Practica 2

Estadística descriptiva

1 Tipus de dades

La primera cosa que necessitem abans de començar a estudiar les dades és saber si aquestes són quantitatives o qualitatives:

- **Dades qualitatives**. Es refereixen a una característica no numèrica de l'individu.
  Tot i no ser variables numèriques, a la pràctica sovint es codifiquen numèricament per facilitar-ne el tractament. En aquest cas els números simplement funcionen com etiquetes assignades a unes categories. Per exemple, el sexe d'un individu es pot codificar amb un 1 o un 0.
  Les variables qualitatives poden ser **nominals** o **ordinals**, és a dir que les etiquetes es poden ordenar. Un exemple d'aquest últim cas són les franges d'edat.

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

2 Taules de freqüències

Per representar les dades utilitzem taules de freqüències. Les instruccions són:

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

La primera ens dona les freqüències, la segona les freqüències relatives, i la tercera les freqüències absolutes acumulades.

Exercici 2. Les dades obtingudes per una variable venen resumides en la taula de freqüències següent:

valors	freqüències
0	40
2	80
4	16
5	4

Calculeu le taules de freqüències absolutes, relatives i acumulades.

3 Representació gràfica de les dades

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.

La instrucció per representar un histograma és la següent:

```
hist(x)
```

Aquesta comanda admet diferents paràmetres optatius:

- Per posar un nombre de classes fixat:

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

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

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

La funció `lines()` afegeix una gràfica (amb línies) a la figura ja existent, resultat de l'histograma. Aquesta opció és útil si volem comprovar en quina mesura l'histograma s'aproxima a la funció de densitat de probabilitat. Per exemple, suposem que hem generat dades d'una llei coneguda i volem comparar-les amb la densitat teòrica:

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

Exercici 3. Considereu les dades `LakeHuron` que té incorporades el programa `R`.

a) Dibuixeu un histograma. Quin és el nombre de classes que fa per defecte?

b) Dibuixeu un histograma canviant el nombre de classes a 10. Us deixa fer el canvi? Si és que no, intenteu justificar el perquè.

c) Dibuixeu un histograma canviant el nombre de classes a 5. Us deixa fer el canvi? Si és que no, intenteu justificar el perquè.

d) Si intentem dibuixar un histograma posant les vores dels rectangles en els punts següents:

$$576 \quad 577 \quad 578 \quad 579 \quad 580 \quad 582$$

l'`R` ens dóna un error. Perquè penseu que no ens ho deixa fer? Dibuixeu un histograma amb uns límits de classe que l'`R` accepti correctament.

Diagrama de barres

El diagrama de barres serveix per la representació mitjançant barres horitzontals o verticals d'unes dades qualitatives o discretes. Cal donar les dades en forma de taula, per tant, si és la variable, primer es calculen

```
x <- table(y)
```

i després es representa el diagrama amb la comanda

```
barplot(x)
```

Diagrama de sectors

El diagrama de sectors serveix per la representació mitjançant un gràfic circular dividit en sectors d'unes dades qualitatives o discretes. També en aquest cas s'entren les dades en forma de taula i després s'executa la comanda

```
pie(x)
```

Plot

La funció genèrica que fa gràfiques és `plot()` i serveix per la representació d'un núvol de punts en 2D:

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

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.

Exercici 4. Considereu les dades proposades a l'Exercici 2:

valors	freqüències
0	40
2	80
4	16
5	4

Representeu-les gràficament amb un diagrama de barres i un diagrama de sectors. Afegiu a cada gràfic el títol i les etiquetes dels eixos corresponents.

4 Descripció numèrica de les dades

Considerem un conjunt de dades $x_1, \ldots, x_n$. La seva descripció numèrica consisteix en un resum de poc estadístics que representin la seva posició, dispersió, etc... Aquestes mesures només són útils per a dades quantitatives.

Un primer resum ens ho dóna la següent instrucció:

```
summary(x)
```

4.1 Mesures de centre

Les mesures de centre ens indiquen on és el centre de la mostra. Les més importants són la mitjana i la mediana.

Mitjana

La mitjana és el valor que s'obté dividint la suma de totes les dades entre el nombre d'observacions:

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

És la mesura que millor ens indica on es troba el centre de les dades, però té com a inconvenient la excessiva dependència de les dades extremes o errònies.
La mitjana es calcula amb la següent comanda:

```
mean(x)
```

Posant l'argument opcional `a`, un nombre entre 0 i 0.5, es produeix la mitjana retallada (*trimmed*), descartant una fracció `a` de dades a cada extrem del vector ordenat:

```
mean(x, trim=a)
```

Per `a` $= 0$ s'obté la mitjana aritmètica ordinària.

També es pot calcular la mitjana ponderada, assignant a cada valor un pes diferent:

```
weighted.mean(x,w)
```

Exercici 5. A l'expedient acadèmic d'un estudiant apareixen les següents dades sobre les assignatures que ha cursat:

Crèdits	Nota
6	6.2
8	7.0
6	7.5
6	9.6
3	5.4
4	8.7
8	8.1
8	6.4
5	5.0
6	7.3

Calculeu la nota mitjana, la nota mitjana del 80% de les assignatures i la mitjana ponderada.

Mediana

La mediana és el valor que queda al centre una vegada ordenat el conjunt de dades de més petit a més gran.

Indiquem amb $x_{(1)}, \ldots, x_{(n)}$ el mateix conjunt de dades ordenades. Si el conjunt és senar, la mediana és la dada del mig:

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

Si el conjunt és parell, la mediana és la mitjana de les dues dades del mig:

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

La instrucció per calcular-la és:

```
median(x)
```

Exercici 6. Calculeu la mediana de les notes de l'exercici anterior.

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 es defineix com la mitjana de les distàncies al quadrat entre les dades i la seva mitjana:

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

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

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

A `R` la funció

```
var(x)
```

retorna la variància corregida.

Exercici 7. Escriviu una funció per calcular la variància d'un conjunt de dades. Anomeneu-la `varp` i utilitzeu-la per calcular la variància de les notes de l'Exercici 5 i comparar-la amb la variància corregida.

Desviació estàndar (o típica)

La desviació típica s és l'arrel de la variància:

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

Igualment que per la variància, la desviació típica corregida és $\widetilde{s} = \sqrt{\widetilde{s}^2}$.

La funció

```
sd(x)
```

retorna la desviació típica corregida.

4.3 Mesures de posició

Les mesures de posició indiquen com es distribueixen les dades a dintre del rang on estan definides.

Percentils

Donat un $p\%$, el p-percentil és el valor numèric tal que almenys un $p\%$ de les dades observades són inferiors o iguals a aquest valor i almenys el $(100-p)\%$ de les dades observades són superiors o iguals a aquest valor.

El procediment per calcular els percentils consisteix en:

- Ordenar les dades de més petita a més gran.
- Buscar el primer amb una freqüència relativa acumulada més gran o igual que $p/100$.

Nota El percentil pot no ser únic.

Exemple 1. Tenim les alçades (en cm) de 10 estudiants:

180 165 168 192 195 187 181 177 175 186.

Primer, les ordenem:

165 168 175 177 180 181 186 187 192 195.

Calculem algun percentil:

- $p = 5$, el percentil és 165. Si calculem el 5% de 10 dades, són 0.5; per tant, hi ha 0.5 dades més petites o iguals a 165 i 9.5 dades més grans o iguals a 165.
- $p = 10$, podem agafar qualsevol valor de l'interval $[165, 168]$.
- $p = 47$, el percentil és 180.

En `R` les instruccions són:

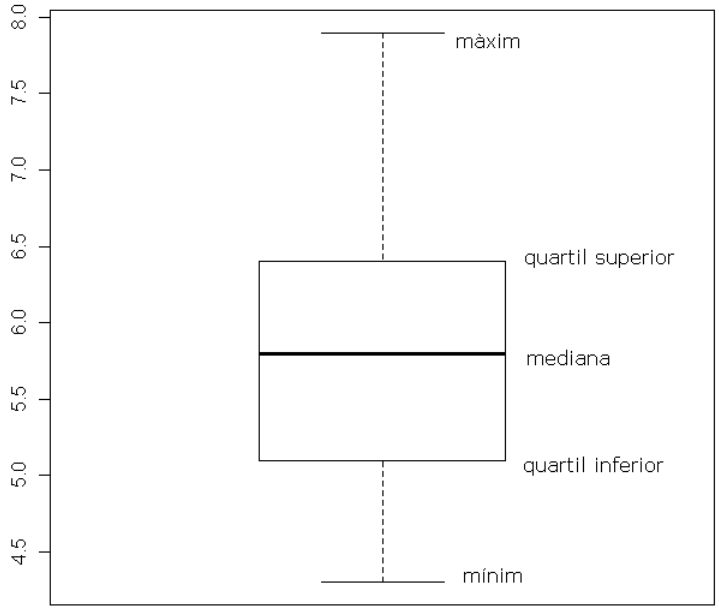
```
quantile(x, probs=c(0.05, 0.1, 0.47))
```

Quartils

Els quartils són els tres punts de tall que divideixen el conjunt de dades en quatre grups de la mateixa mida:

- El percentil per a $p = 25$, s'anomena quartil inferior (Q1).
- El percentil per a $p = 50$ és la mediana (Q2).
- El percentil per a $p = 75$, s'anomena quartil superior (Q3).

A partir dels quartils es construeix el **diagrama de caixa** (o boxplot) que dóna una visió gràfica de com es distribueixen les dades.



Amb R els quartils i el diagrama de caixa s'obtenen amb les següents instruccions:

```
quantile(x)
boxplot(x)
```

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:

$$IR = Q3 - Q1.$$

Exercici 8. Considereu les dades `iris` que té incorporades el programa R. Calculeu els quartils i el rang interquantílic i construïu el diagrama de caixa de la longitud dels sèpals.
Executeu la comanda

```
boxplot(Sepal.Length ~ Species, data = iris)
```

Quina representació obteniu? Podeu afirmar que els iris de la especie setosa tenen els sèpals més petits que els de la especie virgínica?

4.4 Mesures de forma

Les mesures de forma donen informació sobre algunes característiques gràfiques de les dades.

Coeficient d'asimetria o Skewness

El coeficient d'asimetria ens permet identificar si les dades es distribueixen de manera uniforme al voltant de la mitjana aritmètica. Es defineix com

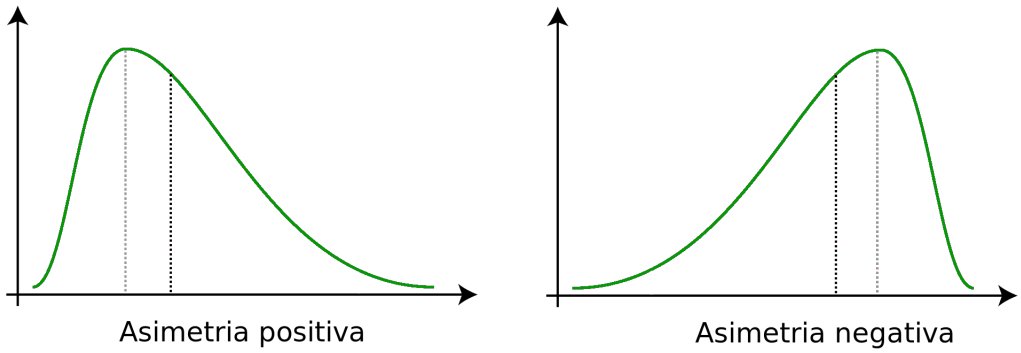
$$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3}.$$

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

A R s'obté amb l'instrucció

```
skewness(x)
```

Per a aquesta funció cal carregar el package `e1071`.



Coeficient d'apuntament o Curtosi

El coeficient de Curtosi indica el grau d'apuntament de les dades. Es defineix com

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

Aquesta mesura només es vàlida per dades simètriques. Quan el seu valor és negatiu la corba és més plana que una campana de Gauss i quan és positiu és més apuntada.

A `R` s'obté amb l'instrucció

```
kurtosis(x)
```

Per a aquesta funció cal carregar el package `e1071`.

Exercici 9. Calculeu els coeficients d'asimetria i d'apuntament de les dades `LakeHuron` incorporades a `R`. Podeu afirmar que les dades són simètriques? Creieu que s'ajusten força bé a una campana de Gauss?

Exercicis

1. Considereu les dades del fitxer `InsectSprays` incorporades a `R`.

   a) Feu un boxplot de les variables.

   b) Feu un boxplot per cada marca, és a dir, feu un boxplot de la variable `count` per cada classe de la variable `spray`. Quines marques han tractat menys insectes?

2. El fitxer `motos.RData` conté les dades de 97 models de motocicletes diferents. Les variables són:

Capacitat	Capacitat del dipòsit de gasolina (en litres).
Consum	Consum cada 100 km (en litres).
CC	Cilindrada (en cm^3).
CV	Potència màxima (en cavalls de vapor).
Pes	Pes de la moto (en kg).
Transmissio	Per cadena o per corretja (*Cadena* o *Corretja*).
Marxes	Canvi de marxes (A = automàtic, M = manual).
Preu	Preu de la moto (en euros).

   Ompliu la taula següent per a la variable `Preu` estratificada per la variable `Marxes`:

Preu	Canvi automàtic	Canvi manual
Mitjana		
Desviació típica		
Mediana		
Percentil 10		
Mínim		
Màxim		

   A partir de la taula, contesteu a les següents preguntes:

   a) Quines motos tenen major variabilitat de preu, les que tenen canvi automàtic o manual?

   b) Quin és el rang de preus per les motos amb canvi manual?

   c) El 90% de les motos amb canvi automàtic tenen un preu més gran de ..... euros.