Pràctica 2

Estadística descriptiva

2.1 Tipus de dades

La primera cosa que necessitem abans de començar a estudiar les dades és saber si aquestes són quantitatives o qualitatives. Entre les variables que podem estudiar n'hi ha de diversos tipus. De manera una mica simple les podem dividir en dues classes:

1. **Dades qualitatives**. Es refereixen a una característica no numèrica de l'individu. En principi no s'expressen numèricament, però a la pràctica moltes vegades es codifiquen numèricament per facilitar-ne el tractament. Els números simplement funcionen com etiquetes assignades a unes categories. Per exemple, el sexe d'un individu és una variable qualitativa que pot prendre dos valors: M (masculí) i F (femení), però moltes vegades es codifica amb un 1 (M) o un 0 (F). En general, no es poden ordenar, però a vegades són qualitatives ordinals, és a dir, les etiquetes es poden ordenar. Un exemple és el cas de les franges d'edat.
2. **Dades quantitatives**. Són les que es refereixen a característiques dels individus que s'expressen numèricament. Dins d'aquestes en podem trobar de **discretes** –quan només poden prendre un nombre discret de valors– o de **contínues** –quan poden prendre qualsevol valor dins d'un interval.

Exercici: Obriu les dades `iris` que té incorporades el programa `R` i decidiu de quin tipus són les 5 variables que conté.

2.2 Taules de freqüències

Per representar les dades podem utilitzar les taules de freqüències, on trobem

- **freqüència absoluta**: nombre de repeticions
- **freqüència relativa**: nombre de repeticions dividit pel nombre total d'observacions,
- **freqüència absoluta acumulada**:suma de les freqüències absolutes dels valors més petits o igual,
- **freqüència relativa acumulada**:suma de les freqüències relatives dels valors més petits o igual.

Les intruccions són:

```
table (x)
table(x)/length(x)
cumsum(table(x))
cumsum(table(x)/length(x))
```

respectivament.

Exemple: Les dades obtingudes per una variable venen resumides en la taula de freqüències següent:

valors	freqüències
0	40
2	80
4	16
5	4

Per obtenir la taula de freqüències absolutes, relatives, absolutes acumulades i relatives acumulades fem:

```
x<- rep(c(0,2,4,5),c(40,80,16,4))
y<-table (x)
z<-y/length(x)
y
z
cumsum(y)
cumsum(z)
```

Obtenim:

```
> y
x
 0  2  4  5
40 80 16  4
> z
x
          0           2           4           5
0.28571429 0.57142857 0.11428571 0.02857143
> cumsum(y)
  0   2   4   5
 40 120 136 140
> cumsum(z)
         0         2         4         5
0.2857143 0.8571429 0.9714286 1.0000000
>
```

Es pot aconseguir una presentació en forma de taula amb el format `dataframe`.:

```
taula<-as.data.frame(table(x))
 taula
  x Freq
1 0   40
2 2   80
3 4   16
4 5    4
```

Per afegir més columnes amb les altres freqüències fem:

```
taula<-transform(taula, Freq.Acum = cumsum(Freq), Freq.Rela = prop.table(Freq),
Freq.Rela.Acum=cumsum(prop.table(Freq)))
taula
```

```
  x Freq Freq.Acum  Freq.Rela Freq.Rela.Acum
1 0   40        40 0.28571429      0.2857143
2 2   80       120 0.57142857      0.8571429
3 4   16       136 0.11428571      0.9714286
4 5    4       140 0.02857143      1.0000000
```

2.3 Representació gràfica de les dades

Diagrama de tija i fulles

Ens permet obtenir simultàniament una distribució de freqüències i la seva representació gràfica.

```
stem(x)
```

Exemple

```
> stem(LakeHuron)

  The decimal point is at the |

  575 |
  576 | 02888899999
  577 | 1224578889
  578 | 011122223444567778899
  579 | 00111111222333444455666667778888999
  580 | 000011444456889
  581 | 0234479
```

Histograma

Per a dades quantitatives contínues la representació gràfica habitual és l'**histograma**. Cal dividir el rang de les dades en classes, si pot ser de la mateixa amplada i dibuixar columnes amb àrea proporcional a la freqüència de les dades d'aquella classe; si les amplades de les classes són iguals, això és equivalent a que les alçades de les columnes siguin proporcionals a les freqüències de les classes.

```
y<-rexp(200,rate=2)
    hist(y)
```

Paràmetres optatius de la funció `hist()`:

- Per posar un nombre de classes fixat

  ```
  hist(y, nclass=12)
  ```

- Per posar les vores dels rectangles en posicions concretes. En particular, això permet tenir un histograma amb intervals desiguals:

  ```
  hist(y, breaks=c(0,0.5,1,2,4))
  ```

- Per fer que les alçades dels rectangles siguin les proporcions (freqüències relatives), en comptes de les freqüències absolutes,

```
hist(y, freq=FALSE)
```

- Aquesta opció és útil si volem comprovar en quina mesura l'histograma s'aproxima a la funció de densitat de probabilitat. Si, com és el cas amb aquests exemples, hem generat dades d'una llei coneguda, podem comparar-les directament:

```
y<-rexp(200,rate=2)
hist(y, freq=FALSE)
x<-seq(0,7,by=0.05)
y1<-dexp(x,rate=2)
lines(x,y1)
```

  La funció `lines()` afegeix una gràfica (amb línies) a la figura ja existent, resultat de l'histograma.

Altres

Altres exemples de gràfiques que podem realitzar amb l'R són:

- **Diagrama de barres:** serveix per la representació mitjançant barres horitzonals o verticals d'unes dades qualitatives o discretes. Cal donar les dades en forma de taula . Si y és la variable primer farem `x<-table(y)` i després `barplot(x)`

- **Diagrama de sectors:** serveix per la representació mitjançant un gràfic circular dividit en sectors d'unes dades qualitatives o discretes. També entrarem les dades en forma de taula.

```
pie(x)
```

- **Strip Chart:** serveix per la representació d'un núvol de punts en una dimensió.

```
stripchart(y)
```

  Per exemple, `stripchart(iris$Sepal.Length)`, o per espècies
  `stripchart(iris$Sepal.Length~iris$Species)`.

- **Plot:** serveix per la representació d'un núvol de punts en 2D.

```
plot(x,y)
```

  Per exemple, `plot(iris$Sepal.Length, iris$Sepal.Width )`

Nota: Tots aquests gràfics els podem personalitzar: afegir títols, canviar els colors... per conèixer les diferents opcions n'hi ha prou posar un signe d'interrogant davant del nom del gràfic.

2.4 Descripció numèrica de les dades

La descripció numèrica de dades consisteix en resumir conjunts de dades $x_1, \ldots, x_n$ amb pocs números que representin la seva posició, dispersió, etc... **Aquestes mesures només són útils per a dades quantitatives**.

Parlarem de les mesures de posició, els quartils i percentils, les mesures de dispersió, les mesures de forma i els diagrames.

Abans, però les primeres funcions que hem de conèixer són:

```
summary(x)
table(x)
```

Aquestes funcions ens donen, la primera un resum de les mesures més habituals que tot seguit trobareu i la segona, com ja hem vist, ens crea una taula de freqüències de les dades.

2.4.1 Mesures de posició

Les mesures de posició ens indiquen d'alguna manera on és el centre de la mostra, quina és la seva posició. N'hi ha moltes, però les dues realment importants són la mitjana i la mediana.

Mitjana

La **mitjana** es defineix com

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

i com tothom sap és una mesura molt utilitzada. El seu inconvenient és la seva excessiva dependència de les dades extremes o errònies.

Calculem la mitjana de la següent manera

```
mean(x)
```

Mediana

Per a resoldre el problema de la mitjana amb la dependència de les dades extremes, es defineix la **mediana**. Abans de definir-la establim que si $x_1, \ldots, x_n$ indica una col.lecció de dades quantitatives, $x_{(1)}, \ldots, x_{(n)}$ indica la mateixa mostra però ordenada de menor a major. Aleshores, la mediana és la dada del mig en el cas d'una mostra de mida senar. Podem escriure

$$Me = x_{(\frac{n+1}{2})}.$$

Si la mostra és de mida parell, la mediana és la mitjana de les dues dades del mig, és a dir,

$$Me = \frac{1}{2}\{x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})}\}.$$

La instrucció per calcular-la és:

```
median(x)
```

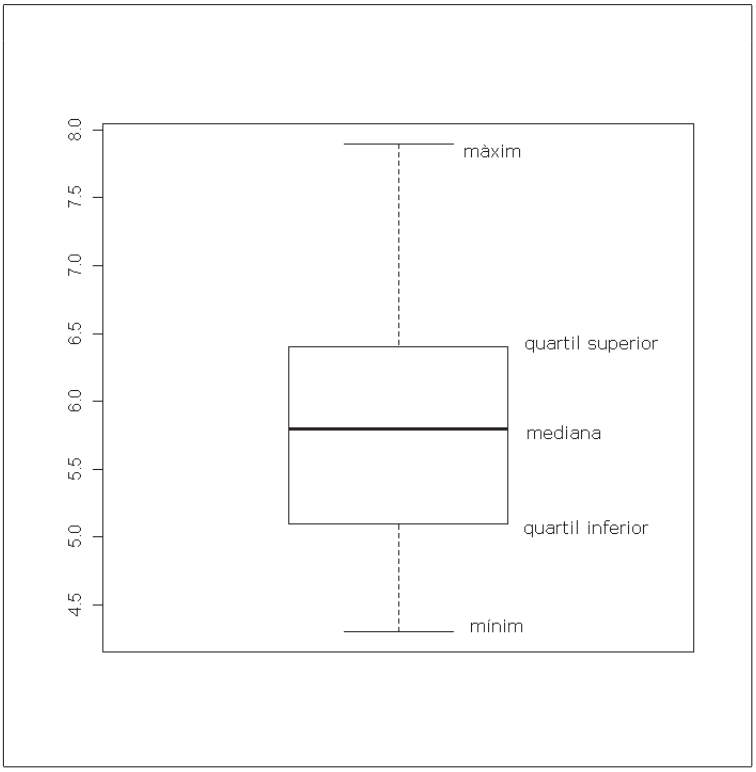
Quartils i percentils

Els quantils d'ordre p són els estadístics que divideixen les dades de manera que deixen el $p\%$ de les dades per sota. Els més utilitzats són:

Quartils. Observem que els quartils divideixen la mostra en quatre parts amb el mateix nombre de dades.

q_0 = mínim de les dades,
q_1 = és el primer quartil, la dada que deixa el 25% del conjunt a l'esquerra i el 75% a la dreta,
q_2 = $Me =$ és el segon quartil (o la mediana), la dada que deixa el 50% a cada costat,
q_3 = és el tercer quartil, la dada que deixa el 75% del conjunt a l'esquerra,
q_4 = màxim de les dades

A partir dels quartils $(q_0, q_1, q_2, q_3, q_4)$, es construeix el **diagrama de caixa** (o boxplot) que dóna una visió gràfica de com es distribueixen les dades.



Amb R utilitzem

```
> quantile(iris$Sepal.Length)
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4  7.9

boxplot(iris)
boxplot(iris$Sepal.Length~iris$Species)
```

Percentils. El percentil d'ordre r ($r = 0, 1, \ldots, 100$) és la dada que deixa el $r\%$ de les dades a l'esquerra i el $(100 - r)\%$ de les dades a la dreta. El primer quartil per exemple és el $25-$percentil.

```
quantile(x, probs=c(0.2,0.6,0.7, 0.9))
```

A més, associat als quartils podem definir una primera mesura de dispersió: el **rang interquartílic** que ens mesura la distància entre el primer i el tercer quartil.

$$IR = q_3 - q_1.$$

2.4.2 Mesures de dispersió

Les mesures de dispersió ens donen informació sobre la variabilitat de les dades entorn a la mitjana:

Variància i variància corregida

La **variància** mostral és:

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

La **variància corregida** és similar a la variància però amb més bones propietats asimptòtiques:

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

Amb `R` utilitzem la funció:

```
var(x)
```

Atenció! aquesta funció retorna la variància corregida $\tilde{s}^2$, amb denominador $n-1$. Si necessitem la variància empírica s, amb denominador n, haurem d'escriure una funció.

Exercici: Escriure una funció per calcular la variància d'un conjunt de dades utilitzant la funció de `R var`. Anomenem varp a aquesta funció.

Desviació estàndar (o típica)

La **desviació típica** s no és més que l'arrel de la variància. Igualment que per la variància, la **desviació típica corregida** és $\tilde{s}$.

```
sd(x)
```

Observació: la desviació típica té les mateixes unitats que les dades, mentre que la variància no.

2.4.3 Mesures de forma

Tenim també dues quantitats que ens ajuden a determinar la forma:

Coeficient d'asimetria o Skewness

Aquesta mesura ens permet identificar si les dades es distribueixen de manera simétrica al voltant de la mitjana.

$$\gamma_3 = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \bar{x})^3.$$

Atenció! Per a aquesta funció i la següent cal, prèviament, carregar el package `e1071`

```
skewness(x)
```

Si el valor d'aquest estadístic és proper a 0 direm que les dades tenen una distribució simètrica, si és positiu les dades tendeixen a l'esquerra de la mitjana i al revés si és negatiu.

Mesura d'apuntament o Curtosi

Aquest estadístic indica el grau d'apuntament de les nostres dades. Quan el valor és negatiu la corba és més plana que una campana de Gauss i quan és positiu és més apuntada.

$$\gamma_4 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3.$$

Per a aquesta funció cal, prèviament, carregar el package `e1071`

```
kurtosis(x)
```

2.5 Dades bivariants

Quan les dades són bivariades, és a dir, tenim dues variables d'interès, podem utilitzar mesures i gráfics per mostrar la relació que existeix entre les variables.

Diagrama de dispersió

Per veure el comportament de les variables i la seva posible relació podem fer un diagrama de dispersió, és a dir, una representació del núvol de punts. Prenent per exemple les dades `iris`, volem explorar la relació entre la longitud del sèpal i del pètal:

```
plot(iris$Sepal.Length,iris$Petal.Length, main="Grafic de dispersio",
xlab="longitud del sepal", ylab="longitud del petal" )
```

Com podem veure, sembla existir una relació lineal.

Covariància i Coeficient de correlació

Una mesura que medeix aquesta dependència entre les variables és la covariància empírica:

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

Amb R utilitzem la funció `cov(x,y)`.

Podem utilitzar també el coeficient de correlació, que és una normalització de la covariància entre -1 i 1, i té una interpretació similar:

$$\rho_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y}.$$

Amb R utilitzem la funció `cor(x,y)`.

Recta de regressió

Els coeficients a b de la recta de regressió $Y = a + bX$ a partir d'unes dades que ja tenim en dos vectors `x` (predictor) i `y` (resposta) s'obtenen fent: `lm(y~x)` En el nostre cas, podem dibuixar-la sobre el diagrama de dispersió fent:

```
x<-lm(iris$Petal.Length~iris$Sepal.Length)
abline(x)
```

Segons la recta de regressió, si el sèpal medeix 8.5mm, el pètal hauria de medir $a + 8.5b = 8.692$.

2.6 Problemes

Aquests exercicis utilitzen bases de dades conegudes que ja estan en el R, com ara LakeHuron, InsectSprays, iris ... Si voleu saber quines són totes les bases de dades feu `data()` i si voleu informació sobre una base de dades, cal que poseu el signe interrogant abans del nom com per exemple

```
?LakeHuron
```

1. Les dades obtingudes per una variable venen resumides en la taula de freqüències següent:

valors	freqüències
0	40
1	52
2	83
3	24
4	12
5	4

   (a) Completeu la taula amb les freqüències absolutes, relatives, absolutes acumulades i relatives acumulades.

   (b) Per aquesta variable calculeu, utilitzant R, la mitjana, la variància, la variància corregida, la desviació estàndard i la desviació estàndard corregida.

2. Per fer aquest exercici utilitzarem les dades del fitxer LakeHuron. Feu primer un diagrama de tija i fulles i un boxplot. Dibuixeu un histograma. Fixeu-vos en el nombre de classes que fa per defecte i intenteu canviar-lo - per exemple fent 5 classes -. Ho fa? Ara podeu provar de posar vosaltres els límits de les classes. Què passa per exemple si poseu

   ```
   hist(LakeHuron, breaks=c(576,577,578,579,580,582))
   ```

   Calculeu el primer quartil i el percentils 20, 45 i 89.

3. Treballarem ara amb les dades de InsectSprays. Feu un boxplot de les variables. És molt més interessant tenir un boxplot per cada marca, això es faria així:

   ```
   boxplot(count~spray, data=InsectSprays)
   ```

   Per la variable `count` feu un histograma.

4. Seguint la idea de la funció varp que us proposen a la pràctica, construïu una funció (sdp) que calculi la desviació estàndard sense corregir.

5. Treballarem ara amb les dades ChickWeight.

   Feu els boxplots del pes(weight) segons la dieta (Diet). Dibuixeu un diagrama de dispersió del pes i els dies d'edat (Time) amb la funció plot.

   Calculeu la recta de regressió de la variable pes respecte l'edat i afegiu-la al gràfic amb la funció

   ```
   abline(a,b)
   ```

   on a és el terme independent (intercept) i b la pendent (slope). Pot ser més pràctic fer directament:

   ```
   abline(lm(pes~edat))
   ```

   Quant pesa un pollet de 9 dies?