdef_lab.xls**, que consta de 108 registres ficticis. Cadascun representa un embaràs gemel·lar.

Dins del projecte «Anàlisi dels efectes del medi ambient sobre la salut i el desenvolupament del fetus en embarassos gemel·lars»' es pretén estudiar la relació pes-sexe dels bessons monozigòtics, així com les diferències de pes existents en els bebès nascuts prematurs segon el sexe.

Parlem de bebès prematurs segons el criteri de la seua edat gestacional. Encara que la duració mitjana d'un embaràs únic és de 40 setmanes, la duració mitjana per a un embaràs de bessons és de 37 setmanes. Per tant, si els bessons naixen abans de la setmana 37, són prematurs. Si naixen abans de la setmana 33, es parla de grans prematurs i entre la setmana 33 i la 37, de prematurs.

Les variables que intervenen en aquest estudi són:

- **sexe**, {xiquet, xiqueta}.
- **pes1,** pes al nàixer del bessó amb major pes.
- **sges,** setmanes de gestació.

El nostre objectiu és contestar a la pregunta:

Podem afirmar que, en el cas de les **xiquetes**, existeix una diferència significativa en el pes de les bessones amb **major pes (pes1)** segons siguen grans prematures, prematures o no prematures?

Per contestar a la pregunta, resol les qüestions següents:

a) Noteu que ens demanen treballar només amb les **xiquetes**, aleshores la primera cosa que hem de fer és filtrar la base de dades i crear un subconjunt de dades XIQUETES, en el qual estarà la informació dels embarassos en què han nascut xiquetes.
b) L'enunciat ens demana estudiar la diferència en el pes de les bessones dins de tres categories, però nosaltres tenim informació de les setmanes de gestació. Per tant, primer hem de disposar d'una variable categòrica CATSGES que indique el tipus d'embaràs.
c) Defineix les tres variables del problema, amb l'objectiu de fer la comparació múltiple.

A continuació, treballarem amb la base de dades filtrada XIQUETES

d) Fes un resum numèric de la variable **pes1** per grups **CATSGES**.
    - Dona estimacions puntuals per a la mitjana de les tres variables definides en l'apartat c).
    - En l'aspecte descriptiu, hi ha diferències entre el pes de les bessones amb major pes segons siguen grans prematures, prematures o no prematures?
e) És normal la variable pes1 en tots els grups generats per les setmanes de gestació?
f) Respon, justificadament, a la pregunta inicial.
   Podem afirmar que, en el cas de les **xiquetes**, existeix una diferència significativa en el pes de les bessones amb **major pes (pes1)** segons siguen grans prematures, prematures o no prematures?
g) Estableix els grups homogenis.
h) Calcula intervals de confiança al 95% per a les diferències entre les mitjanes poblacionals de la variable **pes1** segons el grup generat per la variable **CATSGES**.

PRÀCTICA 10: ANÀLISI DE DADES CATEGÒRIQUES. Part I.

L'objectiu d'aquesta pràctica és la inferència sobre proporcions, en particular:

- Test binomial.
- Test de bondat d'ajust.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

INFERÈNCIA SOBRE PROPORCIONS

La **inferència sobre una proporció** s'utilitza per a obtenir estimacions puntuals/intervals de confiança per a la proporció poblacional i per a resoldre contrastos d'hipòtesis de la forma:

$$\begin{cases} H_0: \pi = \pi_0 \\ H_A: \pi \neq \pi_0 \end{cases} \qquad \begin{cases} H_0: \pi \geq \pi_0 \\ H_A: \pi < \pi_0 \end{cases} \qquad \begin{cases} H_0: \pi \leq \pi_0 \\ H_A: \pi > \pi_0 \end{cases}$$

Per al càlcul de l'interval de confiança per a la proporció poblacional π utilitzem el test binomial. En la finestra d'instruccions de *R-Commander* podem escriure la instrucció següent:

binom.test (r, n, alternative = "two.sided", p = π_0, conf.level = CL)

On r és el nombre d'èxits, n és el nombre total de proves, p és la proporció que es vol comparar (canviarem aquest valor per als contrastos d'hipòtesis; per exemple $\pi_0 = ,30$) i conf.level és el nivell de confiança establert pel càlcul de l'interval de confiança (per exemple CL= ,95). En "alternative" indicarem "two.sided" per a contrastos bilaterals i per al càlcul d'intervals de confiança. En el cas que en la hipòtesi alternativa tinguem un menor que (<) posarem "less" i en el cas d'un major que (>) "greater".

Nota. Si volem calcular un interval de confiança, podem llevar la part p = π_0 o posar qualsevol valor. Això no afecta el resultat final. Anàlogament, per fer un contrast d'hipòtesis i l'opció conf.level.

L'**anàlisi de bondat d'ajust** s'utilitza quan estem interessats el contrast d'hipòtesis següent:

$$\begin{cases} H_0: \pi_i = \pi_{i0}, \quad i = 1,2, \dots k. \\ H_A: \text{hi ha almenys un j, tal que } \pi_j \neq \pi_{j0} \end{cases}$$

Per a resoldre el contrast d'hipòtesis anterior utilitzem el test khi-quadrat. En la finestra d'instruccions podem escriure la següent instrucció següent:

chisq.test (c (15, 26, 15, 0, 8), p = c (3/16, 6/16, 3/16, 1/16, 2/16, 1/16))

On hem indicat els valors observats en cada categoria i les proporcions que es volen contrastar. Si el p-valor és menor que el nivell de significació prefixat α, es rebutja la hipòtesi nul·la.

Cal notar que per poder traure conclusions del test khi-quadrat anterior hem de comprovar primer la condició d'aplicabilitat del test:

Almenys el 80% de les freqüències esperades majors o iguals que 5.

Les freqüències esperades es poden calcular amb *R-Commander* executant l'ordre següent:

```
chisq.test (c (15, 26, 15, 0, 8), p = c (3/16, 6/16, 3/16, 1/16, 2/16, 1/16)) $expected
```

La inferència sobre una proporció considera una única variable categòrica amb dues categories. Observada en una única mostra.

La bondat d'ajust estudia una variable categòrica amb k categories ($k \geq 2$). Observada en una única mostra.

EXERCICI 1

En el pelatge d'algun dels animals d'una determinada població de ratolins *Mus musculus* s'observen pigues blanques en el ventre. En una mostra de 580 ratolins d'aquesta població es va estudiar aquesta característica i es van trobar 28 ratolins amb les taques esmentades.

a) Dona una estimació puntual per a la proporció poblacional de ratolins amb pigues blanques en el ventre.
b) Estableix un interval de confiança del 99% per a la proporció poblacional amb la característica indicada en l'enunciat.
c) Les dades són consistents amb la hipòtesi que menys d'un 5% dels ratolins *Mus musculus* tenen pigues blanques en el ventre?

EXERCICI 2

D'un encreuament entre carabasses blanques i grogues es va obtenir la descendència següent:

COLOR	BLANC	GROC	VERD
NRE. DESCENDENT	155	40	10

a) Quins valors esperaríem observar si la descendència seguira les proporcions 12:3:1 proposades per un determinat model genètic.
b) Les dades són consistents amb la proporció teòrica 12:3:1 proposada per aquest model genètic?

EXERCICI 3

L'aparició en les fulles de les plàntules de cotó de glàndules de pigment pot controlar-se genèticament. D'acord amb una de les teories sobre el control d'aquest mecanisme, el quocient poblacional entre plantes amb glàndules i sense obtingut d'un determinat encreuament hauria de ser 11:5, mentre que per a una altra teoria hauria de ser 13:3. A conseqüència d'un encreuament es van obtenir en un experiment 89 plantes amb glàndules i 36 plantes sense glàndules. Estudia la compatibilitat de les dades amb cadascuna de les teories formulades.

EXERCICI 4

En una secció de 30 metres quadrats d'una àrea sembrada amb *Dentaria* es van comptar totes les plantes i se'n van comptar 296 amb flor i 987 sense.

a) Estableix un interval de confiança del 95% de la proporció teòrica de plantes de *Dentaria* amb flor.

En altres dues mostres diferents de la mateixa àrea es van comptar sis plantes amb flor enfront de vint sense, i 29 amb flor enfront de 485 sense.

b) Estableix en cada cas un interval de confiança del 95% per a la proporció de plantes amb flor en la població. Comenta els resultats obtinguts.

EXERCICI 5

En un experiment sobre la reproducció es va aparellar un grup de pollastres blancs de cresta petita i van donar lloc als 190 descendents que presentem en la taula següent:

TIPUS	NOMBRE DE DESCENDENTS
PLOMATGE BLANC, CRESTA PETITA	111
PLOMATGE BLANC, CRESTA GRAN	37
PLOMATGE FOSC, CRESTA PETITA	34
PLOMATGE FOSC, CRESTA GRAN	8

Les dades són consistents amb les proporcions mendelianes teòriques 8:3:3:1? Considera $\alpha = 0{,}1$.

PRÀCTICA 11: ANÀLISI DE DADES CATEGÒRIQUES. Part II.

L'objectiu d'aquesta pràctica és la inferència sobre proporcions, en particular:

- Test binomial.
- Test de bondat d'ajust.
- Test khi-quadrat per a taules de contingència.
- Test de Fisher i ODDs ratio.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

INFERÈNCIA SOBRE PROPORCIONS

La **inferència sobre una proporció** s'utilitza per a obtenir estimacions puntuals/intervals de confiança per a la proporció poblacional i per a resoldre contrastos d'hipòtesis de la forma:

$$\begin{cases} H_0: \pi = \pi_0 \\ H_A: \pi \neq \pi_0 \end{cases} \qquad \begin{cases} H_0: \pi \geq \pi_0 \\ H_A: \pi < \pi_0 \end{cases} \qquad \begin{cases} H_0: \pi \leq \pi_0 \\ H_A: \pi > \pi_0 \end{cases}$$

Per al càlcul de l'interval de confiança per a la proporció poblacional π utilitzem el test binomial. En la finestra d'instruccions de *R-Commander* podem escriure la instrucció següent:

binom.test (r, n, alternative = "two.sided", p = π_0, conf.level = CL)

On r és el nombre d'èxits, n és el nombre total de proves, p és la proporció que es vol comparar (canviarem aquest valor per als contrastos d'hipòtesis; per exemple: $\pi_0 = ,30$) i conf.level és el nivell de confiança establert pel càlcul de l'interval de confiança (per exemple CL= ,95). En "alternative" indicarem "two.sided" per a contrastos bilaterals i per al càlcul d'intervals de confiança. En el cas que en la hipòtesi alternativa tinguem un menor que (<) posarem "less" i en el cas d'un major que (>) "greater".

Nota. Si volem calcular un interval de confiança, podem llevar la part p = π_0 o posar qualsevol valor. Això no afecta el resultat final. Anàlogament, per fer un contrast d'hipòtesis i l'opció conf.level.

L'**anàlisi de bondat d'ajust** s'utilitza quan estem interessats pel contrast d'hipòtesis següent:

$$\begin{cases} H_0: \pi_i = \pi_{i0}, \quad i = 1,2, \dots k. \\ H_A: \text{hi ha almenys un j, tal que } \pi_j \neq \pi_{j0} \end{cases}$$

Per resoldre el contrast d'hipòtesis anterior utilitzem el test khi-quadrat. En la finestra d'instruccions podem escriure la instrucció següent:

chisq.test (c (15, 26, 15, 0, 8), p = c (3/16, 6/16, 3/16, 1/16, 2/16, 1/16))

On hem indicat els valors observats en cada categoria i les proporcions que es volen contrastar. Si el p-valor és menor que el nivell de significació prefixat α, es rebutja la hipòtesi nul·la.

Cal notar que per poder traure conclusions del test khi-quadrat anterior hem de comprovar primer la condició d'aplicabilitat del test:

Almenys el 80% de les freqüències esperades majors o iguals que 5.

Les freqüències esperades es poden calcular amb *R-Commander* executant l'ordre següent:

chisq.test (c (15, 26, 15, 0, 8), p = c (3/16, 6/16, 3/16, 1/16, 2/16, 1/16)) $expected

ANÀLISI DE TAULES DE CONTINGÈNCIA

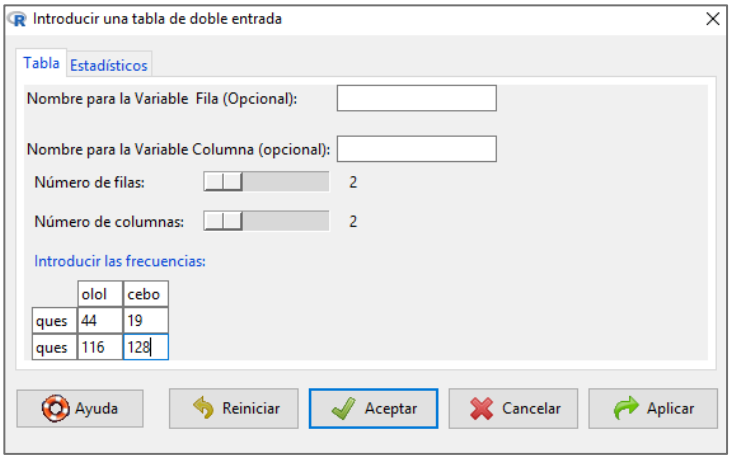
Una taula de contingència, en general, és de la forma:

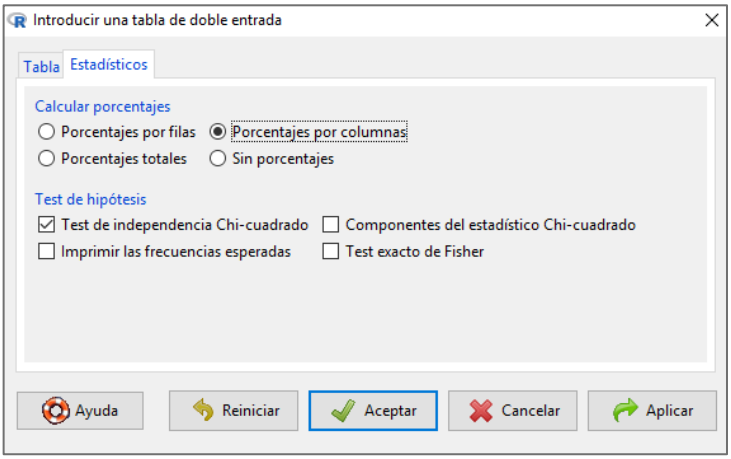
	Categories variable 1		
Categories variable 2	O_{11}	...	O_{1r}
	O_{21}	...	O_{2r}
	...	...	...
	O_{k1}	...	O_{kr}

On O_{ij} és la freqüència observada en la combinació de categories ij (cel·la (i,j) de la taula).

En *R-Commander* les taules de contingència s'introdueixen i analitzen amb l'odre següent del menú:

Estadístics / taules de contingència / introduir i analitzar una taula de doble entrada





Cal notar que per poder traure conclusions del test khi-quadrat anterior hem de comprovar primer la condició d'aplicabilitat del test:

Almenys el 80% de les freqüències esperades majors o iguals que 5.

Les freqüències esperades es poden calcular en *R-Commander* seleccionant l'opció **Imprimir les freqüències esperades**.

TAULES DE CONTINGÈNCIA 2 X 2

En les taules de contingència 2 x 2 podem utilitzar l'ODDs ratio, que indica la direccionalitat de l'associació. El test de Fisher és una alternativa al test khi-quadrat per a taules 2 x 2. El test calcula l'ODDs ratio i resol el contrast d'hipòtesis següent:

$$\begin{cases} H_0: OR = 1 \\ H_A: OR \neq 1 \end{cases}$$

El test de Fisher pot fer-se en *R-Commander* seleccionant l'opció **Test exacte de Fisher**.

Cóm interpretem l'ODDs ratio?

- Si OR = 1, no hi ha associació entre la presència del factor i el succés.
- Si OR > 1, l'associació és positiva (factor de risc), el factor s'associa amb la major ocurrència del succés.
- Si OR < 1, es considera que l'associació és negativa (factor protector), el factor s'associa amb una menor ocurrència del succés.

Per tant, el contrast d'hipòtesis indica si hi ha o no hi ha associació entre els factors. (OR = 1 independents; OR ≠ 1 relacionats.)

PROBLEMA

En aquesta pràctica treballarem sobre la base de dades **datwins2.xls**, que inclou algunes variables que no estan disponibles en el llibre **datwinsdef_lab.xls** (que hem estat usant en les anteriors sessions de pràctiques).

Com a part de l'estudi *Anàlisi dels efectes del medi ambient sobre la salut i el desenvolupament del fetus en embarassos gemel·lars* es pretén estudiar els índexs de poc pes en nàixer del bessó amb major pes, **bajopeso1**, els hàbits tabàquics de les mares durant l'embaràs, **fuma_emb**, així com les relacions existents entre el poc pes en nàixer del bessó amb major pes i l'hàbit tabàquic de la mare a l'inici de l'embaràs, **fuma_ini**.

Les variables amb què treballarem en aquesta pràctica són:

- **bajopeso1**. Variable dicotòmica que indica si el bessó amb major pes té poc pes {Sí, No}.
- **fuma_emb**. Variable amb tres categories que indica si la mare no ha fumat durant l'embaràs, si solament ha fumat en el primer trimestre o si ha fumat en tot l'embaràs {No, primer, tot}.
- **fuma_ini.** Variable dicotòmica que indica si la mare fumava a o no a l'inici de l'embaràs {Sí, No}.

Ajuda. Per poder utilitzar les ordres indicades abans en la inferència sobre proporcions has de calcular primer les freqüències observades. Recorda, que les pots obtenir amb l'ordre de *R-Commander* següent:

> Estadístics / resums / distribució de freqüències

Ajuda. Per poder utilitzar les ordres indicades abans en l'anàlisi de taules de contingència has de calcular primer la taula de contingència. La pots obtenir amb l'ordre ordre de *R-Commander* següent*:*

Estadístics / taules / taules de doble entrada

A continuació respon les preguntes següents:

a) Segons diversos estudis, el percentatge de poc pes en la població general és del 6%, aproximadament. Podem concloure que en parts gemel·lars el percentatge de poc pes en nàixer per al bessó de major pes, **bajopeso1**, és superior que en la població general? Resol i interpreta el contrast corresponent. Dona l'interval de confiança del 95% per al percentatge de poc pes en nàixer del bessó amb major pes.
b) Podem concloure que fumar durant l'embaràs, **fuma_emb**, té una distribució que segueix la regla de proporcionalitat 10:3:2 (No, primer, tot)? Respon detalladament aquesta pregunta i justifica la validesa del test utilitzat.
c) Hi ha relació entre les variables **bajopeso1** i **fuma_ini**? Interpreta els resultats obtinguts i estableix les condicions de validesa del test utilitzat. Proporciona i interpreta l'ODDS ratio.
d) Hi ha relació entre les variables **bajopeso1** i **fuma_emb**? Interpreta els resultats obtinguts i estableix les condicions de validesa del test utilitzat.
e) En relació amb el bessó de major pes, contesta les preguntes següents:
    - Quin és el percentatge de xiquets amb poc pes i mares que van fumar durant tot l'embaràs en la mostra?
    - Quin és el percentatge de xiquets amb poc pes entre les dones que fumaren durant tot l'embaràs?
    - Quin és el percentatge de dones fumadores durant tot l'embaràs entre els xiquets de poc pes?

PRÀCTICA 12: REGRESSIÓ LINEAL. Part I

Els objectius d'aquesta pràctica són:

- Identificar el tipus de problemes estadístics que poden resoldre's per mitjà de la regressió lineal.
- Fer representacions gràfiques de dues variables contínues (núvol de punts) i ajust per mínims quadrats.
- Càlcul i interpretació del coeficient de correlació lineal.
- Obtenció de la recta de regressió. Prediccions.
- Tests estadístics i intervals de confiança sobre els paràmetres de la recta.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

REPRESENTACIÓ GRÀFICA

El diagrama de dispersió és una representació gràfica bidimensional de les observacions. Per a dibuixar-lo en *R-Commander* utilitzem l'ordre següent del menú:

Gràfiques → Diagrama de dispersió

Per a dibuixar la recta de mínims quadrats en la pestanya **opcions** hem de seleccionar l'opció **línia de mínims quadrats**.

COEFICIENT DE CORRELACIÓ LINEAL

Podem calcular la matriu de correlacions amb *R-Commander* seleccionant:

Estadístics → Resums → Matriu de correlacions

Hem de seleccionar totes les variables que volem estudiar. A més, si es vol conèixer si la relació obtinguda és significativa, hem de seleccionar l'opció "**valors p aparellats"**.

En la figura següent, observem que hi ha relació lineal directa entre la longitud i el pes, perquè el coeficient de correlació és r = 0,9437. A més, la relació és significativa, atès que valor p = 0,0001 < α = 0,05.

```
> rcorr.adjust(serpientes[,c("Longitud","Peso")],
 type="pearson", use="complete")

 Pearson correlations:
         Longitud   Peso
Longitud   1.0000 0.9437
Peso       0.9437 1.0000

 Number of observations: 9

 Pairwise two-sided p-values:
         Longitud Peso
Longitud          0.0001
Peso     0.0001
```

En general, el coeficient de correlació s'interpreta de la forma següent:

- Si r > 0, la relació és directa i major com més gran és r. En particular, si r = 1, la relació és directa perfecta.
- Si r < 0, la relació és inversa i major com més petit és r. En particular si r = -1 la relació és inversa perfecta.

Si r = 0, no hi ha relació lineal.

OBTENCIÓ DE LA RECTA DE REGRESSIÓ

En *R-Commander*, la recta de regressió s'obté amb l' opció següent del menú:

> Estadístics → Ajust de models → Regressió lineal

Quan fem la regressió lineal, en la finestra de resultats apareix el coeficient de determinació. El coeficient de determinació és la proporció de variabilitat de Y explicada per la regressió (pel model de regressió). Si és 0, el model no explica res de Y a partir de X. En canvi, si és 1, el model proporciona un ajust perfecte.

Veiem que la recta de regressió quan Y = Pes i X = Longitud és de la forma

Y = - 301,0872 + 7,1919X.

A més el coeficient de determinació és $R^2 = r^2$ = 0,8905. Per tant el 89,05% de variabilitat de Y està explicada pel model de regressió obtingut.

```
> RegModel.1 <- lm(Peso~Longitud, data=serpientes)

> summary(RegModel.1)

Call:
lm(formula = Peso ~ Longitud, data = serpientes)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -301.0872    60.1885  -5.002 0.001561 **
Longitud       7.1919     0.9531   7.546 0.000132 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.5 on 7 degrees of freedom
Multiple R-squared:  0.8905,  Adjusted R-squared:  0.8749
F-statistic: 56.94 on 1 and 7 DF,  p-value: 0.0001321
```

CONTRAST D'HIPÒTESI

Quan fem regressió lineal, hem de comprovar que els coeficients obtinguts són significatius. Per tant, hem de resoldre els contrastos d'hipòtesis següents:

$$\begin{cases} H_0: \beta_0 = 0 \\ H_A: \beta_0 \neq 0 \end{cases} \qquad \begin{cases} H_0: \beta_1 = 0 \\ H_A: \beta_1 \neq 0 \end{cases}$$

Quan fem la regressió lineal, *R-Commander* ens proporciona informació sobre la significativitat dels coeficients. En l'exemple anterior veiem que el p-valor per a β_1 (coeficient de la linealitat) és 0,000132 i, per tant, el coeficient obtingut és significatiu. Anàlogament per a β_0, l'intercepte.

A més de les estimacions puntuals dels paràmetres obtinguts de la regressió lineal, podem calcular també intervals de confiança. Després d'haver fet la regressió lineal en *R-Commander*, hem de seleccionar l'ordre següent:

> Models → Interval de confiança

EXERCICI 1

La taula següent mostra les pèrdues de pes mitjanes, observades en 9 grups de 26 escarabats, després de 6 dies de ser sotmesos a distints graus d'humitat relativa.

Pèrdua de pes (mg)	8,98	8,14	6,67	6,08	5,9	5,83	4,68	4,20	3,72
%Humitat relativa	0	12	29,5	43	53	62,5	75,5	85	93

Es tracta d'estudiar la relació lineal entre ambdues variables i predir la pèrdua de pes mitjana dels escarabats en funció de la humitat.

a) Identifica la variable explicada Y i l'explicativa X.
b) Realitza un gràfic de dispersió per veure com varia Y en funció de X.
   - Com es comporta la pèrdua de pes quan creix la humitat?
   - S'ajusten els punts del gràfic a una recta?
   - Gràficament, podem pensar que un model de regressió lineal és adequat?
c) Determina i interpreta els coeficients de la recta de mínims quadrats.
d) Podem dir que hi ha relació lineal entre les variables? Dona la resposta a nivell 0,01.
e) Calcula el valor pronosticat de la pèrdua de pes si la humitat és del 80%.
f) Podries estimar la pèrdua de pes mitjana si la humitat relativa fora del 98%? Raona la resposta.

EXERCICI 2

S'ha realitzat un estudi de fotoperíode en aus aquàtiques. En aquesta anàlisi, es pretén establir una equació mitjançant la qual es puga predir el temps de reproducció, Y, basant-se en el coneixement del fotoperíode (nombre d'hores de llum per dia) sota el qual es va iniciar la reproducció, X. Es van obtenir dades del comportament d'ànecs bussejadors. Els resultats van ser els següents:

Temps de reproducció	110	54	98	50	67	58	52	50	43	15	28
Fotoperíode	12,8	13,9	14,1	14,7	15	15,1	16	16,5	16,6	17,2	17,9

a) Dibuixa el núvol de punts de Y= Temps de reproducció sobre X = Fotoperíode.
   - És raonable considerar l'existència d'una relació lineal entre les variables?
   - Quin seria el valor esperat del coeficient de correlació, pròxim a 1, a -1, o, a 0? Determina i interpreta els coeficients de la recta de mínims quadrats.
b) Quina és la recta de regressió lineal que permet predir el temps de reproducció a partir del fotoperíode? Són els paràmetres obtinguts significatius?
c) Dona els intervals de confiança al 95% dels paràmetres de la regressió lineal.
d) Podem afirmar que hi ha una relació lineal entre les variables? Formula i resol un contrast d'hipòtesis que et permeta contestar a aquesta pregunta.
e) Calcula i interpreta el coeficient de correlació lineal.
f) Quin percentatge de variabilitat del temps de reproducció NO queda explicat per la recta de regressió obtinguda? Què pots concloure a partir del valor obtingut?
g) Quin canvi cal esperar en el temps de reproducció de les aus aquàtiques per cada hora extra de llum diària en el moment d'iniciar-se la reproducció?

h) Quin seria el temps de reproducció d'una au el fotoperíode de la qual ha sigut 14,5 hores?
i) És possible utilitzar la recta de regressió obtinguda per a realitzar una estimació del temps de reproducció d'una au el fotoperíode de la qual és de 20 hores? Justifica la resposta.

PRÀCTICA 13: REGRESSIÓ LINEAL. Part II

-

Els objectius d'aquesta pràctica són:

- Identificar el tipus de problemes estadístics que poden resoldre's per mitjà de la regressió lineal.
- Fer representacions gràfiques de dues variables contínues (núvol de punts) i ajust per mínims quadrats.
- Càlcul i interpretació del coeficient de correlació lineal.
- Obtenció de la recta de regressió. Prediccions.
- Tests estadístics i intervals de confiança sobre els paràmetres de la recta.

Per a la realització de les pràctiques farem ús del programari estadístic *R* i la seua interfície gràfica *R-Commander.*

REPRESENTACIÓ GRÀFICA

El diagrama de dispersió és una representació gràfica bidimensional de les observacions. Per a dibuixar-lo en *R-Commander* utilitzem l'ordre següent del menú:

> Gràfiques → Diagrama de dispersió

Per a dibuixar la recta de mínims quadrats en la pestanya **opcions** heu de seleccionar l'opció **línia de mínims quadrats**.

COEFICIENT DE CORRELACIÓ LINEAL

Podem calcular la matriu de correlacions *amb R-Commander* seleccionant:

> Estadístics → Resums → Matriu de correlacions

Heu de seleccionar totes les variables que voleu estudiar. A més, si es vol conèixer si la relació obtinguda és significativa, hem de seleccionar l'opció **valors p aparellats**.

En la figura següent, observem que hi ha relació lineal directa entre la longitud i el pes, ja que el coeficient de correlació és r = 0,9437. A més, la relació és significativa, atès que valor p = 0,0001 < α = 0,05.

```
> rcorr.adjust(serpientes[,c("Longitud","Peso")],
 type="pearson", use="complete")

 Pearson correlations:
         Longitud   Peso
Longitud   1.0000 0.9437
Peso       0.9437 1.0000

 Number of observations: 9

 Pairwise two-sided p-values:
         Longitud Peso
Longitud          0.0001
Peso     0.0001
```

En general, el coeficient de correlació s'interpreta de la manera següent:

- Si r > 0, la relació és directa i major com més gran és r. En particular, si r = 1, la relació és directa perfecta.
- Si r < 0, la relació és inversa i major com més petit és r. En particular si r = -1, la relació és inversa perfecta.

Si r = 0, no hi ha relació lineal.

OBTENCIÓ DE LA RECTA DE REGRESSIÓ

En *R-Commander*, la recta de regressió s'obté amb l'opció següent del menú:

> Estadístics → Ajust de models → Regressió lineal

Quan fem la regressió lineal, en la finestra de resultats apareix el coeficient de determinació. El coeficient de determinació és la proporció de variabilitat de Y explicada per la regressió. Si és 0, el model no explica res de Y a partir de X. En canvi, si és 1, el model proporciona un ajust perfecte.

Veiem que la recta de regressió quan Y = pes i X = longitud és de la forma

Y = - 301,0872 + 7,1919X.

A més, el coeficient de determinació és

$R^2 = r^2$ = 0,8905. Per tant el 89,05% de variabilitat de Y està explicada pel model de regressió obtingut.

```
> RegModel.1 <- lm(Peso~Longitud, data=serpientes)

> summary(RegModel.1)

Call:
lm(formula = Peso ~ Longitud, data = serpientes)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -301.0872    60.1885  -5.002 0.001561 **
Longitud       7.1919     0.9531   7.546 0.000132 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.5 on 7 degrees of freedom
Multiple R-squared:  0.8905,  Adjusted R-squared:  0.8749
F-statistic: 56.94 on 1 and 7 DF,  p-value: 0.0001321
```

CONTRAST D'HIPÒTESI

Quan fem regressió lineal, hem de comprovar que els coeficients obtinguts són significatius. Per tant, hem de resoldre els contrastos d'hipòtesis següents:

$$\begin{cases} H_0: \beta_0 = 0 \\ H_A: \beta_0 \neq 0 \end{cases} \qquad \begin{cases} H_0: \beta_1 = 0 \\ H_A: \beta_1 \neq 0 \end{cases}$$

Quan fem la regressió lineal, *R-Commander* ens proporciona informació sobre la significativitat dels coeficients. En l'exemple anterior veiem que el valor p per a β_1 (coeficient de la linealitat) és 0,000132 i, per tant, el coeficient obtingut és significatiu. Anàlogament per a β_0, l'intercepte.

A més de les estimacions puntuals dels paràmetres obtinguts de la regressió lineal, podem calcular també intervals de confiança. Després d'haver fet la regressió lineal en *R-Commander* hem de seleccionar l'ordre següent:

> Models → Interval de confiança

PROBLEMA

Continuem en aquesta pràctica amb l'arxiu **datwinsdef_lab.xlsx**, que consta de 108 registres ficticis, cadascun representant un embaràs gemel·lar.

Com a part de l'estudi "Anàlisi dels efectes del medi ambient sobre la salut i el desenvolupament del fetus en embarassos gemel·lar" es pretén estudiar la relació pes-sexe dels bessons monozigòtics, així com les diferències de pes en els bebès nascuts prematurs segons el sexe.

Les variables que treballarem durant aquesta pràctica són:

- **sexe**. {xiquet, xiqueta}
- **pes1**, pes en nàixer del bessó amb major pes
- **talla1**, talla en nàixer del bessó amb major pes
- **pesm,** pes de la mare
- **tallap,** talla del pare

Contesta a les preguntes següents:

a) Realitza un diagrama de dispersió per estudiar la relació lineal entre el pes (X, variable explicativa) i la talla (Y, variable explicada) en nàixer del bessó amb major pes. Interpreta la gràfica.
b) Realitza un diagrama de dispersió amb la línia de mínims quadrats per estudiar la relació lineal entre el pes i la talla en nàixer del bessó amb major pes, separant per sexe (és a dir, fes la gràfica per grups. D'aquesta manera et dibuixarà en la mateixa gràfica el diagrama de dispersió per a xiquetes i per a xiquets). Interpreta la gràfica.
c) Realitza un diagrama de dispersió amb la línia de mínims quadrats per estudiar la relació lineal entre el pes del pare i la talla de la mare. Interpreta la gràfica.
d) Calcula el coeficient de correlació lineal entre les variables pes1 i talla1. Si hi ha relació, és directa o inversa? És significativa al 5%?
e) Calcula la recta de regressió que consideres adequada per a estudiar la relació entre el pes i la talla en nàixer del bessó amb major pes.
    - Interpreta la recta.
    - Realitza i interpreta el contrast de linealitat.
    - Obté l'interval de confiança pel pendent.
f) Quin és el percentatge de variabilitat de la talla explicada mitjançant la recta de regressió? I la no explicada?
g) Quina és la talla estimada per a un bessó amb un pes en nàixer d'1 kg?
h) Si considerem únicament les xiquetes, augmenta la força de l'associació lineal?
Ajuda: com que pregunten per les xiquetes, en aquest cas caldrà filtrar i treballar sobre el subconjunt de dades que continga únicament les xiquetes.
i) Calcula i interpreta el coeficient de correlació lineal entre la variable pesm i tallap (*en tota la mostra*).
j) Calcula la recta de regressió entre el pes de la mare i la talla del pare. Elabora i interpreta el contrast de linealitat.