380 TEXTOS DOCENTS

PROBABILITATS I ESTADÍSTICA

Mireia Besalú
Carles Rovira

Departament de Probabilitat, Lògica
i Estadística

Universitat de Barcelona

PROBABILITATS I ESTADÍSTICA

Mireia Besalú
Carles Rovira

Departament de Probabilitat,
Lògica i Estadística

380 TEXTOS DOCENTS

Universitat de Barcelona
Publicacions i Edicions

Índex

Introducció

Aquest manual recull uns apunts preparats per a fer l'assignatura Probabilitats i Estadística del grau d'Informàtica de la Universitat de Barcelona.

Es tracta d'una assignatura de segon curs, en la qual l'estudiant rep coneixement d'estadística descriptiva, probabilitats –càlcul de probabilitats, Teorema del Límit Central– i inferència estadística –intervals de confiança, contrastos d'hipòtesis, model lineal.

Aquest manual fa referències constants al programari lliure R. Més concretament, al llarg de tots els capítols es van donant les comandes de R per implementar les diverses tècniques presentades.

1. ESTADÍSTICA DESCRIPTIVA

En aquest capítol suposem que tenim unes dades i que volem descriure-les. Això ho farem emprant dos tipus de tècniques: representant-les gràficament i buscant valors numèrics que les resumeixin. També donarem les fórmules i les instruccions amb R.

1.1. Dades univariants

Començarem estudiant una única variable. Tindrem, per tant, n observacions: $x_1, x_2, \ldots, x_n$.

La primera cosa que necessitem saber abans de començar a estudiar les dades és si les dades que tenim 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. 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).
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.

Un cop tenim clar de quin tipus de dades es tracta, podem començar a estudiar-les. I una bona manera és fent l'histograma.

1.1.1. Representació de dades: histogrames

L'**histograma**, el podem utilitzar quan tinguem un nombre prou elevat de dades –per exemple, més de cent–. Vegem com es construeix en el cas de dades quantitatives. Per al cas de dades qualitatives, també es pot adaptar.

Per construir-lo necessitem primer una taula de freqüències.

Per construir la taula de freqüències:

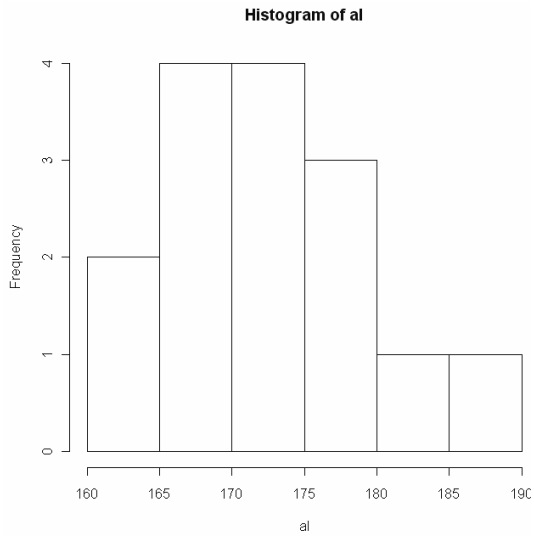
- Agrupem les observacions en classes (intervals). Han de cobrir tot el rang de les dades i normalment s'agafen totes les classes de la mateixa longitud.
- Calculem la freqüència absoluta de cada classe (el nombre d'observacions de cada classe).

Un cop tenim la taula de freqüències només cal que fem un diagrama de barres amb aquests valors.

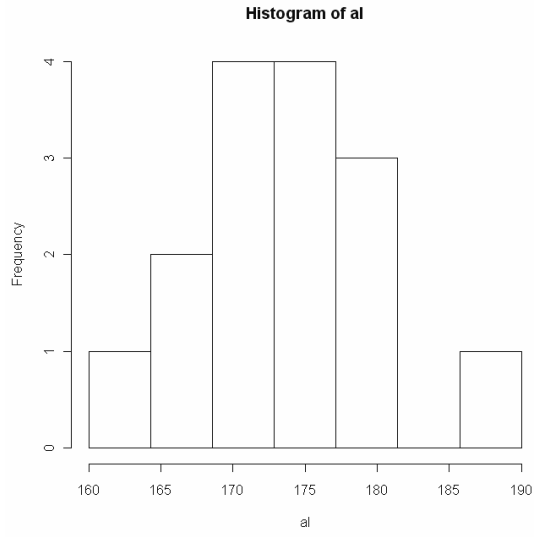
De la mateixa manera que hem calculat les freqüències absolutes de cada classe podríem calcular també la freqüència relativa –la freqüència absoluta dividida per la mida de la mostra–, la freqüència absoluta acumulada –la freqüència absoluta d'aquesta classe més la de totes les anteriors– o la freqüència relativa acumulada –la freqüència relativa d'aquesta classe més la de totes les anteriors. En cada un d'aquests casos obtenim un histograma diferent. L'histograma de freqüències relatives es coneix també com a *histograma de densitat.*

Observeu que el principal problema és com escollir les classes –nombre de classes i mida de cada classe–; aquest problema no té una resposta clara i depèn de cada situació concreta. Donem només una pista: normalment s'agafen com a mínim 6 classes i com a màxim 25. Vegem-ne un exemple.

```
> al<-c(180,170,165,190,170,181,175,174,167,160,177,180,172,169,174)
> al
 [1] 180 170 165 190 170 181 175 174 167 160 177 180 172 169 174
> hist(al, breaks=6)
```




```
> hist(al, breaks=seq(160,190, length=8))
```



La interpretació dels histogrames ens ha de permetre, entre d'altres coses, detectar simetries, pics o l'existència de classes buides que separin la població en grups.

1.1.2. Representació numèrica: estadístics

També podem intentar descriure les dades utilitzant uns valors numèrics, anomenats **estadístics**, que recullen en un sentit la informació que ens dóna la mostra. Aquestes mesures només les podem utilitzar quan treballem amb dades quantitatives i que a més puguin prendre bastants valors diferents. Les més importants són les mesures de centre i les de dispersió.

Mesures de centre

Es tracta de mesures que ens indiquen, en algun sentit, on es troba el centre de les dades. En destaquem la mitjana, la mediana i la moda.

Definició 1. *La* **mitjana** *és el valor que s'obté dividint la suma de totes les dades entre el nombre d'observacions. És a dir,*

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Definició 2. *La* **mediana** *és el valor en què, un cop les dades han estat ordenades de més petita a més gran, la meitat de les dades són més petites o iguals a aquest valor i l'altra meitat són més grans o iguals a aquest valor.*

Procediment per a calcular-la: S'ordenen les dades de la més petita a la més gran. Si el nombre de dades és senar, la mediana és el valor central. Si el nombre de dades és parell, la mediana és qualsevol valor entre les dues dades més properes al centre. En aquest cas, normalment s'agafa la mitjana de les dues dades centrals.

La diferència important entre aquestes dues mesures de centre de les dades és que la mediana és més robusta, és a dir, no es veu tan afectada per valors extrems. Considerem, per exemple, l'alçada de 6 estudiants:

$$180 \quad 187 \quad 181 \quad 177 \quad 175 \quad 186.$$

Podem calcular fàcilment la mitjana, que és $\overline{x} = 181$, i la mediana, que és 180.5. Ara suposem que arriba un estudiant que fa 220 cm d'alçada; aleshores la mitjana passa a ser $\overline{x} = 186.571$, mentre que la mediana passa a ser 181. És a dir, un valor extrem afecta molt més la mitjana que la mediana.

Definició 3. *La* **moda** *és el valor de les dades que es dóna amb més freqüència en la mostra.*

Si continuem amb l'exemple d'abans,

```
> mean(al)
 [1] 173.6
> median(al)
 [1] 174
```

L'R, però, no té cap funció per a calcular la moda d'un conjunt de dades; en tot cas, podríem calcular la taula de freqüències i mirar el valor amb la freqüència absoluta més gran.

```
> table(al)
al
160 165 167 169 170 172 174 175 177 180 181 190
  1   1   1   1   2   1   2   1   1   2   1   1
```

En aquest cas, tindrem 3 modes diferents: 170, 174 i 180.

Mesures de dispersió

Un cop coneixem on és el centre de les dades, també és important conèixer el grau de dispersió respecte del centre que hem calculat. Imagineu que els sous mensuals dels treballadors d'una empresa A són: 1400, 1600 i 1800, mentre que els de l'empresa B són 400, 600 i 3800. El sou mitjà a les dues empreses és el mateix, però, com veieu, la situació no és la mateixa!

Definició 4. *La* **variància mostral** *es calcula com la mitjana de les distàncies al quadrat entre les dades i la seva mitjana. És a dir,*

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

Per qüestions tècniques, normalment s'utilitza el que s'anomena *variància mostral corregida*:

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

El problema que observem és que les unitats de la variància mostral no són les mateixes unitats que les de les dades inicials, sinó que estan elevades al quadrat. Per poder treballar amb les mateixes unitats utilitzem l'anomenada *desviació típica mostral.*

Definició 5. *La* **desviació típica mostral** *es defineix com l'arrel quadrada de la variància mostral. És a dir,*

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

Podem calcular aquests valors per a les empreses A i B, i obtenim:

```
> A<-c(1400,1600,1800)
> B<-c(400,600,3800)
> mean(A)
[1] 1600
> mean(B)
[1] 1600
```

```
> (3-1)/3*var(A)
[1] 26666.67 > (3-1)/3*var(B)
[1] 2426667
> sqrt((3-1)/3*var(A))
[1] 163.2993
> sqrt((3-1)/3*var(B))
[1] 1557.776
```

$$\overline{x}_A = 1600 \quad s_A^2 = 26666.66 \quad s_A = 163.29$$
$$\overline{x}_B = 1600 \quad s_B^2 = 2426667 \quad s_B = 1557.77.$$

Observem que l'R ens retorna la variància corregida; és per això que la multipliquem per $\frac{n-1}{n}$.

Com ja hem comentat, el sou mitjà a les dues empreses és el mateix, però els sous a l'empresa B tenen una dispersió molt més gran que a l'empresa A –aquest fet es veu tant en la variància com en la desviació típica.

Hi ha altres mesures de dispersió. Necessitem, però, introduir primer el concepte de *percentil*.

Mesures de posició

Ens referirem bàsicament als percentils. Es tracta de donar la dada que ocupa una determinada posició.

Definició 6. *Donat un tant per cent determinat* p%*, el seu* **percentil** *és el valor* x *tal que almenys el* p% *de les dades observades són inferiors o iguals a* x *i almenys el (100*− p)% *de les dades observades són superiors o iguals a* x.

Vegem-ne un exemple. Imaginem que tenim les alçades (en cm) de 10 estudiants:

180 165 168 192 195 187 181 177 175 186.

El primer que hem de fer és ordenar-les:

165 168 175 177 180 181 186 187 192 195.

Ara ja podem calcular percentils:

- $p = 5$, el percentil és 165 –si calculem el 5% de 10 dades, són 0.5; per tant, veiem que 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.

```
> al2<-c(180,165,168,192,195,187,181,177,175,186)
> al2
 [1] 180 165 168 192 195 187 181 177 175 186
```

```
> quantile(al2,0.05)
  5%
166.35
> quantile(al2,0.1)
  10%
167.7
> quantile(al2,0.47)
  47%
180.23
```

Tingueu en compte que aquesta definició fa que els percentils no estiguin unívocament determinats; és a dir, que, donat un valor p, hi pot haver més d'un percentil. Això fa que segons el programari estadístic que utilitzem obtinguem valors diferents. De totes maneres, quan hi hagi moltes dades i pocs forats entre aquestes dades, els resultats que obtindrem seran sempre molt semblants.

Observeu que quan agafem $p = 50$ recuperem la definició de la mediana. Hi ha altres percentils que també tenen nom propi.

Definició 7. *El percentil per a* p $=25$ *s'anomena* **quartil inferior**, *i correspon al valor numèric tal que almenys el 25% de les dades observades són inferiors o iguals a aquest valor numèric i almenys el 75% de les dades observades són superiors o iguals a aquest valor numèric. El percentil per a* p $=75$ *s'anomena* **quartil superior**, *i correspon al valor numèric tal que almenys el 75% de les dades observades són inferiors o iguals a aquest valor numèric i almenys el 25% de les dades observades són superiors o iguals a aquest valor numèric.*

Si seguim amb l'exemple anterior,

```
> median(al2)
[1] 180.5
```

Un cop tenim definits els quartils inferiors i superiors, podem definir una nova mesura de dispersió, que és el *rang interquartílic*.

Definició 8. *El* **rang interquartílic** *és la diferència entre el quartil superior i el quartil inferior.*

Entre el quartil inferior i el quartil superior tindrem com a mínim el 50% de les observacions centrals de la nostra mostra. Així, quan el rang interquartílic sigui petit indicarà que les dades centrals estan poc disperses.

```
> IQR(al2)
[1] 11.25
```

Mesures de forma

Hi ha encara altres mesures que ens expliquen com són les dades o, el que és el mateix, com és la població d'on provenen aquestes dades. No entrarem aquí en l'estudi d'aquests valors.

En citarem només dos com a cultura general.

1) **Coeficient d'asimetria** o *skewness*. Quan les dades són simètriques, aquest coeficient val 0 i com més asimètriques són més gran és el seu valor. Es calcula de la manera següent:

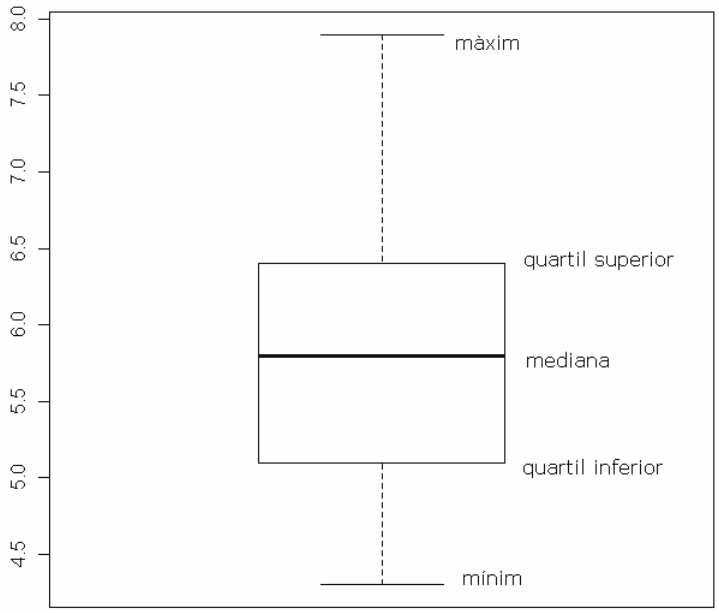
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{s^3}.$$

2) **Coeficient de curtosi** o **mesura d'apuntament**. 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. Es calcula com

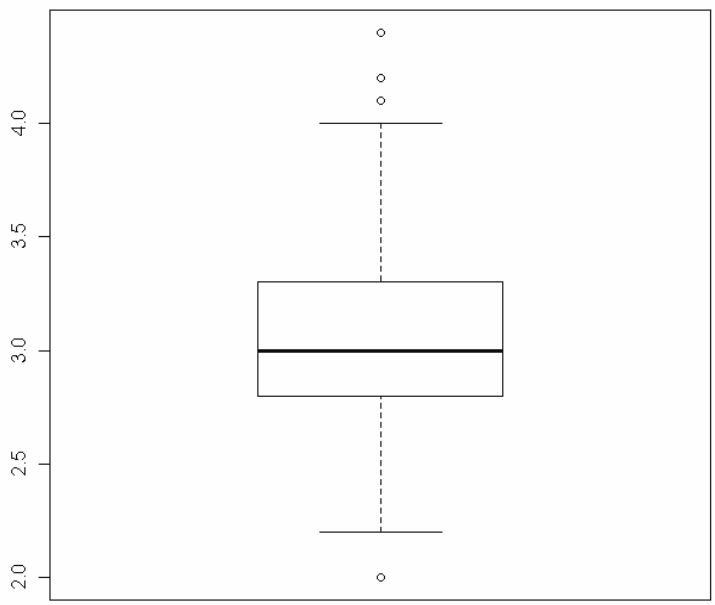
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{s^4}-3.$$

1.1.3. Altres mètodes de representació gràfica: el diagrama de caixa

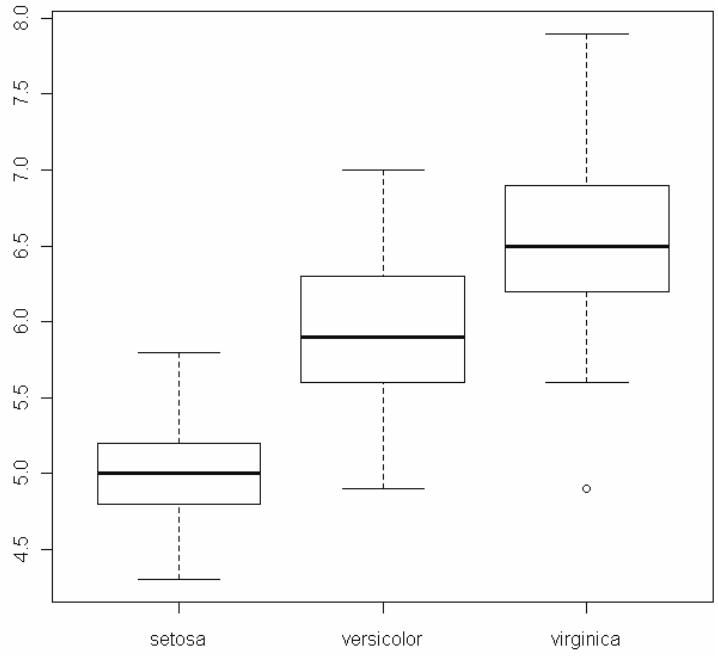
Un mètode de representació gràfica força utilitzat és l'anomenat **diagrama de caixa** o *box-plot*. És un diagrama en el qual es representen 5 nombres que ens donen bastanta informació de com són les dades: el mínim (l'inici de la semirecta), el quartil inferior (l'inici de la caixa), la mediana (la ratlla que parteix la caixa), el quartil superior (el final de la caixa) i el màxim (el final de la semirecta).



Quan hi ha valors que considerem estranys, és a dir, que són molt diferents de la resta, aquests valors s'especifiquen explícitament fora del *box-plot*. Vegem-ho a l'exemple següent:



Aquest tipus de gràfic és força útil per a detectar simetries i per a comparar dues o més variables.



1.2. Dades bivariants

També parlarem breument de com estudiar les dades quan s'observa més d'una variable alhora. És clar que podem estudiar cada variable separadament, però moltes vegades ens aporta més informació estudiar-les conjuntament, així podem intentar establir si hi ha alguna relació entre elles.

Ens limitarem al cas més senzill, quan tenim només dues variables (cas bivariant). Suposarem, per tant, que tenim n observacions $(x_i, y_i), i = 1, \ldots, n$ d'una parella de variables X, Y.

1.2.1. Taules creuades

Una taula creuada és una taula de freqüències de parelles de valors. La podem utilitzar tant per a variables qualitatives com quantitatives, com per a combinacions de les dues, encara que quan apareguin variables quantitatives contínues els valors s'hauran d'agrupar per intervals –de manera semblant a com ho fèiem per als histogrames univariants.

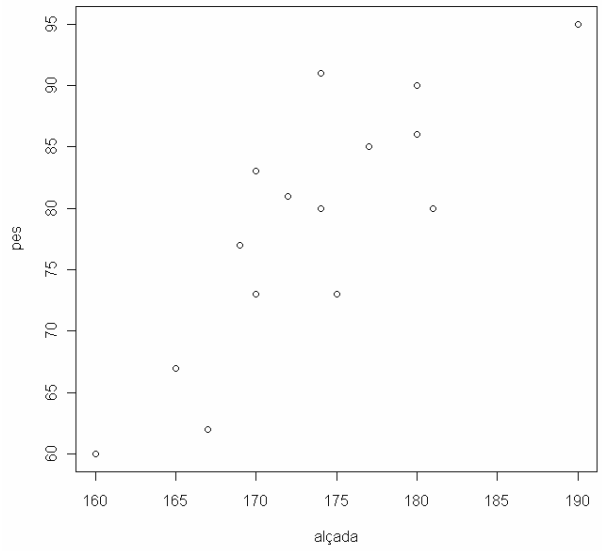
Exemple 1. Imagineu que tenim la variable X sexe –que és qualitativa– i la variable Y alçada –que és quantitativa–. Aleshores, agrupant la variable Y en intervals (menys de 150, entre 150 i 160, entre 160 i 170, entre 170 i 180, entre 180 i 190, més de 190), podem construir una taula creuada del tipus:

	Home	**Dona**
Menys de 150	n_{11}	n_{12}
Entre 150 i 160	n_{21}	n_{22}
Entre 160 i 170	n_{31}	n_{32}
Entre 170 i 180	n_{41}	n_{42}
Entre 180 i 190	n_{51}	n_{52}
Més de 190	n_{61}	n_{62}

1.2.2. Representacions gràfiques

Hi ha diversos tipus de diagrames per a representar dues variables alhora. En aquest cas, l'objectiu és que el diagrama ens permeti o bé entendre la relació entre les dues variables –podem utilitzar un diagrama de dispersió– o bé comparar les dues variables –per exemple, amb un diagrama de barres múltiple o amb un diagrama de caixa múltiple.

```
> al<-c(180,170,165,190,170,181,175,174,167,160,177,180,172,169,174)
> pes<-c(90,73,67,95,83,80,73,80,62,60,85,86,81,77,91)
> plot(al,pes, xlab="alçada", ylab="pes")
```



1.2.3. Mesures de dependència

Hi ha dos valors que ens mesuren el grau de dependència lineal entre dues variables: la *covariància mostral* i el *coeficient de correlació lineal*, però només ens serveixen per a treballar amb variables quantitatives.

Definició 9. *La* **covariància mostral** *s'obté com la mitjana dels productes de les desviacions de les dades respecte de la seva mitjana. És a dir:*

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

Una covariància mostral positiva ens indica que com més grans són els valors de la primera variable més grans són també els de la segona. Quan és negativa ens indica el contrari, és a dir, que com més grans són els valors de la variable X més petits són els de la variable Y.

```
> al<-c(180,170,165,190,170,181,175,174,167,160,177,180,172,169,174)
> pes<-c(90,73,67,95,83,80,73,80,62,60,85,86,81,77,91)
> cov(al,pes)
[1] 63.51429
```

La covariància mostral té el problema que el seu valor depèn de les unitats de mesura de les variables X i Y. Per exemple, les mateixes variables –alçada i pes– amb els mateixos valors però canviant la unitat de mesura de la variable alçada de metres a centímetres ens multiplicarà per 100 el valor de la covariància mostral. Per solucionar aquest problema, es defineix el *coeficient de correlació lineal.*

```
> al2<-al*0.01
> al2
 [1] 1.80 1.70 1.65 1.90 1.70 1.81 1.75 1.74 1.67 1.60 1.77 1.80 1.72
1.69 1.74
> cov(al2,pes)
[1] 0.6351429
```

Definició 10. *El* **coeficient de correlació lineal** *entre dues variables* X *i* Y *es defineix com la covariància mostral entre les dues variables dividit pel producte de les seves desviacions típiques mostrals. És a dir:*

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

```
> cor(al,pes)
[1] 0.8271535
> cor(al2,pes)
[1] 0.8271535
```

El gran avantatge d'aquest coeficient és que no depèn de les unitats de mesura. Sempre pren valors entre -1 i 1. Quan valgui 1 indicarà que hi ha una relació lineal del tipus $y = a\,x + b$ i quan sigui -1, una relació del tipus $y = -a\,x + b$. Si $r_{x,y}$ és, en valor absolut, més gran que 0.8 s'entén que la relació entre les dues variables és gran. Si $r_{x,y}$ és, en valor absolut, més petit que 0.8 s'entén que la relació és feble. Quan val 0 és quan no hi ha cap relació entre les dues variables.