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 7

1. Estadística descriptiva 9

1.1. Dades univariants 9

1.1.1. Representació de dades: histogrames 9

1.1.2. Representació numèrica: estadístics 11

1.1.3. Altres mètodes de representació gràfica: el diagrama de caixa 15

1.2. Dades bivariants 16

1.2.1. Taules creuades 17

1.2.2. Representacions gràfiques 17

1.2.3. Mesures de dependència 18

2. Probabilitats 21

2.1. Combinatòria 21

2.2. Espais de probabilitat 24

2.2.1. Esdeveniments 25

2.2.2. Espai de probabilitat finit 26

2.2.3. Propietats de la probabilitat 27

2.3. Probabilitat condicionada 28

2.4. Esdeveniments independents 29

2.5. Fórmula de les probabilitats totals 30

2.6. Fórmula de Bayes 31

3. Variables aleatòries 35

3.1. Variable aleatòria: primers conceptes 35

3.2. Variable aleatòries discretes 37

3.2.1. Independència de variables aleatòries discretes 41

3.2.2. Esperança i variància de variables aleatòries discretes 42

3.2.3. Exemples de variables aleatòries discretes 44

3.3. Variables aleatòries absolutament contínues 48

3.3.1. Independència de variables absolutament contínues 51

3.3.2. Esperança i variància de variables aleatòries absolutament contínues 51

3.3.3. Exemples de variables aleatòries absolutament contínues 52

3.3.4. La distribució normal 55

Universitat de Barcelona. Dades catalogràfiques

Besalú, Mireia
Probabilitats i estadística. – (Textos docents ; 380)

A la portada: Departament de Probabilitat, Lògica i Estadística
Bibliografia
ISBN 978-84-475-3653-5

I. Rovira Escofet, Carles
II. Universitat de Barcelona. Departament de Probabilitat, Lògica i Estadística
III. Títol. IV. Col·lecció: Textos docents (Universitat de Barcelona) ; 380
1. Probabilitats 2. Estadística matemàtica

© Publicacions i Edicions de la Universitat de Barcelona
Adolf Florensa, s/n
08028 Barcelona
Tel. 934 035 530
Fax 934 035 531
www.publicacions.ub.edu
comercial.edicions@ub.edu

ISBN	978-84-475-3653-5
Dipòsit Legal	B-2.683-2013
Impressió i enquadernació	Gráficas Rey

És rigorosament prohibida la reproducció total o parcial d'aquesta obra. Cap part d'aquesta publicació, inclòs el disseny de la coberta, no pot ser reproduïda, emmagatzemada, transmesa o utilitzada per cap mitjà o sistema, sense l'autorització prèvia per escrit de l'editor.

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.

4. Teorema del límit central ... 59

4.1. La mitjana mostral ... 59

4.1.1. Mitjana mostral per a variables normals ... 60

4.2. Teorema del límit central ... 62

4.2.1. Cas general ... 62

4.2.2. Cas binomial ... 63

5. Intervals de confiança ... 65

5.1. Exemple ... 65

5.2. Intervals de confiança per a una població normal ... 67

5.2.1. Intervals de confiança per a μ, on σ és coneguda ... 67

5.2.2. Intervals de confiança per a μ, on σ és desconeguda ... 68

5.2.3. Interval de confiança per a la variància σ^2 ... 69

5.3. Intervals de confiança per a la proporció ... 71

5.4. Intervals de confiança per a mostres grans ... 73

5.5. Intervals de confiança per a dues mostres ... 74

5.5.1. Intervals de confiança per a dues mostres de dades aparellades ... 74

5.5.2. Intervals de confiança per a dues mostres de dades no aparellades ... 75

6. Contrastos d'hipòtesis ... 83

6.1. Teoria general del contrast d'hipòtesis ... 83

6.2. Contrastos per a una població normal ... 86

6.2.1. Contrast sobre la mitjana quan la variància és coneguda ... 86

6.2.2. Contrast sobre la mitjana quan la variància és desconeguda ... 89

6.2.3. Contrast sobre la variància ... 91

6.3. Contrast sobre una proporció ... 94

6.4. Contrast sobre la mitjana per a mostres grans ... 96

6.5. Contrastos per a dues mostres normals ... 98

6.5.1. Contrast sobre la diferència de mitjanes $\mu_x - \mu_y$, cas de variàncies σ_x^2 i σ_y^2 conegudes ... 98

6.5.2. Contrast sobre la diferència de mitjanes $\mu_x - \mu_y$, cas de variàncies σ_x^2 i σ_y^2 desconegudes però iguals ... 100

6.5.3. Contrast sobre la raó de variàncies ... 102

7. El model de regressió ... 107

7.1. Càlcul de la recta de regressió ... 108

7.1.1. Mètode dels mínims quadrats ... 109

7.2. Qualitat de l'ajustament ... 112

7.2.1. Relació amb el coeficient de correlació lineal ... 113

7.2.2. Anàlisi dels residus ... 114

7.3. Inferència en el model de regressió ... 115

7.4. Altres models ... 115

Bibliografia ... 119

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:

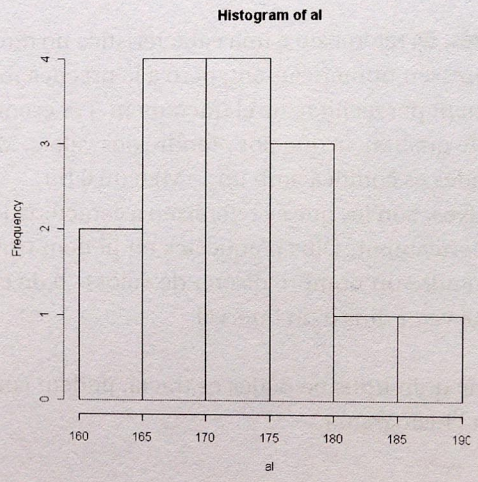
- Agrupem les observacions en classes (intervals). Han de cobrir tot el rang de les dades i normalment s'agafen totes les classes de la mateixa longitud.
- Calculem la freqüència absoluta de cada classe (el nombre d'observacions de cada classe).

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

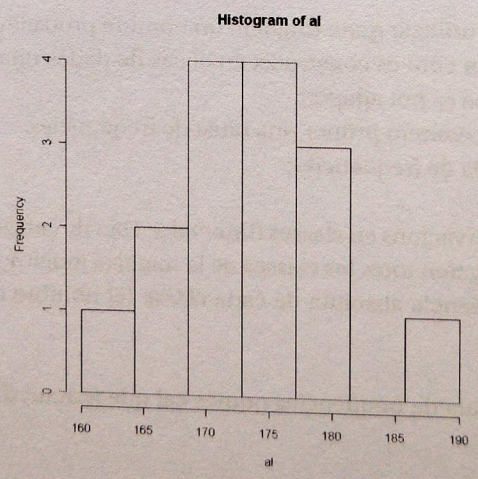
De la mateixa manera que hem calculat les freqüències absolutes de cada classe podríem calcular també la freqüència relativa –la freqüència absoluta dividida per la mida de la mostra–, la freqüència absoluta acumulada –la freqüència absoluta d'aquesta classe més la de totes les anteriors– o la freqüència 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:

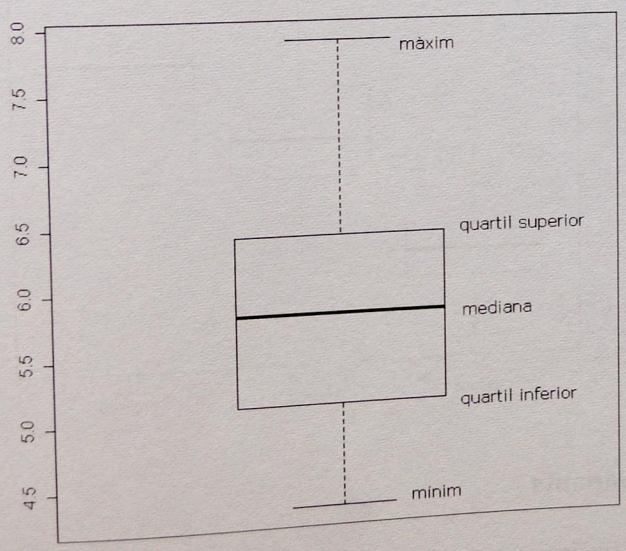
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{s^3}.$$

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

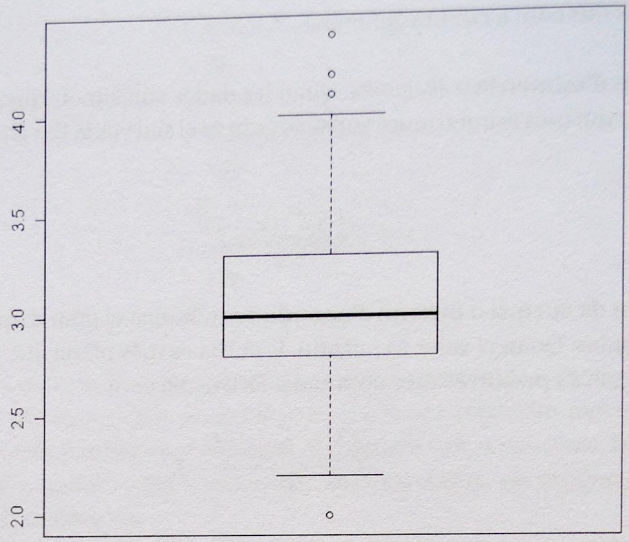
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^4}{s^4}-3.$$

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

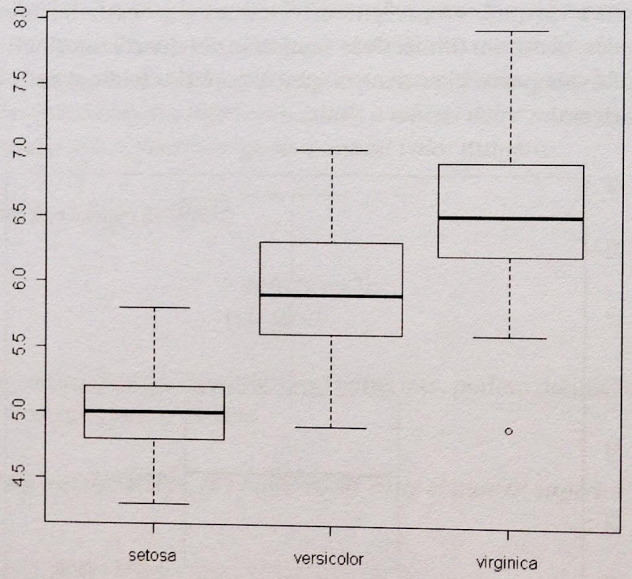
Un mètode de representació gràfica força utilitzat és l'anomenat **diagrama de caixa** o *box-plot*. És un diagrama en el qual es representen 5 nombres que ens donen bastanta informació de com són les dades: el mínim (l'inici de la semirecta), el quartil inferior (l'inici de la caixa), la mediana (la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

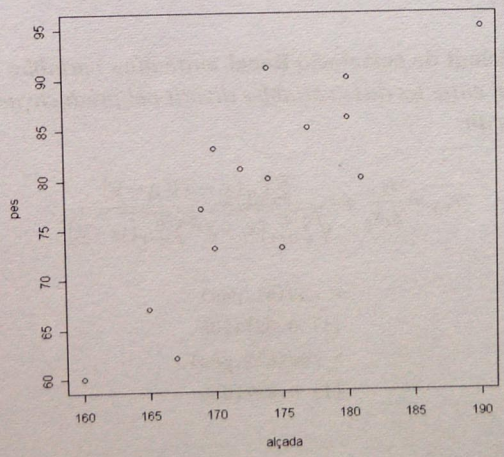
Exemple 1. Imagineu que tenim la variable X sexe –que és qualitativa– i la variable Y alçada –que és quantitativa–. Aleshores, agrupant la variable Y en intervals (menys de 150, entre 150 i 160, entre 160 i 170, entre 170 i 180, entre 180 i 190, més de 190), podem construir una taula creuada del tipus:

	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 $y = ax + b$ i quan sigui -1, una relació del tipus $y = -ax + b$. 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.

2. PROBABILITATS

El càlcul de probabilitats és l'estudi matemàtic de fenòmens que depenen de l'atzar i la incertesa. Són fenòmens dels quals coneixem els possibles resultats que es poden donar, però, durant el seu instant inicial, no podem predir exactament quin dels possibles resultats es produirà. Aquests fenòmens els anomenarem *experiments aleatoris.*

Un **experiment aleatori** serà aquell experiment que, repetit en les mateixes condicions, pot donar resultats diferents, coneguts, i per al qual no podem predir quin d'aquests resultats es produirà realment.

Exemple 2. *El resultat de la cara superior en la tirada d'un dau.*

Exemple 3. *El nombre d'encerts en una travessa de 15 partits.*

Exemple 4. *El resultat obtingut en la tirada de tres monedes diferents al mateix temps.*

Exemple 5. *El resultat obtingut en la tirada al mateix temps d'una moneda i d'un dau.*

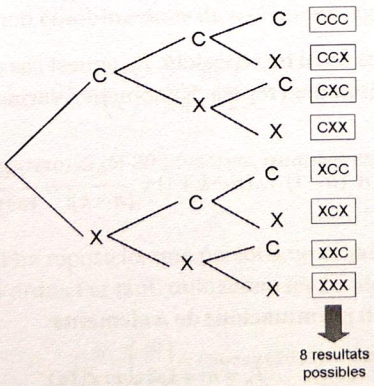
Exemple 6. *El nombre de visites que rep una pàgina web durant un dia.*

Abans de començar a parlar de probabilitats, necessitarem conèixer algunes tècniques de recompte per tal de poder calcular el nombre de resultats possibles d'un experiment aleatori.

2.1. Combinatòria

Ens proposem ara fer un recompte de possibles resultats. Per exemple:

Exemple 7. *Tenim 3 monedes diferents, i volem saber quants resultats diferents podem obtenir si les tirem totes tres a la vegada.*



Observem que en total tenim 8 possibles resultats. Això també ho hauríem pogut calcular fent

$$2 \cdot 2 \cdot 2 = 8.$$

Aquest exemple anterior era fàcil i fins i tot podíem escriure tots els casos, però necessitem algunes eines per si hem de fer recomptes d'experiments més complicats. Per tant, la primera tècnica que necessitem entendre és la **regla multiplicativa**: suposem que hem de realitzar k tries en les condicions següents:

- La primera tria té m_1 possibilitats,
- La segona tria té m_2 possibilitats,
- La tercera tria té m_3 possibilitats,
- ...,
- La k-èsima tria té m_k possibilitats.

Aleshores, observem que el nombre total de possibilitats és

$$m_1 \cdot m_2 \cdot m_2 \cdot \ldots \cdot m_k.$$

Exemple 8. *Quantes matrícules diferents de quatre números hi haurà que tinguin totes les xifres diferents?*

$$m_1 = 10,\ m_2 = 9,\ m_3 = 8,\ m_4 = 7.$$

Per tant, hi haurà $10 \cdot 9 \cdot 8 \cdot 7 = 5040$ matrícules diferents que tinguin totes les xifres diferents.

Exemple 9. *I quantes matrícules diferents de quatre números hi haurà que tinguin les dues primeres xifres iguals i les altres dues diferents (entre elles i de les primeres)?*

$$m_1 = 10,\ m_2 = 1,\ m_3 = 9,\ m_4 = 8.$$

Per tant, hi haurà $10 \cdot 1 \cdot 9 \cdot 8 = 720$ matrícules diferents.

Suposarem a partir d'ara que tenim un conjunt de n elements diferents i que en volem escollir k. Ens plantejarem diferents maneres de fer aquesta tria. Bàsicament, abans de cada extracció haurem de respondre dues preguntes: ens importa l'ordre? (és el mateix «AB» que «BA»?) i hi ha reposició? (es poden repetir els elements?). Segons les respostes que donem, tenim els diferents casos.

1) **Importa l'ordre però no hi ha reposició.** En aquest cas anem escollint els elements per ordre, però no els podem repetir. S'anomenen **variacions de *n* elements agafats de *k* en *k***:

$$V_k^n = n \cdot (n-1) \cdot \ldots (n-k+1) = \frac{n!}{(n-k)!} = \frac{\texttt{factorial(n)}}{\texttt{factorial(n-k)}}.$$

CAS PARTICULAR

Si $n = k$, s'anomenen **permutacions de *n* elements**:

$$P_n = n! = \texttt{factorial(n)}.$$

Exemple 10. *En un departament de la universitat s'ha de triar un director i un secretari. Si hi ha 20 professors que poden optar als càrrecs, quants resultats diferents es poden donar?*

Observem que ens importa l'ordre; per exemple, escollim primer el director i segon el secretari, i en canvi no hi ha reposició: la mateixa persona no pot tenir els dos càrrecs. Per tant, hem d'utilitzar variacions:

$$V_2^{20} = \frac{20!}{(20-2)!} = \frac{\texttt{factorial(20)}}{\texttt{factorial(20-2)}} = 380.$$

Exemple 11. *De quantes maneres diferents podem ordenar les vocals?*

Igual que abans, ens importa l'ordre i no podem repetir les vocals, però observem que, a més, hem d'utilitzar tots els elements. Per tant, en aquest cas seran permutacions:

$$P_5 = 5! = \texttt{factorial(5)} = 120.$$

2) **Importa l'ordre i hi ha reposició.** Ara escollim els elements per ordre i a cada tria podem repetir els elements que han sortit anteriorment. S'anomenen **variacions amb repetició de *n* elements agafats de *k* en *k***:

$$VR_k^n = n^k.$$

Exemple 12. *Tirem un dau tres vegades, i després de cada llençament anotem el resultat. Quants resultats diferents podem obtenir?*

Ara es poden repetir els resultats: en cada llençament ens pot sortir qualsevol número, i ens importa l'ordre. Per tant, utilitzarem variacions amb repetició:

$$VR_3^6 = 6^3 = 216.$$

3) **No importa l'ordre ni hi ha reposició.** En aquesta situació no podem repetir els resultats obtinguts anteriorment i dos conjunts amb els mateixos elements però amb ordres diferents els considerarem iguals. De fet, és com si escollíssim d'un sol cop k elements. S'anomenen **combinacions de *n* elements agafats de *k* en *k***:

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \texttt{choose(n,k)}.$$

Exemple 13. *En una reunió de 20 persones, quantes encaixades de mans es poden donar?*

Ens fixem que no hi ha reposició: una persona no es pot donar la mà amb ella mateixa, i no importa l'ordre. Per tant, utilitzarem les combinacions:

$$C_2^{20} = \binom{20}{2} = \texttt{choose(20,2)} = 190.$$

4) **No importa l'ordre però hi ha reposició.** Aquest cas no el tractarem en aquest curs de probabilitats, però en donem la fórmula per cultura general. Hem de triar els elements de manera que no ens importa si es repeteixen, però els mateixos conjunts d'elements amb ordres diferents els considerem iguals. El problema, per tant, consisteix a trobar el nombre de conjunts no ordenats que puc formar repetint elements. S'anomenen **combinacions amb repetició de n elements agafats de k en k**:

$$CR_k^n = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!} = \texttt{choose(n+k-1,k)}.$$

Exemple 14. *Tinc 5 boles numerades de l'1 al 5, i en trec 3 amb reposició. Quantes combinacions puc fer si no m'importa l'ordre amb què les he tret?*

Ens fixem que hi ha reposició i que no importa l'ordre. Per tant, utilitzarem les combinacions amb repetició:

$$CR_3^5 = \binom{7}{3} = \texttt{choose(7,3)} = 35.$$

Un cop coneixem algunes de les tècniques de recompte, podem continuar parlant dels experiments aleatoris i d'alguns conceptes relacionats que necessitarem.

2.2. Espais de probabilitat

El primer concepte que necessitem definir és l'*espai mostral.*

Definició 11. *L'***espai mostral** Ω *es defineix com el conjunt de tots els resultats possibles d'un experiment aleatori.*

Mirem ara quins són els espais mostrals dels exemples que hem vist abans.

- Espai mostral de l'exemple 2: *Resultat de la cara superior en la tirada d'un dau.*

$$\Omega = \{1,2,3,4,5,6\}.$$

- Espai mostral de l'exemple 3: *El nombre d'encerts en una travessa de 15 partits.*

$$\Omega = \{0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15\}.$$

- Espai mostral de l'exemple 4: *El resultat obtingut en la tirada de tres monedes diferents al mateix temps.*

$$\Omega = \{(C,C,C),(C,C,X),(C,X,C),(C,X,X),(X,C,C),(X,C,X),(X,X,C),(X,X,X)\}.$$

- Espai mostral de l'exemple 5: *El resultat obtingut en la tirada al mateix temps d'una moneda i d'un dau.*

$$\Omega = \{(C,1),(C,2),(C,3),(C,4),(C,5),(C,6),(X,1),(X,2),(X,3),(X,4),(X,5),(X,6)\}.$$

- Espai mostral de l'exemple 6: *El nombre de visites que rep una pàgina web durant un dia.*

$$\Omega = \{0, 1, 2, 3, 4, \ldots\}.$$

Fixeu-vos que els espais mostrals poden ser molt diferents segons quin sigui l'experiment aleatori que estigueu tractant. Pot ser que siguin numèrics o que no ho siguin i també pot ser que siguin finits o infinits.

A la pràctica, no sempre estarem interessats en tot l'espai mostral, sinó en uns quants dels possibles resultats que formen part de l'espai mostral. És per això que ara necessitem definir què és un *esdeveniment* o *succés*.

2.2.1. Esdeveniments

Definició 12. *Un* **esdeveniment** *o* **succés** *és un subconjunt de l'espai mostral* Ω.

Exemple 15. A partir de l'experiment de l'exemple 4 (el resultat obtingut en la tirada de tres monedes diferents al mateix temps), podem considerar els següents conjunts que són exemples d'esdeveniments:

$$\begin{aligned} A &= \{\text{El resultat de la moneda 1 és cara}\}, \\ B &= \{\text{Ens han sortit exactament dues cares}\}, \\ D &= \{\text{El resultat de la moneda 1 és creu}\}, \\ E &= \{\text{Ens ha sortit com a mínim una creu}\}. \end{aligned}$$

Hi ha alguns esdeveniments que s'utilitzen més i que són importants de conèixer. Dos d'aquests són:

- El **succés segur** és el format per tots els resultats possibles, o sigui Ω.
- El **succés impossible** és el que no es dóna mai, o sigui $\emptyset$.

Ara bé, donats dos esdeveniments A i B, podem definir diverses **operacions** entre ells:

- L'esdeveniment que es dóna quan passa A o quan passa B s'anomena A **unió** B, i ho escriurem $A \cup B$.
- L'esdeveniment que es dóna quan passa A i al mateix temps passa B s'anomena A **intersecció** B, i ho escriurem $A \cap B$.
- L'esdeveniment que es dóna quan no passa A s'anomena el **complementari** d'A, i ho escriurem A^c.

I finalment, tenim les relacions següents:

- Dos esdeveniments A i B són **disjunts** si no tenen resultats en comú, o sigui $A \cap B = \emptyset$.
- Dos esdeveniments A i B són **exhaustius** si entre els dos tenim tots els resultats possibles, o sigui $A \cup B = \Omega$.
- Si tots els resultats de B estan inclosos en A, es diu que B **està inclòs** en A, i escriurem $B \subset A$.

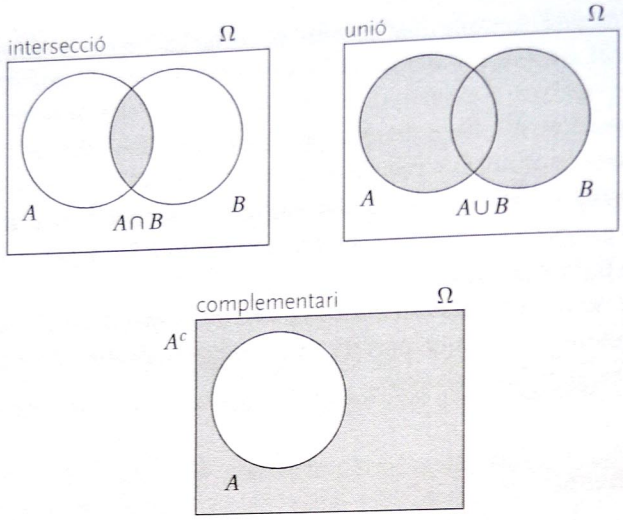


Figura 2.1. Operacions entre esdeveniments.

Observem que hi ha esdeveniments que poden complir les dues condicions: poden ser exhaustius i a l'hora ser disjunts. L'exemple més fàcil d'esdeveniments que compleixen les dues condicions és: un esdeveniment A i el seu complementari A^c.

Observació 1. Com que els esdeveniments són subconjunts de l'espai mostral, les operacions que hem definit són les que podem fer habitualment entre subconjunts. I per tant, podem utilitzar el llenguatge conjuntista sempre que treballem amb els esdeveniments.

Exemple 16. *Recuperem els esdeveniments que hem definit a l'exemple 15; així, doncs, tenim que*

$$\begin{aligned}
A &= \{(C,C,C),(C,C,X),(C,X,C),(C,X,X)\},\\
B &= \{(C,C,X),(C,X,C),(X,C,C)\},\\
D &= \{(X,C,C),(X,C,X),(X,X,C),(X,X,X)\}\\
E &= \{(C,C,X),(C,X,C),(C,X,X),(X,C,C),(X,C,X),(X,X,C),(X,X,X)\}.
\end{aligned}$$

I podem, per exemple, fer les operacions següents:

$$\begin{aligned}
A\cap B &= \{(C,C,X),(C,X,C)\},\\
A\cup B &= \{(C,C,C),(C,C,X),(C,X,C),(C,X,X),(X,C,C)\},\\
E^c &= \{(C,C,C)\},\\
A\cup E &= \Omega;\ \text{per tant, aquests dos conjunts són exhaustius,}\\
A\cap D &= \emptyset;\ \text{per tant, aquests dos conjunts són disjunts.}
\end{aligned}$$

2.2.2. Espai de probabilitat finit

Un cop tenim clares les definicions d'espai mostral i d'esdeveniment, ens falta parlar de la probabilitat que assignarem a cada resultat o esdeveniment.

La definició rigorosa d'espai probabilitat és molt tècnica i abstracta pel que necessitarem en aquest curs; és per això que introduirem només els espais de probabilitat finits, és a dir, quan l'espai mostral és finit.

Si l'espai mostral és $\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$, una **probabilitat** P serà una mesura numèrica de la possibilitat que passi un cert esdeveniment. Aquesta probabilitat P queda definida quan assignem a cada un dels resultats ω_i una probabilitat $P(\omega_i)$, $1 \leq i \leq k$. Aquesta probabilitat ha de complir unes certes propietats:

1) $\sum_{i=1}^{k} P(\omega_i) = 1$,

2) per a qualsevol i, $i = 1, \ldots, k$, $0 \leq P(\omega_i) \leq 1$.

Cas particular

Quan tots els elements tenen la mateixa probabilitat, tenim el cas equiprobable.

$$P(\omega_1) = \cdots = P(\omega_k) = \frac{1}{k}.$$

En aquest cas, calcular les probabilitats dels esdeveniments és fàcil. Si A és un esdeveniment

$$P(A) = \frac{card(A)}{card(\Omega)}$$

on $card(\cdot)$ denota el **cardinal** del conjunt, o sigui, el nombre d'elements que té el conjunt. Per calcular el cardinal d'un esdeveniment utilitzarem sovint la combinatòria.

Exemple 17. *Si seguim amb l'exemple 4, recordem que tenim l'espai mostral*

$$\Omega = \{(C,C,C),(C,C,X),(C,X,C),(C,X,X),(X,C,C),(X,C,X),(X,X,C),(X,X,X)\},$$

i observem que tots els elements tenen la mateixa probabilitat de donar-se; per tant, és un espai de probabilitat equiprobable. Veiem també que $card(\Omega) = 8$, i si considerem com abans l'esdeveniment $A = \{$com a mínim han sortit 2 cares$\}$, aleshores

$$A = \{(C,C,C),(C,C,X),(C,X,C),((X,C,C)\},$$

i per tant $card(A) = 4$ i podem calcular la probabilitat d'A

$$P(A) = \frac{4}{8} = \frac{1}{2}.$$

2.2.3. Propietats de la probabilitat

Algunes de les propietats bàsiques de la probabilitat són:

1) $P(\emptyset) = 0$.

2) $P(\Omega) = 1$.

3) Si A és un esdeveniment, aleshores $P(A^c) = 1 - P(A)$.

4) Si $\{B_k, 1 \leq k \leq n\}$ és una família d'esdeveniments disjunts 2 a 2, és a dir, tals que per a tot $i \neq j$ es compleix que $B_i \cap B_j = \emptyset$, aleshores

$$P(B_1 \cup \cdots \cup B_n) = P(B_1) + \cdots + P(B_n).$$

Això ens diu en particular que si A i B són dos esdeveniments disjunts, aleshores

$$P(A \cup B) = P(A) + P(B).$$

5) Si A i B són dos esdeveniments qualssevol, aleshores

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

I en particular tenim la desigualtat

$$P(A \cup B) \leq P(A) + P(B).$$

Exemple 18. *Dues preguntes per a resoldre utilitzant les propietats anteriors.*

1) *Si* $P(A) - P(A^c) = \frac{2}{3}$, *quant val* $P(A)$*?*

Com que sabem que $P(A^c) = 1 - P(A)$, tenim que $P(A) - (1 - P(A)) = \frac{2}{3}$; per tant,

$$P(A) = \frac{1}{2}\left(1 + \frac{2}{3}\right) = \frac{5}{6}.$$

2) *Si* $P(A^c) = \frac{1}{3}$, $P(B) = \frac{1}{6}$ *i sabem que* A *i* B *són disjunts, quant val* $P(A \cup B)$*?*

Com que A i B són disjunts, sabem que

$$P(A \cup B) = P(A) + P(B) = \left(1 - \frac{1}{3}\right) + \frac{1}{6} = \frac{5}{6}.$$

2.3. Probabilitat condicionada

Quan volem calcular una probabilitat, el fet de conèixer una informació prèvia sobre l'esdeveniment ens modifica la situació? Per exemple, si tirem dos daus, la probabilitat que els dos resultats obtinguts siguin parells serà la mateixa que si la volem calcular sabent que el resultat del segon dau és un dos? Evidentment, en aquest cas, el fet de saber informació fa que la probabilitat augmenti. Per poder calcular les probabilitats en aquests casos cal definir el concepte de *probabilitat condicionada.*

Definició 13. *Siguin* A *i* B *dos esdeveniments tals que* P(B)>*0, aleshores la* **probabilitat de A condicionada per/coneguda B** *és*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Fixeu-vos que la fórmula anterior també es pot escriure de la manera següent

$$P(A \cap B) = P(A|B)P(B).$$

A més, el paper dels esdeveniments A i B es pot intercanviar. Si $P(A) > 0$, es compleix que

$$P(A \cap B) = P(B|A)P(A),$$

i és certa la relació

$$P(A|B)P(B) = P(B|A)P(A).$$

Vegem ara un exemple en què utilitzarem la probabilitat condicionada.

Exemple 19. *El 80% dels clients d'una companyia telefònica tenen contractat un servei de telefonia fixa* (T), *el 75% tenen contractat un servei d'ADSL* (A) *i el 65% tenen contractats els dos serveis. Ens preguntem quina és la probabilitat que un client que tingui contractat un servei de telefonia fixa tingui contractat un servei d'ADSL.*

$$P(A|T) = \frac{P(A \cap T)}{P(T)} = \frac{0.65}{0.8} = 0.8125.$$

Observació 2. Hem de tenir en compte que la probabilitat $P(\cdot|B)$ té totes les propietats d'una probabilitat. Així, doncs, és cert que

$$P(A^c|B) = 1 - P(A|B),$$

o en el cas que A i C siguin esdeveniments disjunts,

$$P(A \cup C|B) = P(A|B) + P(C|B).$$

2.4. Esdeveniments independents

A vegades, però, el fet de tenir més informació no ens fa canviar les probabilitats. Pensem, per exemple, en l'experiment de la tirada de dos daus. Volem calcular la probabilitat que el resultat del primer dau sigui parell. Si sabem que el resultat del segon dau és un 2, aquesta nova informació no ens fa variar la probabilitat que havíem calculat. En aquest cas, direm que els esdeveniments: «treure un resultat parell en el primer dau» és independent de l'esdeveniment «treure un 2 en el segon dau».

Definició 14. *Donats dos esdeveniments* A *i* B, *direm que són* **esdeveniments independents** *si:*

$$P(A \cap B) = P(A)P(B).$$

Aquesta seria la definició rigorosa d'esdeveniments independents, però hi ha altres igualtats equivalents a aquesta.

Proposició 1. *Si* A *i* B *són dos esdeveniments tals que* P(A)>*0 i* P(B)>*0, aleshores les afirmacions següents són equivalents:*

1) *A i B són independents,*
2) $P(A|B) = P(A)$,
3) $P(B|A) = P(B)$.

Observació 3. Fixem-nos que les igualtats 2 i 3 ens estan dient que quan A i B són independents, condicionar no ens dóna cap nova informació.

Exemple 20. *Suposem que tenim una urna amb 7 boles blanques i 3 boles vermelles. En traiem dues boles amb reposició, i considerem els esdeveniments següents:*

$$A = \{\textit{la primera bola és vermella}\}$$
$$B = \{\textit{la segona bola és vermella}\}.$$

I ens preguntem si els esdeveniments A *i* B *són independents.*

Per decidir-ho, hem de calcular les probabilitats: $P(A \cap B)$, $P(A)$ i $P(B)$.

Observem que $P(A) = \frac{3}{10}$, $P(B) = \frac{3}{10}$ i finalment $P(A \cap B) = \frac{9}{100}$. Per tant, com que $P(A) \cdot P(B) = P(A \cap B)$, podem dir que els esdeveniments A i B són independents.

Propietats dels esdeveniments independents:

1) Els esdeveniments Ω i $\emptyset$ són independents de qualsevol altre esdeveniment.

2) Les afirmacions següents són equivalents:

a) A i B són independents,

b) A^c i B són independents,

c) A i B^c són independents,

d) A^c i B^c són independents.

Hem de vigilar amb el concepte d'independència, perquè a vegades no obtenim els resultats que esperàvem.

Exemple 21. *Tenim un dau una mica particular: a les 6 cares hi trobem respectivament: un 1 blanc, un 2 blanc, un 3 blanc, un 2 verd, un 3 verd i un 4 verd. Tirem el dau una vegada i considerem els esdeveniments*

$$A = \{\text{surt un número verd}\}$$
$$B = \{\text{surt un número parell}\}$$
$$C = \{\text{surt un 2}\}.$$

Aleshores, veiem que els esdeveniments A i B no són independents, però en canvi els esdeveniments A i C sí que ho són.

$$P(A \cap B) = \frac{2}{6} \text{ mentre que } P(A) = \frac{3}{6} \text{ i } P(B) = \frac{3}{6} \text{ per tant}$$
$$P(A \cap B) \neq P(A)P(B), \text{ és a dir, no són independents.}$$

En canvi,

$$P(A \cap C) = \frac{1}{6} \text{ mentre que } P(A) = \frac{3}{6} \text{ i } P(C) = \frac{2}{6} \text{ per tant}$$
$$P(A \cap C) = P(A)P(C), \text{ és a dir, són independents.}$$

2.5. Fórmula de les probabilitats totals

Per introduir la fórmula de les probabilitats totals necessitem definir un nou concepte, el de *partició.*

Una **partició** de l'espai mostral Ω és una família d'esdeveniments $\{A_1, A_2, \ldots, A_m\}$ que compleixen dues condicions:

1) són disjunts dos a dos,
2) són exhastius $\left(\Omega = \cup_{i=1}^{m} A_i\right)$.

Un exemple simple de partició és: $A_1 = A$ i $A_2 = A^c$ per a qualsevol A.

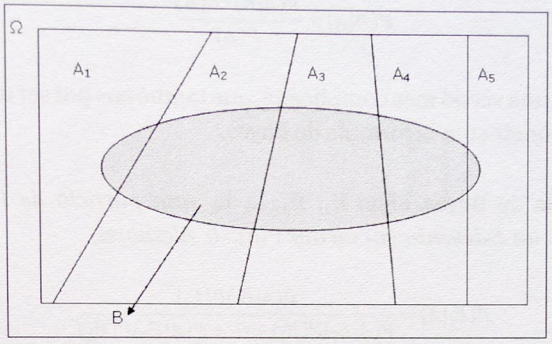


Figura 2.2. Exemple de partició.

Teorema 1. Fórmula de les probabilitats totals. *Sigui* $\{A_1, A_2, \ldots, A_m\}$ *una partició de l'espai mostral* Ω, *tal que* $P(A_i) > 0$ *per a tot* i. *Aleshores, donat qualsevol altre esdeveniment* B, *tenim la igualtat següent:*

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_m) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_m)P(A_m). \end{aligned}$$

Exemple 22. *El material d'oficina d'una empresa es compra en dues pàgines web diferents. A la primera pàgina web es compra el 40% del material, mentre que la resta es compra a la segona pàgina. El percentatge de devolucions pel material comprat a la primera pàgina és del 8%, mentre que el percentatge de devolucions per a la segona és del 10%. Agafeu un material a l'atzar i calculeu la probabilitat que s'hagi de retornar.*

Definim els esdeveniments següents:

$$\begin{aligned} D &= \{\text{s'ha de retornar el material}\}, \\ W1 &= \{\text{el material s'ha comprat a la pàgina } 1\}, \\ W2 &= \{\text{el material s'ha comprat a la pàgina } 2\}. \end{aligned}$$

Aleshores, volem calcular $P(D)$:

$$P(D) = P(D|W1)P(W1) + P(D|W2)P(W2) = 0.08 \cdot 0.4 + 0.1 \cdot 0.6 = 0.092.$$

2.6. Fórmula de Bayes

En alguns casos, necessitem calcular una probabilitat condicionada quan justament coneixem la probabilitat amb la condició inversa. O sigui, suposem que donats dos esdeveniments A i B volem calcular $P(B|A)$ i coneixem:

- $P(A)$ i $P(B)$,
- i a més també coneixem $P(A|B)$.

Aleshores, podem invertir les condicions.

Teorema 2. *Siguin* A *i* B *dos esdeveniments tals que* $P(A)>0$ *i* $P(B)>0$. *Aleshores,*

$$P(B|A)=\frac{P(A|B)\cdot P(B)}{P(A)}.$$

Observació 4. Hi ha una versió més complicada, que també ens pot ser útil en determinades situacions i que es coneix com la fórmula de Bayes.

Teorema 3. Fórmula de Bayes. *Sigui* $E_1, E_2, \ldots, E_k$ *una partició de* Ω *tal que* $P(E_i)>0$ $\forall i=1,\ldots,k$, *i sigui* A *un esdeveniment tal que* $P(A)>0$. *Aleshores,*

$$P(E_i|A)=\frac{P(A|E_i)P(E_i)}{P(A|E_1)P(E_1)+\cdots+P(A|E_k)P(E_k)}.$$

Exemple 23. *Una botiga d'electrodomèstics té classificats els aparells que ven segons l'eficiència energètica. Saben que es venen un 25% d'aparells bons, un 30% d'aparells regulars i un 45% d'aparells dolents, energèticament parlant. Dels aparells que venen, alguns necessiten una reparació abans de dos anys, cosa que passa en un 10% dels aparells bons, un 20% dels aparells regulars i un 40% dels aparells dolents. Ens preguntem quina és la probabilitat que un aparell que necessita una reparació abans de dos anys sigui un aparell bo.*

En primer lloc, definim els esdeveniments següents:

$$\begin{aligned}
B &= \{\text{l'electrodòmestic és bo}\},\\
R &= \{\text{l'electrodòmestic és regular}\},\\
D &= \{\text{l'electrodòmestic és dolent}\},\\
G &= \{\text{l'electrodòmestic necessita una reparació abans de dos anys}\}.
\end{aligned}$$

Un cop definits aquests esdeveniments, les probabilitats que ens dóna l'enunciat són les següents:

$$P(B)=0.25,\ P(R)=0.3\ P(D)=0.45,$$
$$P(G|B)=0.1,\ P(G|R)=0.2\ P(G|D)=0.4.$$

I volem calcular la probabilitat $P(B|G)$. Aleshores utilitzarem la fórmula de Bayes,

$$P(B|G)=\frac{P(G|B)P(B)}{P(G)}=\frac{0.1\cdot 0.25}{P(G)}.$$

La probabilitat $P(G)$, la podem calcular utilitzant la fórmula de probabilitats totals. Així, doncs,

$$P(G)=P(G|B)\cdot P(B)+P(G|R)\cdot P(R)+P(G|D)\cdot P(D)=0.1\cdot 0.25+0.2\cdot 0.3+0.4\cdot 0.45=0.265.$$

Aleshores,

$$P(B|G) = \frac{0.1 \cdot 0.25}{0.265} = 0.0943.$$

Fixeu-vos que, de fet, hem utilitzat la fórmula de Bayes.

3. VARIABLES ALEATÒRIES

En el capítol anterior hem parlat de l'espai mostral. L'espai mostral pot ser molt diferent en funció de l'experiment que estiguem realitzant, i de fet pot ser no numèric. A nosaltres ens interessarà, però, treballar amb valors numèrics, i és per això que introduirem el concepte de *variable aleatòria*.

3.1. Variable aleatòria: primers conceptes

Una **variable aleatòria** és una funció que fa correspondre a cada element de l'espai mostral un valor numèric.

La definició rigorosa de variable aleatòria és més complicada i tècnica que els objectius d'aquest curs, però el que a nosaltres ens interessarà és tenir-ne clar el concepte.

Exemple 24. *Alguns exemples de variables aleatòries podrien ser:*

- *El nombre de clients que arriba a una cua durant una hora.*
- *El nombre de vegades que hem de tirar un dau fins a obtenir el primer 6.*
- *El temps que dura una bombeta fins que falla.*
- *Els ingressos d'una família.*

Comentem ara més detalladament un exemple sobre variables aleatòries.

Exemple 25. *Si tirem un dau 5 vegades, podem definir una variable aleatòria* X *com el nombre d'uns que hem obtingut. Un altre exemple de variable aleatòria* Y *podria ser el valor de la suma del primer i l'últim resultat. I d'aquesta manera, podríem anar definint moltes variables aleatòries.*

Observem que la variable aleatòria X pot prendre qualsevol valor entre 0 i 5; en canvi, la variable aleatòria Y pot prendre els valors entre 2 i 12.

Per a la variable aleatòria X ja podem calcular probabilitats. Per exemple, ens podem preguntar quant val la probabilitat que de les 5 tirades que fem del dau tots els resultats siguin 1; en aquest cas ens estem demanant la $P(X = 5)$. Com que estem suposant que les tirades són independents,

$$P(X=5)=P(\{1,1,1,1,1\})=\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}=\left(\frac{1}{6}\right)^5.$$

Igualment ens podem preguntar la probabilitat que la suma del primer i l'últim resultat sigui 4, i per tant el que volem calcular és $P(Y = 4)$. Com que el resultat de les tirades 2, 3 i 4 no ens interessa, podem pensar que només tenim dos daus.

$$P(Y=4)=P(\{1,3\},\{2,2\},\{3,1\})=\frac{3}{36}.$$

Observem que tenim 3 casos favorables i d'un total de $6 \cdot 6 = 36$ de possibles, totes les combinacions dels resultats dels dos daus.

Remarquem que les operacions habituals entre variables aleatòries donen lloc a una altra variable aleatòria.

Proposició 2. *Si* X *i* Y *són dues variables aleatòries i* a *i* b *dos valors reals, aleshores*

$$aX+bY,\ XY,\ \text{i}\ \frac{X}{Y}\ (\text{si } Y \neq 0)$$

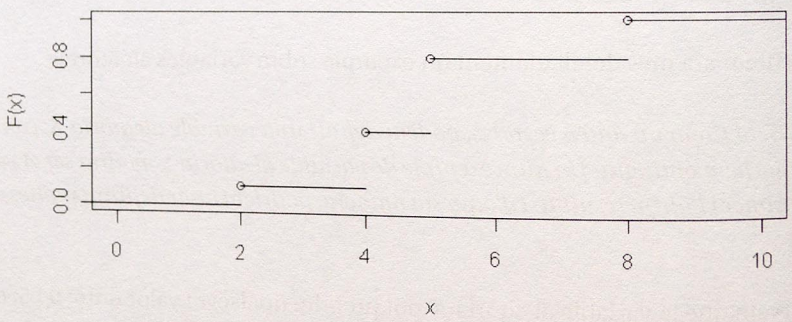
són també variables aleatòries.

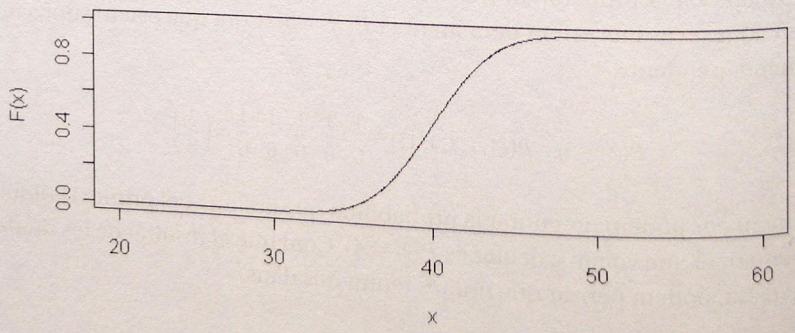
Una de les eines bàsiques quan estudiem variables aleatòries és la *funció de distribució.*

Definició 15. *Sigui* X *una variable aleatòria; aleshores, una funció* $F:\mathbb{R}\to[0,1]$ *serà la seva* **funció de distribució** *si* F *li fa correspondre a cada valor* x *la probabilitat que la variable aleatòria prengui un valor inferior o igual a* x, *és a dir,*

$$F(x):=P(X\leq x).$$

Fixeu-vos que a partir de la definició les funcions de distribució poden ser molt diferents. Per exemple,





En tot cas, la funció de distribució compleix sempre les propietats següents:

- $\lim_{x\to+\infty} F(x)=1$ i $\lim_{x\to-\infty} F(x)=0$,
- $F(x)$ és una funció creixent: si $a \leq b$, aleshores $F(a) \leq F(b)$,
- $F(x)$ no sempre serà una funció contínua, però en canvi sempre serà contínua per la dreta, és a dir,

$$\lim_{y\downarrow x} F(y)=F(x).$$

Observem que podem fer servir la definició de la funció de distribució per calcular probabilitats de la variable X:

- $P(a < X \leq b)=F(b)-F(a)$,
- $P(a \leq X \leq b)=F(b)-F(a^-)$,
- $P(a \leq X < b)=F(b^-)-F(a^-)$,
- $P(a < X < b)=F(b^-)-F(a)$,
- $P(X < a)=F(a^-)$,
- $P(X = a)=F(a)-F(a^-)$, on $F(a^-)=\lim_{x\uparrow a} F(x)$.

Com podem veure a partir de les gràfiques anteriors, les funcions de distribució poden ser bàsicament esglaonades o contínues. Per tant, segons com sigui la funció de distribució, direm que les variables aleatòries són variables aleatòries discretes o variables aleatòries absolutament contínues.

Observació 5. Quan coneguem el comportament d'una variable aleatòria, parlarem indistintament de la **llei** o de la **distribució** de la variable aleatòria.

Comencem estudiant el cas de les variables aleatòries discretes.

3.2. Variables aleatòries discretes

Definició 16. *Una* **variable aleatòria** *és* **discreta** *si pot prendre només un nombre finit o numerable de valors de manera que cadascun d'aquests valors té una probabilitat més gran que zero de ser observat.*

A partir de la definició de variable aleatòria discreta, observem que una variable d'aquest tipus queda determinada en el moment en què coneixem la probabilitat que la variable prengui cadascun dels seus possibles valors. Precisament, la funció que ens dóna la probabilitat de cada valor s'anomena *funció de massa de probabilitat.*

Definició 17. *La* **funció de massa de probabilitat** *d'una variable aleatòria discreta és la funció que ens dóna per cada valor real* x *la seva probabilitat corresponent, o sigui* $P(X = x)$.

Observació 6. Tot sovint obviarem l'expressió «de massa» i parlarem només de funció de probabilitat.

PROPIETAT

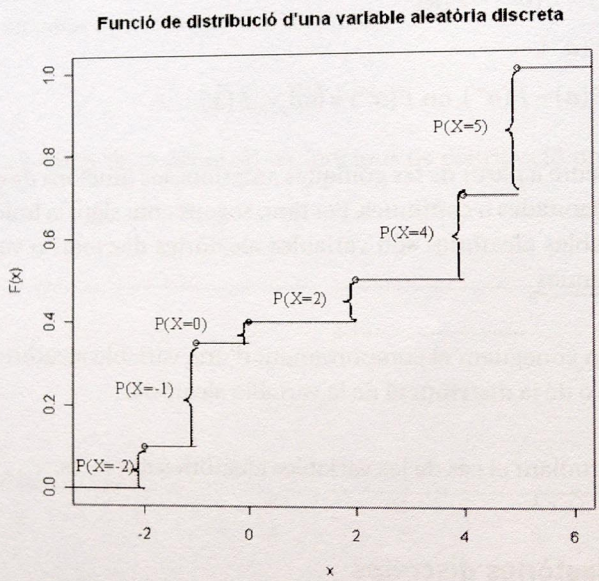
La suma de les probabilitats $P(X = x)$ per a tots els valors x que pugui prendre la nostra variable serà 1,

$$\sum_x P(X = x) = 1.$$

Fixem-nos també que a partir de la funció de massa de probabilitat d'una variable aleatòria discreta podem calcular-ne la funció de distribució, utilitzant que

$$F(x) = P(X \leq x) = \sum_{y \leq x} P(X = y).$$

Observem que per una variable aleatòria discreta la funció de distribució sempre serà una funció esglaonada.



A partir de la figura anterior podem veure que només tenim salts en els valors que pren la variable X, i a més l'alçada d'aquests salts és justament la probabilitat del punt.

En el cas de la funció de distribució de la gràfica, veiem que es tracta d'una variable X amb: $P(X = -2) = 0.1$, $P(X = -1) = 0.25$, $P(X = 0) = 0.05$, $P(X = 2) = 0.1$, $P(X = 4) = 0.2$, $P(X = 5) = 0.25$.

Exemple 26. *L'exemple més senzill de variable aleatòria és la variable aleatòria constant* $X = c$, *on* c *pot ser qualsevol valor.*

Aquesta variable aleatòria té per funció de probabilitat:

$$P(X = c) = 1.$$

I la seva funció de distribució és:

$$F(x)=\begin{cases} 0 & \text{si} \quad x<c \\ 1 & \text{si} \quad c\leq x. \end{cases}$$

Observació 7. Tenim la mateixa informació tant si coneixem la funció de distribució com la funció de probabilitat d'una variable aleatòria discreta. De fet, a partir d'una de les dues podem trobar l'altra.

Vegem en primer lloc, com calcular la funció de distribució a partir de la funció de probabilitat.

Exemple 27. *Tirem dos daus iguals de 5 cares, primer un i després l'altre, i ens fixem si les cares superiors són parells o senars.* Per tant, el nostre espai mostral serà:

$$\Omega=\{(\text{parell, parell}), (\text{parell, senar}), (\text{senar, parell}), (\text{senar, senar})\},$$

i com que una tirada és independent de l'altra, la probabilitat de cada element és:

$$\begin{aligned} P(\text{parell, parell}) &= \frac{2}{5}\cdot\frac{2}{5}=\frac{4}{25}, \\ P(\text{parell, senar}) &= P(\text{senar, parell})=\frac{3}{5}\cdot\frac{2}{5}=\frac{6}{25}, \\ P(\text{senar, senar}) &= \frac{3}{5}\cdot\frac{3}{5}=\frac{9}{25}. \end{aligned}$$

La variable que considerarem serà X = nombre de resultats parells obtinguts. Observem, per tant, que la variable aleatòria X pot prendre els valors 0, 1 o 2. I la seva funció de probabilitat serà:

$$\begin{aligned} P(X=0) &= P(\text{parell, parell})=\frac{4}{25}, \\ P(X=1) &= P(\text{parell, senar})+P(\text{senar, parell})=\frac{12}{25}, \\ P(X=2) &= P(\text{senar, senar})=\frac{9}{25}. \end{aligned}$$

Ara, si volem calcular la funció de distribució a partir de la funció de probabilitat, recordem que $F(x)=P(X\leq x)$, i aleshores observem que

1) Si $x<0$, aleshores $F(x)=P(X\leq x)=0$,

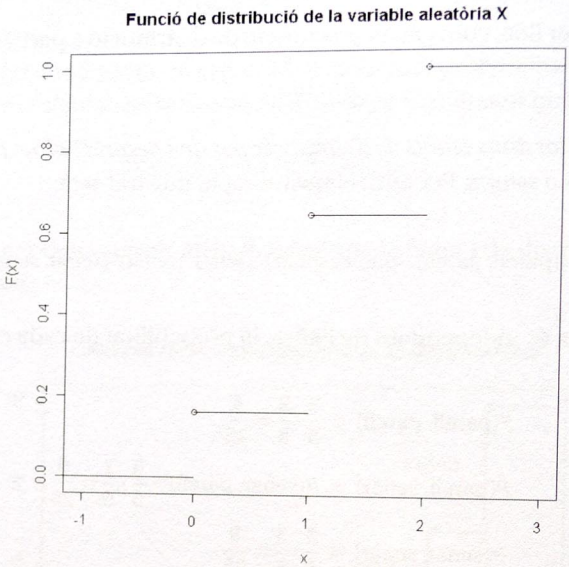
2) Si $0\leq x<1$, aleshores $F(x)=P(X\leq x)=P(X=0)=\frac{4}{25}$,

3) Si $1\leq x<2$, aleshores $F(x)=P(X\leq x)=P(X=0)+P(X=1)=\frac{4}{25}+\frac{12}{25}=\frac{16}{25}$,

4) Si $x\geq 2$, aleshores $F(x)=P(X\leq x)=P(X=0)+P(X=1)+P(X=2)=\frac{16}{25}+\frac{9}{25}=1$.

Per tant, la funció de distribució de la variable X és:

$$F(x)=\begin{cases} 0 & \text{si} \quad x<0 \\ \frac{4}{25} & \text{si} \quad 0\leq x<1 \\ \frac{16}{25} & \text{si} \quad 1\leq x<2 \\ 1 & \text{si} \quad 2\leq x \end{cases}$$



Ara utilitzarem el següent exemple per veure com funciona el procés invers, és a dir, calcular la funció de probabilitat a partir de la funció de distribució.

Exemple 28. *Sigui* X *una variable aleatòria amb funció de distribució*

$$F(x)=\begin{cases} 0 & \text{si} \quad x<-2 \\ \frac{1}{6} & \text{si} \quad -2\leq x<1 \\ \frac{1}{2} & \text{si} \quad 1\leq x<2 \\ \frac{3}{4} & \text{si} \quad 2\leq x<5 \\ 1 & \text{si} \quad 5\leq x. \end{cases}$$

Si volem calcular la funció de probabilitat de la variable X, el primer que hem de tenir clar és quins valors pot prendre la nostra variable. Els valors que pot prendre seran sempre els punts de salt, en aquest cas -2, 1, 2 i 5. Si recordem que una de les propietats que hem donat de la funció de distribució era que $P(X=a)=F(a)-F(a^-)$, si la utilitzem en el nostre cas ens quedarà que:

$$P(X=-2)=F(-2)=\frac{1}{6},$$

$$P(X=1)=F(1)-F(1^-)=\frac{1}{2}-\frac{1}{6}=\frac{1}{3},$$

$$P(X=2)=F(2)-F(2^-)=\frac{3}{4}-\frac{1}{2}=\frac{1}{4},$$

$$P(X=5)=F(5)-F(5^-)=1-\frac{3}{4}=\frac{1}{4}.$$

Observació 8. Si ens fixem en aquest últim exemple, observem que no coneixem l'espai mostral amb què estem treballant, sinó que només coneixerem la llei (funció de distribució o de probabilitat) de la variable aleatòria amb què tractem. Aquesta serà la situació amb què ens trobarem habitualment.

3.2.1. Independència de variables aleatòries discretes

Recordem que en el capítol anterior hem parlat d'independència d'esdeveniments. Vegem ara com entenem el mateix concepte per a variables aleatòries.

Definició 18. *Siguin* X *i* Y *dues variables aleatòries discretes; direm que* X *i* Y *són* **independents** *si es compleix que*

$$P(X=x, Y=y)=P(X=x)\cdot P(Y=y)$$

per tots els valors x *i* y *que puguin prendre les variables* X *i* Y *respectivament.*

Fixeu-vos que en aquesta definició hi ha la mateixa idea d'independència que quan treballàvem amb esdeveniments.

Exemple 29. *Tirem dues vegades un dau i considerem les variables aleatòries següents:* X = *nombre de resultats parells que hem obtingut i* Y = *nombre de dosos que hem obtingut.*

Observem, per exemple, que

$$P(X=0)=P(\text{senar, senar})=\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4},$$

$$P(Y=2)=P(2,2)=\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{36},$$

$$P(X=0, Y=2)=0.$$

Com que $P(X=0, Y=2)\neq P(X=0)\cdot P(Y=2)$, podem afirmar que X i Y no són variables aleatòries independents.

Observació 9. Observem que n'hi ha prou que un producte no sigui igual que la intersecció per negar la independència; en canvi, hauríem de comprovar tots els productes per demostrar la independència.

Exemple 30. *Tenim una caixa amb 3 boles numerades de l'1 al 3. Traiem una primera bola, i sigui* X *el número d'aquesta bola. Tornem la bola a la caixa i traiem una segona bola i diem* Y *al número d'aquesta segona bola. Són independents les variables* X *i* Y*?*

Per veure-ho, hem de calcular les probabilitats $P(X=i, Y=j)$, $P(X=i)$ i $P(Y=j)$ per a $i, j=1,2,3$. Observem que

$$P(X=i, Y=j)=\frac{1}{9}, \qquad \text{si } i, j = 1, 2, 3.$$

I també que

$$P(X=i)=P(Y=j)=\frac{1}{3}, \qquad \text{si } i, j = 1, 2, 3.$$

Per tant, $P(X=i, Y=j)=P(X=i)\cdot P(Y=j)$ per a tots els valors possibles de i i j, i aleshores podem dir que les variables X i Y són independents.

3.2.2. Esperança i variància de variables aleatòries discretes

En el cas de variables aleatòries, tenim un concepte semblant al de mitjana mostral que s'utilitza en l'estadística descriptiva que anomenarem *esperança*.

Definició 19. *L'***esperança** *d'una variable aleatòria discreta* X, *que pren els valors* $x_1, \ldots, x_n$, *és la mitjana ponderada dels valors que pren i es calcula com*

$$E(X)=\sum_{i=1}^{n} x_i \cdot P(X=x_i).$$

Observació 10. Hem de tenir en compte que n pot ser ∞, una variable aleatòria discreta pot prendre un nombre numerable de valors.

Aquesta definició es pot estendre en el cas que vulguem calcular l'esperança d'una transformació g de la variable aleatòria X. En aquest cas, calcularem l'esperança de $g(X)$ com

$$E(g(X))=\sum_{i=1}^{n} g(x_i)\cdot P(X=x_i).$$

Observació 11. Si la variable pren un nombre no finit (numerable) de valors, tots els resultats i les definicions d'aquest apartat són certs quan els sumatoris que apareixen tenen sentit. Per exemple, hi ha variables en què, quan intentem calcular l'esperança, ens surt un sumatori que no està ben definit. En aquests casos, direm que l'esperança no existeix.

Observeu que, donada una variable aleatòria X, la seva esperança és, sempre que existeixi, un valor real.

Donarem algunes propietats bàsiques de l'esperança.

Propietats de l'esperança

1) Siguin X i Y dues variables aleatòries i a, b dos nombres reals. Aleshores:

$$E(aX+bY)=aE(X)+bE(Y).$$

2) Sigui k una constant; aleshores, $E(k)=k$.

3) $E(-X)=-E(X)$.

En estadística descriptiva, a més del concepte de mitjana mostral, també hi ha el de variància mostral i el de desviació típica, dues mesures de dispersió. Els mateixos tipus de càlculs els poden fer per a variables aleatòries.

Definició 20. *La* **variància** *d'una variable aleatòria* X *es defineix com*

$$Var(X) = E\left((X - E(X))^2\right) = E(X^2) - (E(X))^2.$$

A partir de la definició d'esperança, tenim

$$Var(X) = \sum_{i=1}^{n} (x_i - E(X))^2 \cdot P(X = x_i),$$

o bé la forma alternativa,

$$Var(X) = \sum_{i=1}^{n} x_i^2 \cdot P(X = x_i) - \left(\sum_{i=1}^{n} x_i \cdot P(X = x_i)\right)^2.$$

Observació 12. Hem de tenir en compte, com en el cas de l'esperança, que n pot ser ∞.

I també podem definir la desviació típica d'una variable aleatòria X.

Definició 21. *La* **desviació típica** *d'una variable aleatòria* X *es defineix com*

$$s(X) = \sqrt{Var(X)}.$$

Propietats de la variància

1) Sigui X una variable aleatòria i a, b dos nombres reals. Aleshores:

$$Var(aX + b) = a^2\, Var(X).$$

2) Sigui k una constant; aleshores, $Var(k) = 0$.
3) $Var(-X) = Var(X)$.
4) $Var(X) \geq 0$ i $Var(X) = 0$ si i només si la variable X és constant.
5) Siguin X i Y dues variables aleatòries **independents**. Aleshores:

$$Var(X + Y) = Var(X) + Var(Y).$$

Exemple 31. *Tenim dos daus iguals que tenen dues cares vermelles i les altres liles. Tirem els dos daus i considerem la variable* X = *nombre de cares liles obtingudes.*

Aleshores la funció de massa de probabilitat de X és

$$\begin{aligned} P(X=0) &= \frac{2}{6} \cdot \frac{2}{6} = \frac{4}{36}, \\ P(X=1) &= \frac{2}{6} \cdot \frac{4}{6} + \frac{4}{6} \cdot \frac{2}{6} = \frac{16}{36}, \\ P(X=2) &= \frac{4}{6} \cdot \frac{4}{6} = \frac{16}{36}, \end{aligned}$$

i podem calcular:

- La seva esperança

$$E(X) = \sum_{k=0}^{2} k \cdot P(X=k) = 0 \cdot P(X=0) + 1 \cdot P(X=1) + 2 \cdot P(X=2)$$
$$= 0 \cdot \frac{4}{36} + 1 \cdot \frac{16}{36} + 2 \cdot \frac{16}{36} = \frac{48}{36} = \frac{4}{3}.$$

- $E(X^2)$, que ens servirà per a calcular la variància

$$E(X^2) = \sum_{k=0}^{2} k^2 \cdot P(X=k) = 0^2 \cdot \frac{4}{36} + 1^2 \cdot \frac{16}{36} + 2^2 \cdot \frac{16}{36} = \frac{80}{36} = \frac{20}{9}.$$

- L'esperança d'una transformació de la variable aleatòria X

$$E(-5X+7) = -5E(X) + 7 = -5 \cdot \frac{48}{36} + 7 = \frac{1}{3}.$$

Podem calcular també la variància de dues maneres:

$$Var(X) = \sum_{k=0}^{2} \left(k - \frac{4}{3}\right)^2 \cdot P(X=k)$$
$$= \left(-\frac{4}{3}\right)^2 \cdot \frac{4}{36} + \left(-\frac{1}{3}\right)^2 \cdot \frac{16}{36} + \left(\frac{2}{3}\right)^2 \cdot \frac{16}{36} = \frac{4}{9},$$

o per l'expressió alternativa

$$Var(X) = E(X^2) - (E(X))^2 = \frac{20}{9} - \left(\frac{4}{3}\right)^2 = \frac{4}{9}.$$

3.2.3. Exemples de variables aleatòries discretes

En aquest apartat presentarem les distribucions discretes més habituals.

Distribució de Bernoulli. Una variable aleatòria X que pot prendre dos possibles resultats $\{0, 1\}$ segueix una **distribució de Bernoulli** de paràmetre p, amb $p \in (0,1)$, si

$$P(X=0) = 1-p,$$
$$P(X=1) = p.$$

Escriurem que $X \sim Ber(p)$.

L'esperança i la variància d'una distribució Bernoulli són

$$E(X) = p, \quad Var(X) = p \cdot (1-p).$$

Observació 13. Una distribució de Bernoulli serveix per a modelar qualsevol experiment amb probabilitat p d'èxit i probabilitat $1-p$ de fracàs. Aleshores,

$$X = \begin{cases} 1 & \text{si és un èxit} \\ 0 & \text{si és un fracàs.} \end{cases}$$

Exemple 32. *Tirem un dau no trucat i observem si surt un 2.* Aleshores definim la variable aleatòria

$$X = \begin{cases} 1 & \text{si ha sortit un 2} \\ 0 & \text{en cas contrari.} \end{cases}$$

Aleshores, $X \sim Ber(\frac{1}{6})$, i per tant $E(X) = \frac{1}{6}$ i $Var(X) = \frac{1}{6}\left(1 - \frac{1}{6}\right) = \frac{5}{36}$.

Distribució binomial. Fem n (nombre natural) experiments Bernoulli independents amb la mateixa probabilitat d'èxit p ($p \in (0,1)$). Aleshores, una variable aleatòria Y que compta el nombre d'èxits en aquests n experiments segueix una **distribució binomial** de paràmetres n i p. La variable Y pren valors en el conjunt $\{0, 1, \ldots, n\}$ amb funció de massa de probabilitat:

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \; k = 0, 1, \ldots, n.$$

Escriurem que $Y \sim B(n, p)$.

L'esperança i la variància d'una distribució binomial són

$$E(Y) = n \cdot p, \quad Var(Y) = n \cdot p \cdot (1-p).$$

Exemple 33. *Resolem un test de 100 preguntes a l'atzar, en què cada pregunta té tres respostes diferents. Contestem totes les preguntes. Quina és la probabilitat que haguem respost correctament exactament 38 preguntes? I que n'haguem respost correctament més de 20? I menys de 57? I entre 30 i 34?*

En primer lloc, observem que si definim la variable aleatòria $X =$ «nombre de respostes correctes que hem obtingut», $X \sim B(100, \frac{1}{3})$, ja que en total tenim $n = 100$ preguntes per contestar i la probabilitat d'èxit (de donar la resposta correcta) és $\frac{1}{3}$. Aleshores les probabilitats que ens demanen són:

$$\begin{aligned}
P(X = 38) &= \binom{100}{38}\left(\frac{1}{3}\right)^{38}\left(1 - \frac{1}{3}\right)^{100-38} = \texttt{dbinom(38,100,1/3)} = 0.05073657, \\
P(X > 20) &= 1 - P(X \leq 20) = 1 - F(20) = \texttt{1-pbinom(20,100,1/3)} = 0.9976302, \\
P(X < 57) &= P(X \leq 56) = \texttt{pbinom(56,100,1/3)} = 0.999999 \\
P(30 < X < 34) &= P(X = 31) + P(X = 32) + P(X = 33) = \texttt{dbinom(31,100,1/3)} \\
&\quad + \texttt{dbinom(32,100,1/3)} + \texttt{dbinom(33,100,1/3)} = 0.2422495, \\
&= P(30 < X \leq 33) = F(33) - F(30) \\
&= \texttt{pbinom(33,100,1/3)} - \texttt{pbinom(30,100,1/3)} = 0.2422495
\end{aligned}$$

I $E(X) = 100 \cdot \frac{1}{3} = \frac{100}{3}$ i $Var(X) = 100 \cdot \frac{1}{3}\left(1 - \frac{1}{3}\right) = \frac{200}{9}$.

Proposició 3. *Si* $X_1, X_2, \ldots, X_n$ *són variables aleatòries independents idènticament distribuïdes amb llei* $\mathrm{Ber(p)}$, *aleshores* $Y = X_1 + X_2 + \cdots + X_n$ *té una distribució* $\mathrm{B(n, p)}$.

Distribució geomètrica. Realitzem ara experiments Bernoulli independents amb la mateixa probabilitat d'èxit p ($p \in (0,1)$). Una variable aleatòria Z que compta el nombre d'expe-

riments necessaris per tal d'observar el primer èxit segueix una **distribució geomètrica** de paràmetre p. La variable aleatòria Z pren valors en els nombres naturals (no inclou el 0) amb funció de massa de probabilitat:

$$P(Z=k)=(1-p)^{k-1}p,\ k\geq 1.$$

Escriurem que $Z \sim Geom(p)$.

L'esperança i la variància d'una distribució geomètrica són:

$$E(Y)=\frac{1}{p},\quad Var(Y)=\frac{1-p}{p^2}.$$

Observació 14. En el cas de la distribució geomètrica, quan utilitzem R, hem d'anar amb compte. Hi ha una manera alternativa de definir la distribució geomètrica: comptar el nombre de fracassos fins a l'èxit. Per tant, en aquesta segona definició la variable pot prendre els valors $n = 0, 1, \ldots$ i és la definició que fa servir R. Així, doncs, quan vulguem calcular probabilitats d'una distribució geomètrica amb R, posarem el valor anterior al que necessitem calcular.

Exemple 34. *La probabilitat que una fotocopiadora faci malament una fotocòpia és de 0.03. Quina és la probabilitat que la primera fotocòpia defectuosa ens aparegui quan fem la fotocòpia 20? I la probabilitat que la primera fotocòpia defectuosa ens aparegui abans de fer-ne 15?*

Anomenem Y la variable aleatòria que compta el nombre de fotocòpies correctes fetes fins que apareix la primera defectuosa. Aleshores volem calcular

$$P(Y=20) = (1-0.03)^{19}\cdot 0.03 = \texttt{dgeom(19,0.03)} = 0.01681838$$
$$P(Y<15) = P(Y\leq 14) = \texttt{pgeom(13,0.03)} = 0.3471637.$$

I $E(Y)=\frac{1}{0.03}=33.33$ i $Var(Y)=\frac{1-0.03}{0.03^2}=1077.778$.

Exemple 35. *Tenim un dau trucat, de manera que la probabilitat de treure un 6 és de $\frac{2}{3}$, mentre que la probabilitat de treure qualsevol dels altres valors és de $\frac{1}{15}$.*

1) Anomenem X la variable aleatòria

$$X=\begin{cases} 1 & \text{si ha sortit un 6} \\ 0 & \text{en cas contrari} \end{cases}$$

Aleshores, $X \sim Ber(\frac{2}{3})$.

2) Fem 10 tirades del dau i anomenem Y la variable aleatòria que compta el nombre de sisos que hem obtingut. Tindrem, per tant, que $Y \sim B(10, \frac{2}{3})$. Podem calcular probabilitats com:

a) Probabilitat de treure exactament 7 sisos.

$$P(Y=7)=\binom{10}{7}\left(\frac{2}{3}\right)^7\left(1-\frac{2}{3}\right)^3 = \texttt{dbinom(7,10,2/3)} = 0.260123.$$

b) Probabilitat de treure almenys un sis.

$$P(Y \geq 1) = 1 - P(Y=0) = 1 - \binom{10}{0}\left(\frac{2}{3}\right)^0\left(1-\frac{2}{3}\right)^{10}$$

$$= 1 - \left(\frac{1}{3}\right)^{10} = \texttt{1-dbinom(0,10,2/3)} = 0.999983.$$

c) Probabilitat de treure més de 8 sisos.

$$P(Y > 8) = 1 - P(Y \leq 8) = \texttt{1-pbinom(8,10,2/3)} = 0.104049.$$

3) Ara anem fent tirades del dau i anomenem Z la variable aleatòria que compta el nombre de tirades que ens cal fer fins a obtenir el primer sis. Així, la variable Z seguirà una distribució geomètrica de paràmetre $\frac{2}{3}$, $Z \sim Geom(\frac{2}{3})$. Aleshores podem calcular:

a) Probabilitat de necessitar exactament tres tirades.

$$P(Z=3) = \left(1-\frac{2}{3}\right)^2\left(\frac{2}{3}\right) = \texttt{dgeom(2,2/3)} = 0.074074.$$

b) Probabilitat de necessitar més d'una tirada.

$$P(Z > 1) = 1 - P(Z=1) = 1 - \left(1-\frac{2}{3}\right)^0\left(\frac{2}{3}\right)$$

$$= \texttt{1-dgeom(0,2/3)} = 0.33333.$$

c) Probabilitat de necessitar més de 5 tirades.

$$P(Z > 5) = \sum_{k=6}^{\infty} P(Z=k) = \sum_{k=6}^{\infty}\left(1-\frac{2}{3}\right)^{k-1}\left(\frac{2}{3}\right)$$

$$= 1 - P(Z \leq 5) = \texttt{1-pgeom(4,2/3)} = 0.00411.$$

Distribució de Poisson. La distribució de Poisson serveix en el recompte d'esdeveniments estranys, poc freqüents, etc., i que normalment estan relacionats amb el pas del temps. Una variable aleatòria T segueix una **distribució de Poisson** de paràmetre λ, $\lambda > 0$, si pren valors en els enters no negatius amb funció de massa de probabilitat:

$$P(T=k) = \exp(-\lambda)\frac{\lambda^k}{k!},\ k \geq 0.$$

Escriurem $X \sim Pois(\lambda)$.
L'esperança i la variància d'una distribució de Poisson són:

$$E(T) = \lambda,\quad Var(T) = \lambda.$$

Proposició 4. *Si* $X_1 \sim$ Pois (λ_1), $X_2 \sim$ Pois (λ_2),..., $X_n \sim$ Pois (λ_n) *on* $X_1, X_2, \ldots, X_n$ *són independents, aleshores*

$$X_1 + X_2 + \cdots + X_n \sim Pois(\lambda_1 + \lambda_2 + \cdots + \lambda_n).$$

Exemple 36. *El nombre de copes que es trenquen en un restaurant durant una hora segueix una distribució* Pois(*0.1*). *Volem calcular la probabilitat que es trenqui alguna copa durant la primera hora de servei i la probabilitat que es trenquin 6 copes durant 4 hores.*

Definim X = nombre de copes que es trenquen durant una hora. Aleshores, $X \sim Pois(0.1)$, i, per tant, ens demanen que calculem

$$P(X>0)=1-P(X=0)=1-e^{-0.1}=\texttt{1-dpois(0,0.1)}=0.09516258.$$

Pel que fa a la segona probabilitat que ens demanen, si X_i = nombre de copes que es trenquen durant l'hora per $i=1,2,3,4$, i $Y=X_1+X_2+X_3+X_4 \sim Pois(0.1+0.1+0.1+0.1)$. Hem de calcular:

$$P(Y=6)=e^{-0.4}\cdot\frac{0.4^6}{6!}=\texttt{dpois(6,0.4)}=3.813376\cdot 10^{-6}.$$

I, a més, $E(Y)=Var(Y)=0.4$.

3.3. Variables aleatòries absolutament contínues

Fins ara ens hem centrat en l'estudi de variables aleatòries que prenien un conjunt numerable de valors; ara, en canvi, estudiarem les variables aleatòries que prenen tots els valors d'un interval. D'aquest tipus de variables aleatòries només estudiarem aquelles que tenen una funció de distribució «bona», és a dir, que sigui contínua i derivable (considerarem que és derivable encara que no ho sigui en un nombre finit de punts).

Com que la funció de distribució F és derivable, ens apareix una nova funció: la derivada de la funció de distribució $f=F'$. Aleshores serà cert que

$$F(x)=P(X\le x)=\int_{-\infty}^{x} f(y)\,dy.$$

Així doncs quan parlem de variables absolutament contínues hem de tenir en compte les definicions següents:

Definició 22. *Una funció f és una* **funció de densitat** *si:*

1) $f(x)\ge 0$ *per qualsevol valor de* x,

2) f *és integrable i, a més,* $\int_{-\infty}^{\infty} f(x)\,dx=1$.

Definició 23. *Una variable aleatòria* X *és* **absolutament contínua** *si té associada una funció de densitat* f *tal que per a tot* a ≤ b*:*

$$P(a<X<b)=\int_a^b f(y)\,dy.$$

De fet, la probabilitat que X es trobi entre dos valors a i b és l'àrea que ens queda sota la corba de densitat entre els dos punts.

Observació 15. La funció de densitat ens determina la llei de la variable aleatòria, igual que ho fa la funció de distribució. Observem que si coneixem la funció de densitat podem calcular la funció de distribució i viceversa.

$$f(x) = F'(x)$$
$$F(x) = \int_{-\infty}^{x} f(y)\,dy.$$

I, per tant, tenim les següents relacions per a qualssevol valors a i b,

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(y)\,dy,$$
$$P(a \leq X) = 1 - F(a) = \int_a^{\infty} f(y)\,dy = 1 - \int_{-\infty}^{a} f(y)\,dy,$$
$$P(X \leq b) = F(b) = \int_{-\infty}^{b} f(y)\,dy,$$
$$P(X = a) = \int_a^a f(y)\,dy = 0.$$

Observació 16. Quan treballem amb variables aleatòries absolutament contínues, podrem canviar sempre $\leq$ per $<$ i $\geq$ per $>$, ja que la probabilitat que la variable sigui igual a un valor és sempre 0. Per tant,

$$\begin{aligned} P(a \leq X \leq b) &= P(a < X < b) = P(a < X \leq b) = P(a \leq X < b), \\ P(a \leq X) &= P(a < X), \\ P(X \leq b) &= P(X < b). \end{aligned}$$

Abans de fer un exemple, definirem la funció indicador:

$$\mathbb{1}_{(a,b)}(x) = \begin{cases} 1, & \text{si } a < x < b, \\ 0, & \text{altrament.} \end{cases}$$

Exemple 37. *Sigui* X *una variable aleatòria amb funció de densitat*

$$f(x) = c\,x\,\mathbb{1}_{(2,4)}(x).$$

Ens preguntem:

1) *Quin valor ha de prendre* c *per tal que* f(x) *sigui efectivament una funció de densitat?*
2) *Calculeu la funció de distribució de* X.
3) *Calculeu les següents probabilitats:* $P(X < 3)$, $P(X > 5)$, $P\left(3 < X \leq \frac{7}{2}\right)$ *i* $P\left(X \leq \frac{7}{2} | X > 3\right)$.

1) Per tal que f sigui una funció de densitat cal que sigui no-negativa i que compleixi que $\int_2^4 f(x)\,dx = 1$. Utilitzarem aquesta última condició per a trobar el valor de c:

$$1 = \int_2^4 c\,x\,dx = c\left[\frac{x^2}{2}\right]_2^4 = 6 \cdot c \Rightarrow c = \frac{1}{6}.$$

2) Calculem la funció de distribució utilitzant que $F(x)=\int_{-\infty}^{x} f(y)\,dy$:

$$F(x)=\begin{cases} \displaystyle\int_{-\infty}^{x} 0\,dy=0, & x<2, \\ \displaystyle\int_{-\infty}^{2} 0\,dy+\int_{2}^{x}\frac{1}{6}y\,dy=\left(\frac{x^2}{12}-\frac{1}{3}\right), & 2\le x<4, \\ \displaystyle\int_{-\infty}^{0} 0\,dy+\int_{2}^{4}\frac{1}{6}y\,dy+\int_{4}^{x} 0\,dy=1, & 4\le x. \end{cases}$$

3) Ara, per calcular les probabilitats que ens demanen, podem utilitzar la funció de densitat:

$$\begin{aligned} P(X<3) &= \int_{2}^{3}\frac{1}{6}x\,dx=\frac{9}{12}-\frac{4}{12}=\frac{5}{12}, \\ P(X>5) &= \int_{5}^{+\infty} 0\,dx=0, \\ P\left(3<X\le\frac{7}{2}\right) &= \int_{3}^{\frac{7}{2}}\frac{1}{6}x\,dx=\frac{49}{48}-\frac{9}{12}=\frac{13}{48} \end{aligned}$$

o bé la funció de distribució

$$\begin{aligned} P(X<3) &= F(3)=\frac{5}{12}, \\ P(X>5) &= 1-F(5)=1-1=0, \\ P\left(3\le X<\frac{7}{2}\right) &= F\left(\frac{7}{2}\right)-F(3)=\frac{33}{48}-\frac{5}{12}=\frac{13}{48}. \end{aligned}$$

I podem calcular l'última probabilitat que ens demanen a partir de les que ja tenim:

$$P\left(X\le\frac{7}{2}|X>3\right)=\frac{P\left(3<X\le\frac{7}{2}\right)}{P(X>3)}=\frac{\frac{13}{48}}{1-\frac{5}{12}}=\frac{13}{28}.$$

Exemple 38. *Sigui* X *una variable aleatòria amb funció de distribució*

$$F(x)=\begin{cases} 0 & si \quad x<0 \\ \dfrac{x^2}{2} & si \quad 0\le x<1 \\ 1-\dfrac{(2-x)^2}{2} & si \quad 1\le x<2 \\ 1 & si \quad x\ge 2 \end{cases}$$

X *és una variable aleatòria absolutament contínua? Podem trobar la funció de densitat?*

Observem que efectivament la variable aleatòria X és una variable absolutament contínua, ja que la seva funció de distribució és contínua i derivable; per tant, podem calcular la seva funció de densitat. La funció de densitat serà la derivada de la funció de distribució.

$$f(x)=F'(x)=\begin{cases} 0 & \text{si} \quad x<0 \\ x & \text{si} \quad 0\leq x<1 \\ 2-x & \text{si} \quad 1\leq x<2 \\ 0 & \text{si} \quad x\geq 2 \end{cases}$$

3.3.1. Independència de variables absolutament contínues

Observem que la definició de variables aleatòries discretes independents que hem explicat anteriorment

$$P(Y=y, Z=z)=P(Y=y)P(Z=z)$$

ara no té sentit, ja que totes aquestes probabilitats ara seran sempre zero. Necessitem per tant una nova definició que ens permeti generalitzar la donada per al cas discret.

Definició 24. *Dues variables aleatòries* Y *i* Z *són* **independents** *si*

$$P(Y\leq y, Z\leq z)=P(Y\leq y)P(Z\leq z),$$

per a tota possible parella de valors y *i* z.

Observació 17. Aquesta definició d'independència també ens serveix per a les variables aleatòries discretes.

3.3.2. Esperança i variància de variables aleatòries absolutament contínues

La idea d'esperança i variància d'una variable aleatòria absolutament contínua és exactament la mateixa que en el cas discret: el que varia és la manera de calcular-la.

Definició 25. *L'***esperança** *d'una variable aleatòria absolutament contínua* X *és la mitjana ponderada dels valors que pren i es calcula com*

$$E(X)=\int_{-\infty}^{\infty} x f(x)\,dx,$$

on f(x) *és la funció de densitat associada a la variable aleatòria* X.

I aquesta definició es pot estendre al cas que vulguem calcular l'esperança de $g(X)$; llavors tindrem que

$$E(g(X))=\int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

Pel que fa a la variància, recordem que la definició era

$$Var(X)=E\left((X-E(X))^2\right)=E(X^2)-E(X)^2;$$

per tant,

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\,dx;$$

o bé, equivalentment,

$$Var(x) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - E(X)^2.$$

Observació 18. Totes les propietats que hem donat amb referència a l'esperança i la variància per al cas de variables aleatòries discretes es compliran també en el cas de variables aleatòries absolutament contínues.

Exemple 39. *Calculeu l'esperança i la variància de la variable aleatòria* X *que té per funció de densitat*

$$f(x) = \frac{1}{12}(7-2x)\mathbb{1}_{(0,3)}(x).$$

$$\begin{aligned}
E(X) &= \int_0^3 x f(x)\,dx = \int_0^3 x\frac{1}{12}(7-2x)\,dx = \int_0^3 \frac{7}{12}x\,dx - \int_0^3 \frac{1}{6}x^2\,dx = \frac{9}{8},\\
Var(X) &= E(X^2) - \left(\frac{9}{8}\right)^2 = \int_0^3 x^2 f(x)\,dx - \frac{81}{64} = \int_0^3 x^2\frac{1}{12}(7-2x)\,dx - \frac{81}{64}\\
&= \int_0^3 \frac{7}{12}x^2\,dx - \int_0^3 \frac{1}{6}x^3\,dx - \frac{81}{64} = \frac{15}{8} - \frac{81}{64} = \frac{9}{64}.
\end{aligned}$$

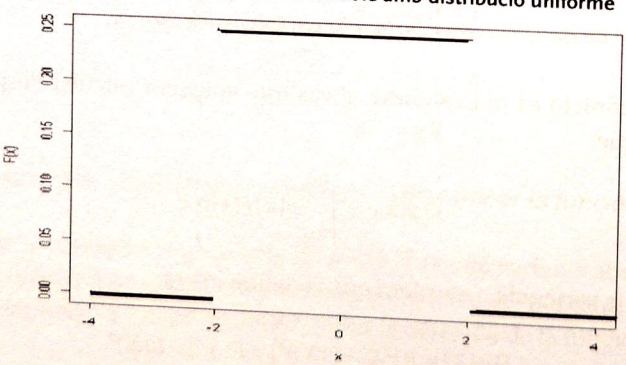
3.3.3. Exemples de variables aleatòries absolutament contínues

Estudiarem ara algunes de les distribucions contínues més utilitzades.

- **Distribució uniforme.** Una variable aleatòria X segueix una distribució uniforme, $U(a,b)$, on $-\infty < a < b < \infty$, si té la següent funció de densitat:

$$f(x) = \frac{1}{b-a}\mathbb{1}_{(a,b)}(x).$$

Funció de densitat d'una variable amb distribució uniforme



I en aquest cas direm que $X \sim U(a,b)$.

Aleshores podem calcular fàcilment la seva funció de distribució. Com que per a $x \in (a,b)$ tindrem

$$F(x) = \int_a^x \frac{1}{b-a}\, dy = \frac{x-a}{b-a},$$

ens queda que

$$F(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x < b, \\ 1, & b \le x. \end{cases}$$

Aquest tipus de llei apareix quan volem agafar un punt a l'atzar en un interval.

Pel que fa a l'esperança i la variància d'una variable aleatòria $X \sim U(a,b)$ són:

$$[E(X) = \frac{a+b}{2}, \qquad Var(X) = \frac{(b-a)^2}{12}.$$

Exemple 40. *Imagineu que per a fer un estudi sobre seguretat hem escollit un punt a l'atzar a la via del TAV entre Barcelona i Madrid que suposarem que té 600 km.*

Aquí tindríem una variable $X \sim U(0,600)$. Si ens demanen quina és la probabilitat que el punt estigui a menys de 150 km de Barcelona, és clar, per la uniformitat, que la probabilitat serà $\frac{1}{4}$. Fixeu-vos també que si ens demanen quina és la probabilitat d'escollir el punt 183.4265 haurem de dir que 0.

Podem fer aquests càlculs d'una manera més explícita: la variable aleatòria X té densitat

$$f(x) = \frac{1}{600}\, \mathbb{1}_{(0,600)}(x)$$

i funció de distribució

$$F(x) = \begin{cases} 0, & x < 0, \\ \dfrac{x}{600}, & 0 \le x < 600, \\ 1, & 600 \le x. \end{cases}$$

I volem calcular

$$\begin{aligned} P(X \le 150) &= F(150) = \frac{150}{600} = \frac{1}{4} \\ &= \texttt{punif(150,0,600)} = 0.25 \end{aligned}$$

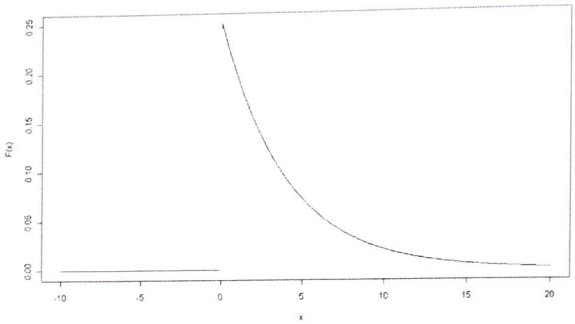
i

$$P(X = 183.4265) = 0.$$

- **Distribució exponencial.** Una variable aleatòria Y segueix una distribució exponencial de paràmetre λ, on $\lambda > 0$, $Y \sim \exp(\lambda)$, si té la següent funció de densitat:

$$f(x) = \lambda \exp(-\lambda x)\, \mathbb{1}_{(0,\infty)}(x).$$

Funció de densitat d'una variable aleatòria amb distribució exponencial



Aleshores la seva funció de distribució és:

$$F(x) = \begin{cases} 0, & x < 0, \\ \displaystyle\int_0^x \lambda \exp(-\lambda y) = 1 - \exp(-\lambda x), & 0 \leq x. \end{cases}$$

Aquest tipus de llei apareix bàsicament en dos tipus de situacions: en el temps de vida d'algun element –per exemple, el temps que tarda a fondre's una bombeta o el temps que tarda a espatllar-se un disc dur– o en la durada d'un servei –el temps que es passa una persona a la perruqueria o la durada d'una trucada telefònica.

Pel que fa a l'esperança i la variància d'una variable aleatòria $X \sim Exp(\lambda)$ són:

$$E(X) = \frac{1}{\lambda}, \qquad Var(X) = \frac{1}{\lambda^2}.$$

Exemple 41. *Suposem que la durada d'una trucada telefònica segueix una distribució exponencial de paràmetre $\frac{1}{2}$. Anomenem* X *aquesta variable aleatòria; aleshores sabem que té funció de densitat*

$$f(x) = \frac{1}{2} \exp\left(-\frac{1}{2}x\right) \mathbb{1}_{(0,\infty)}(x)$$

i funció de distribució

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - \exp\left(-\frac{1}{2}x\right), & 0 \leq x. \end{cases}$$

Ara ja podem calcular probabilitats. Les farem de tres maneres –utilitzant la funció de densitat, utilitzant la de distribució i amb R–, com per exemple:

1) Probabilitat que la trucada duri més de 3 minuts:

$$P(X>3)=\int_3^{\infty}\frac{1}{2}\exp\left(-\frac{1}{2}x\right)dx=\exp\left(-\frac{3}{2}\right);$$

o també:

$$P(X>3)=1-P(X\leq 3)=1-F(3)=\exp\left(-\frac{3}{2}\right);$$

o bé:

$$P(X>3)=1-P(X\leq 3)=1-\texttt{pexp(3,1/2)}=0.2231302.$$

2) Probabilitat que la trucada duri exactament 3 minuts:

$$P(X=3)=0.$$

3) Probabilitat que, si sabem que la durada ha durat més d'1 minut, en duri menys de 2:

$$P(X<2|X>1)=\frac{P(1<X<2)}{P(X>1)},$$

on

$$\begin{aligned}P(1<X<2) &= \int_1^2\frac{1}{2}\exp\left(-\frac{1}{2}x\right)dx=\exp\left(-\frac{1}{2}\right)-\exp(-1)\\ &= F(2)-F(1)=\exp\left(-\frac{1}{2}\right)-\exp(-1)\\ &= \texttt{pexp(2,1/2)}-\texttt{pexp(1,1/2)}=0.2386512,\\ P(X>1) &= \int_1^{+\infty}\frac{1}{2}\exp\left(-\frac{1}{2}x\right)dx=\exp\left(-\frac{1}{2}\right)\\ &= 1-P(X\leq 1)=1-F(1)=\exp\left(-\frac{1}{2}\right)\\ &= 1-\texttt{pexp(1,1/2)}=0.6065307.\end{aligned}$$

Per tant,

$$P(X<2|X>1)=\frac{\exp\left(-\frac{1}{2}\right)-\exp(-1)}{\exp\left(-\frac{1}{2}\right)}=1-\exp\left(-\frac{1}{2}\right).$$

Estudiarem a part la distribució normal o gaussiana atesa la seva importància i les seves múltiples aplicacions.

3.3.4. La distribució normal

Aquesta és sense cap mena de dubte la llei més utilitzada en l'estadística. Això és perquè apareix en molts fenòmens naturals, com en quasi totes les característiques biològiques que

vulgueu estudiar –pes dels porcs, glòbuls vermells a la sang, etc. Precisament el nom de *llei normal* es deu al fet que es pensava que era el patró «natural» de les distribucions.

Definició 26. *Direm que una variable aleatòria* X *segueix una* **llei normal** *o* **gaussiana** *amb paràmetres* μ *i* σ^2*, és a dir que* $X \sim N(\mu, \sigma^2)$*, si té densitat*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2),$$

per a tot x *real.*

Observació 19:

1) Una variable aleatòria normal és una variable aleatòria absolutament contínua.
2) La distribució normal queda determinada a partir dels dos paràmetres μ *i* σ^2.

$$\text{si } X \sim N(\mu, \sigma^2), \text{ aleshores } E(X) = \mu \text{ i } Var(X) = \sigma^2.$$

   És a dir, la μ ens indica el centre de la distribució mentre que σ ens explica la seva dispersió. Observeu que la μ pot ser qualsevol valor real mentre que la σ s'agafa sempre positiva.
3) La funció de distribució de la llei normal no té una fórmula explícita. És clar que per a tot x es té

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) dy.$$

   Aquesta integral, però, no la sabem calcular explicítament. De totes maneres, avui en dia tenim el programari que ens calcula aquesta funció.

Definició 27. *Una variable aleatòria* Z *segueix una llei* **normal estàndard** *si* $Z \sim N(0, 1)$*. És a dir, es tracta d'una distribució normal amb esperança 0 i variància 1 i amb funció de densitat*

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right),$$

per a tot x *real.*

Donada qualsevol distribució normal la podem convertir en una distribució normal estàndard per mitjà d'una transformació lineal. Aquest procés s'anomena *estandardització.*

Proposició 5. *Si* X *és una variable aleatòria* $N(\mu, \sigma^2)$*, aleshores la variable aleatòria* $Z = \frac{X-\mu}{\sigma}$ *segueix una llei normal estàndard.*

Una conseqüència bàsica d'aquest resultat és que els càlculs de probabilitats al voltant de l'esperança en termes de la desviació són els mateixos per a totes les distribucions normals. És a dir, que si X és una variable aleatòria $N(\mu, \sigma^2)$ i Z una variable aleatòria normal estàndard, aleshores per a tot k positiu es té:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P\left(-k \leq \frac{X-\mu}{\sigma} \leq k\right) = P(-k \leq Z \leq k).$$

La propietat d'estandardització és un cas particular del resultat més general següent:

Proposició 6:

1) *Si* X *i* Y *són dues variables aleatòries independents amb lleis* $N(\mu_1, \sigma_1^2)$ *i* $N(\mu_2, \sigma_2^2)$ *respectivament, aleshores la variable aleatòria* $Z = X + Y$ *segueix una llei* $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
2) *Si* X *és una variable aleatòria* $N(\mu, \sigma^2)$, *aleshores la variable aleatòria* $Z = aX + b$ *segueix una llei* $N(a\mu + b, a^2\sigma^2)$.

Exemple 42. *El pes dels inquilins d'un bloc de pisos segueix una distribució normal de mitjana 65 i variància 25. Quina és la probabilitat que si agafem un veí a l'atzar pesi menys de 77 kg? I que pesi exactament 56 kg?*

Si X = pes d'un inquilí qualsevol, aleshores X és una variable aleatòria que segueix una distribució $N(65, 25)$.

$$P(X < 77) = P\left(\frac{X-65}{5} \leq \frac{77-65}{5}\right) = P\left(N(0,1) < \frac{12}{5}\right) = \texttt{pnorm(12/5)} = 0.9918025$$

$$P(X = 56) = 0$$

4. TEOREMA DEL LÍMIT CENTRAL

El Teorema del Límit Central ens diu que la distribució de la suma de moltes variables aleatòries independents amb la mateixa distribució, independentment de quina sigui aquesta, és aproximadament una llei normal.

4.1. La mitjana mostral

Volem estudiar com es comporta la mitjana mostral. Considerem $X_1, X_2, \ldots, X_n$ una mostra aleatòria d'una distribució determinada, o sigui, n variables aleatòries independents idènticament distribuïdes (v.a.i.i.d). Aleshores podem definir-ne la mitjana mostral.

Definició 28. *La* **mitjana mostral** $\overline{X}_n$ *d'una mostra aleatòria* $X_1, X_2, \ldots, X_n$ *d'una distribució determinada és una nova variable aleatòria que es defineix com*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Ens plantegem ara quins valors tindran l'esperança i la variància de la mitjana mostral. Com que $X_1, X_2, \ldots, X_n$ són v.a.i.i.d, totes tenen la mateixa esperança i variància: $E(X_i) = \mu$ i $Var(X_i) = \sigma^2$ per a $i = 1, \ldots, n$. Utilitzant les propietats de l'esperança i de la variància podem calcular el que volíem:

$$E(\overline{X}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu,$$

$$Var(\overline{X}_n) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Observem així que l'esperança de la mitjana mostral coincideix amb l'esperança de X_i. Pel que fa a la desviació estàndard de la mitjana mostral (que rep el nom d'**error estàndard de la mostra**), serà $\frac{\sigma}{\sqrt{n}}$.

Observació 20. Si augmentem la mida de la mostra, aleshores la variància de $\overline{X}_n$ es fa més petita, és a dir, la dispersió de la mitjana mostral al voltant de μ disminueix.

El problema és que necessitem una mica més d'informació de la mitjana mostral; no en fem prou coneixent-ne l'esperança i la variància. Estudiem ara en diferents casos com es comporta la mitjana mostral.

4.1.1. Mitjana mostral per a variables normals

Recordem en primer lloc el següent resultat que hem donat en el capítol anterior.

Proposició 7. *Si* X *i* Y *són dues variables aleatòries independents amb lleis* $N(\mu_1, \sigma_1^2)$ *i* $N(\mu_2, \sigma_2^2)$ *respectivament, aleshores la variable aleatòria*

$$Z = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Observem que aquest resultat es pot estendre al cas que tinguem n variables que segueixin una distribució normal. Aleshores ens quedarà:

Proposició 8. *Si* $X_i \sim N(\mu_i, \sigma_i^2)$ *per a* $i = 1, 2, \ldots, n$, *són* n *variables aleatòries independents, aleshores la variable aleatòria*

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

A partir d'aquests resultats obtenim fàcilment el resultat de la següent proposició sobre la distribució de la mitjana mostral.

Proposició 9. *Si* $X_1, X_2, \ldots, X_n$ *és una mostra aleatòria d'una distribució normal* $N(\mu, \sigma^2)$ *(això vol dir que* $X_i \sim N(\mu, \sigma^2)$*), aleshores la mitjana mostral*

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Aquest resultat és molt interessant, però el problema amb què ens trobarem quan el vulguem utilitzar és que en general coneixerem el valor de μ, però, en canvi, no coneixerem el de σ. Veiem ara com varia el resultat segons si el valor de σ és conegut o no ho és.

Cas de σ coneguda

En el cas en què μ i σ són conegudes, podem aplicar simplement la proposició anterior i per tant tenim que

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim N(0, 1).$$

Cas de σ desconeguda

En el cas en què el valor de σ no és conegut, no podem utilitzar els resultats anteriors com hem fet abans, ja que ens apareixeria σ i no sabem quant val. Aleshores el que farem serà substituir el valor de σ per una estimació. L'estimació que farem servir és la desviació mostral:

$$S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}.$$

Un cop solucionat aquest problema, ens trobem que el que no coneixem és la distribució de $\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}}$, ja que ara no és cert que segueixi una llei normal. Per això, en el cas que el valor de σ sigui desconegut i utilitzem en el seu lloc la desviació mostral, farem servir el resultat següent:

Proposició 10. *Si $X_1, X_2, \ldots, X_n$ és una mostra aleatòria simple d'una distribució normal $N(\mu, \sigma^2)$, aleshores*

$$\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}} \sim t_{n-1}.$$

on t_{n-1} és una distribució **t de Student amb n – *1* graus de llibertat**.

La distribució t_n, **t de Student amb *n* graus de llibertat** és una distribució:

- simètrica al voltant del 0 (igual que $N(0,1)$),
- la seva variància és més gran que la d'una $N(0,1)$, $Var(t_n) = \frac{n}{n-2}$,
- per a valors de n grans ($n > 30$) es pot aproximar per una $N(0,1)$.

Exemple 43. *El nivell de glucosa a la sang quan ens aixequem al matí segueix una distribució normal amb mitjana $\mu = 87$ mg/dL i desviació estàndard $\sigma = 10$. Agafem 150 persones a l'atzar i els analitzem el nivell de glucosa quan s'aixequen al matí. Quina és la probabilitat que la mitjana observada sigui menor que 83?*

La distribució de la mitjana mostral del nivell de glucosa és una distribució normal amb esperança 87 i desviació estàndard $\frac{10}{\sqrt{150}} = 0.8165$, és a dir que

$$\frac{\overline{X}_n - 87}{0.8165} \sim N(0,1).$$

Així, hem de calcular

$$P(\overline{X}_n < 83) = P\left(\frac{\overline{X}_n - 87}{0.8165} < \frac{-4}{0.8165}\right) = P(N(0,1) < -4.8989) = \texttt{pnorm(-4.8989)} = 0.00000048.$$

Exemple 44. *Volem mesurar també el nivell de bilirubina. Sabem que el nivell de bilirubina segueix una distribució normal amb desviació estàndard desconeguda. Agafem ara 17 persones a l'atzar i els analitzem el nivell de bilirubina. Si a partir de les dades obtingudes ens surt una desviació típica mostral de 0.2, quina és la probabilitat que la distància entre la mitjana poblacional i la mitjana mostral sigui més gran que 0.1?*

En aquest cas tenim que la desviació estàndard és desconeguda i que la desviació mostral és $s = 0.2$. Aleshores fixeu-vos que

$$\frac{\overline{X}_n - \mu}{\frac{0.2}{\sqrt{16}}} = \frac{\overline{X}_n - \mu}{0.05} \sim t_{16}.$$

Ens demanen que calculem

$$\begin{aligned} P(|\overline{X}_n - \mu| > 0.1) &= 1 - P(-0.1 \leq \overline{X}_n - \mu \leq 0.1) \\ &= 1 - P\left(\frac{-0.1}{0.05} \leq \frac{\overline{X}_n - \mu}{0.05} \leq \frac{0.1}{0.05}\right) \\ &= 1 - P(-2 \leq t_{16} \leq 2), \end{aligned}$$

on t_{16} és una t de Student amb 16 graus de llibertat. Utilitzant

$$P(-2 \leq t_{16} \leq 2) = P(t_{16} \leq 2) - P(t_{16} < -2) = 2 \cdot P(t_{16} \leq 2) - 1 = 2*\texttt{pt}(2,16) - 1 = 0.9372,$$

obtenim que

$$P\left(\left|\overline{X}_n - \mu\right| > 0.1\right) = 1 - 0.9372 = 0.0628.$$

4.2. Teorema del límit central

El teorema del límit central ens servirà per estudiar el comportament de la suma de n variables aleatòries independents i idènticament distribuïdes per a valors de n grans, sigui quina sigui la distribució de les variables aleatòries. Fixeu-vos que els resultats de l'apartat anterior els podem utilitzar només en el cas en què les variables aleatòries segueixin una distribució normal.

4.2.1. Cas general

Ara veurem el cas general del teorema del límit central, però la importància d'aquest teorema rau en el fet que es pot utilitzar sigui quina sigui la distribució de la nostra mostra.

Una versió general del teorema és la següent:

Teorema 4. *Sigui* $X_1, X_2, \ldots, X_n$ *una mostra aleatòria d'una distribució amb esperança* μ *i variància* σ^2. *Aleshores, quan* n *és prou gran* ($n > 30$),

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}},$$

es comporta com una distribució normal estàndard.

Observació 21. Igual que hem fet en el cas en què hem estudiat la mitjana mostral de variables aleatòries normals, en el cas que no coneguéssim σ podríem utilitzar la desviació típica mostral i aleshores tindríem que

$$\frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}}$$

es comporta com una $N(0,1)$ (com que els valors de n són grans, la t de Student amb n graus de llibertat es comporta com una $N(0,1)$).

Exemple 45. *Un servei tècnic de reparació d'ordinadors tarda una mitjana de 3 hores a arreglar un ordinador, amb una desviació estàndard de 100 minuts. Durant la setmana vinent aquest servei ha d'arreglar 100 ordinadors. Per fer-ho, disposa de 7 treballadors, és a dir, de 280 hores de treball. Quina és la probabilitat que els pugui arreglar tots?*

Considerem ara, per a cada ordinador, una variable X_i que ens doni el temps que tarda el servei a arreglar aquest ordinador. Per les dades que ens donen ja sabem que la seva es-

perança d'aquesta variable serà 180 –posem totes les unitats en minuts– i la seva desviació estàndard 100. No sabem, però, si aquesta variable segueix una distribució normal.

Com que 280 hores representen 16800 minuts, el que nosaltres volem calcular és

$$P\left(\sum_{i=1}^{100} X_i \leq 16800\right).$$

Per calcular aquesta probabilitat, tindrem en compte que, segons el teorema del límit central, la variable aleatòria

$$\frac{\sum_{i=1}^{100} X_i - 18000}{100\sqrt{100}} = \frac{\sum_{i=1}^{100} X_i - 18000}{1000}$$

es comporta com una normal estàndard. Aleshores:

$$\begin{aligned} P\left(\sum_{i=1}^{100} X_i \leq 16800\right) &= P\left(\frac{\sum_{i=1}^{100} X_i - 18000}{1000} \leq \frac{16800-18000}{1000}\right) \\ &= P(N(0,1) \leq -1.2) = \texttt{pnorm(-1.2)} = 0.1150. \end{aligned}$$

És a dir, la probabilitat que es puguin arreglar tots els ordinadors és força petita.

Estudiem ara el cas en què les variables aleatòries segueixin una distribució binomial.

4.2.2. Cas binomial

Sigui $X_1, X_2, \ldots, X_n$ una mostra aleatòria d'una distribució $Ber(p)$. Si considerem la suma d'aquestes variables aleatòries Bernouilli independents, sabem que $X_1 + \cdots + X_n \sim B(n,p)$. Recordem que l'esperança i la variància d'una $B(n,p)$ són respectivament np i $np(1-p)$.

En aquest cas, tenim el resultat següent:

Teorema 5. *Sigui* $X_1, X_2, \ldots, X_n$ *una mostra aleatòria d'una distribució* Ber(p). *Aleshores, per a tot* a < b *es compleix:*

$$\lim_{n\to\infty} P\left(a < \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \leq b\right) = P(a < N(0,1) \leq b).$$

El que ens diu aquest teorema és que quan el valor de n ($n > 30$) és prou gran, aleshores

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$$

es comporta com una distribució $N(0,1)$, o sigui que, a efectes pràctics, si la n és gran, podem substituir una llei per l'altra.

Observació 22. Podem reescriure el resultat del teorema anterior de diverses maneres. Sabem que $\sum_{i=1}^{n} X_i$ es distribueix com una $B(n,p)$, però quan n és prou gran ($n > 30$), el teorema ens diu que:

- $\dfrac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$ es comporta com una distribució $N(0,1)$.

- $\displaystyle\sum_{i=1}^{n} X_i - np$ es comporta com una distribució $N(0, np(1-p))$.

- $\displaystyle\sum_{i=1}^{n} X_i$ es comporta com una distribució $N(np, np(1-p))$.

- $\overline{X}_n$ es comporta com una distribució $N\left(p, \frac{p(1-p)}{n}\right)$.

Exemple 46. *Si tirem un dau perfecte 1000 vegades, quina és la probabilitat que traguem més de 150 sisos?*

Considerem la variable aleatòria $X =$ *nombre de sisos obtinguts*. És clar que $X \sim B(1000, \frac{1}{6})$. Podem, per tant, calcular la probabilitat que ens demanen, utilitzant la distribució binomial

$$\begin{aligned} P(X > 150) &= 1 - P(X \leq 150) = 1 - \sum_{k=0}^{150} P(X = k) \\ &= 1 - \sum_{k=0}^{150} \binom{150}{k} \left(\frac{1}{6}\right)^k \left(1 - \frac{1}{6}\right)^{1000-k} = 0.916311, \end{aligned}$$

però aquesta quantitat sembla una mica llarga de calcular.

També ho podem calcular utilitzant l'aproximació per la distribució normal: observeu que una distribució $B(1000, \frac{1}{6})$ es comporta com una distribució

$$N\left(1000 \cdot \frac{1}{6}, 1000 \cdot \frac{1}{6} \cdot \frac{5}{6}\right).$$

Així, si considerem una variable aleatòria $Y \sim N\left(\frac{1000}{6}, \frac{5000}{36}\right)$, tindrem

$$\begin{aligned} P(X > 150) &\simeq P(Y > 150) = P\left(\frac{Y - \frac{1000}{6}}{\sqrt{\frac{5000}{36}}} > \frac{150 - \frac{1000}{6}}{\sqrt{\frac{5000}{36}}}\right) \\ &= P(N(0,1) > -1.4142) = \texttt{1-pnorm(-1.4142)} = 0.921349. \end{aligned}$$

A partir d'ara tindrem unes dades $x_1, \ldots, x_n$ (una mostra) i intentarem a partir d'elles obtenir informació sobre el procés que les ha generat. Quan parlem d'obtenir informació no volem dir descriure les dades, com es fa en estadística descriptiva, sinó que entendrem que les nostres dades són una realització d'una mostra amb una distribució desconeguda. Aquesta distribució no serà desconeguda del tot, sinó que, de fet, dependrà d'un paràmetre desconegut que serà el nostre principal objectiu. Aquest paràmetre podrà ser: la mitjana, la variància, una proporció...

El primer mètode que introduirem serà el càlcul d'intervals de confiança. L'objectiu és trobar un interval amb un percentatge bastant alt de probabilitats, de manera que el valor real del paràmetre que estem estimant estigui dins el nostre interval.

A continuació veurem aquestes idees amb exemples i donarem alguns mètodes per a construir aquests intervals.

5.1. Exemple

Exemple 47. *Hem observat les hores treballades mensualment per 10 treballadors d'una empresa:*

180 165 168 192 195 187 181 177 175 186.

Suposem que sabem que la desviació estàndard de la població val 4. Suposem també que les dades són normals.

Tenim, per tant, una mostra aleatòria simple $X_1, \ldots, X_n$ *d'una distribució* $N(\mu, 16)$, *és a dir, amb* $\sigma = 4$. *Volem trobar un interval de confiança al 95% per a* μ.

1) Sabem que:

$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}_n - \mu}{\frac{4}{\sqrt{10}}} \sim N(0,1)$$

(recordem per l'enunciat que: $\sigma = 4$ i $n = 10$).

2) Com que volem un nivell de confiança del 95%, hem de determinar u_1 i u_2 tals que

$$P(u_1 \leq Z \leq u_2) = 0.95.$$

Podem agafar, per exemple, $u_1 = -1.96$ i $u_2 = 1.96$, que són dos valors que compleixen la propietat anterior (normalment utilitzarem intervals simètrics). Tenim així que

$$-1.96 \leq \frac{\overline{X}_n - \mu}{\frac{4}{\sqrt{10}}} \leq 1.96,$$

que podem reescriure com

$$-1.96\frac{4}{\sqrt{10}} \leq \overline{X}_n - \mu \leq 1.96\frac{4}{\sqrt{10}}$$

o

$$\overline{X}_n - 1.96\frac{4}{\sqrt{10}} \leq \mu \leq \overline{X}_n + 1.96\frac{4}{\sqrt{10}}.$$

És a dir, que obtenim l'interval per a μ

$$\left[\overline{X}_n - 1.96\frac{4}{\sqrt{10}}, \overline{X}_n + 1.96\frac{4}{\sqrt{10}}\right].$$

3) Substituint segons les nostres observacions, obtenim l'interval

$$\left[\overline{x} - 1.96\frac{4}{\sqrt{10}}, \overline{x} + 1.96\frac{4}{\sqrt{10}}\right].$$

Com que a partir de les dades podem calcular la mitjana $\overline{x} = 180.6$, obtenim l'interval

$$\left[180.6 - 1.96\left(\frac{4}{\sqrt{10}}\right), 180.6 + 1.96\left(\frac{4}{\sqrt{10}}\right)\right] = [178.121,\ 183.079].$$

Observació 23. Sobre la notació i les idees generals dels intervals de confiança.

1) Quan parlem d'interval de confiança al 90%, serà el mateix que si diem que estem fent un interval de confiança amb nivell de confiança $\gamma = 0.9$.

2) L'interval que s'obté depèn de les observacions inicials. És a dir, per a cada conjunt d'observacions –en el nostre exemple, per a cada grup de 10 treballadors– obtindrem un interval diferent! Aleshores, com hem d'interpretar el nivell de confiança? Un interval de confiança per a μ amb un nivell de confiança del 95% ens indica que, per cada 100 mostres que agafem, dels 100 intervals diferents que obtindrem, 95 d'aquests intervals contindran el valor real de μ, desconegut per nosaltres.

3) La tria de u_1 i de u_2 hauria pogut ser diferent; podrien ser qualssevol valors que complissin

$$P(u_1 \leq Z \leq u_2) = 0.95.$$

De totes maneres, per comoditat acostumem a agafar valors en què l'interval sigui simètric.

4) Si l'interval és simètric, com passa en aquest cas, la longitud de l'interval és:

$$L = 2 \cdot 1.96\frac{4}{\sqrt{10}}.$$

5) Si ens interessa estar molt segurs d'on és el paràmetre i augmentem el nivell de confiança, a mesura que augmentem la confiança obtindrem intervals més grans. En el nostre exemple, el cas extrem seria una confiança del 100%, en què hauríem d'agafar $u_1 = -\infty$ i $u_2 = \infty$ i obtindríem un interval amb tots els valors reals.

6) La manera d'obtenir intervals petits –que és el que ens interessa– és augmentar la mida de la mostra. Això és degut al factor $\frac{1}{\sqrt{n}}$ que apareix. Però el fet d'augmentar la mida de la mostra és car; per tant, hem de trobar l'equilibri entre el cost que representa augmentar la mida de la mostra i el nivell de confiança que necessitem.

7) La funció Z que hem utilitzat,

$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1),$$

s'anomena **funció pivotant**. Una funció és una funció pivotant si la seva distribució és coneguda i no depèn del paràmetre μ.

5.2. Intervals de confiança per a una població normal

En aquest apartat suposarem que tenim una mostra aleatòria $X_1, \ldots, X_n$ d'una $N(\mu, \sigma^2)$.

5.2.1. Intervals de confiança per a μ, on σ és coneguda

Volem construir un interval de confiança per a μ suposant que el valor de σ és conegut amb un nivell de confiança γ. Estem, doncs, en el cas de l'exemple anterior.

Els passos que seguim per calcular-lo són:

1) Utilitzarem que

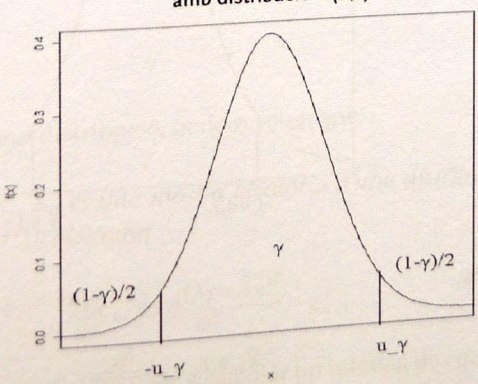
$$Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

2) Calculem el valor crític u_γ tal que

$$P\left(-u_\gamma \leq Z \leq u_\gamma\right) = \gamma \Rightarrow P(Z < u_\gamma) = 1 - \frac{1-\gamma}{2},$$

en aquest cas $u_\gamma = \texttt{qnorm}\left(1 - \frac{1-\gamma}{2}\right)$.

Funció de densitat d'una variable aleatòria amb distribució N(0,1)



3) L'interval serà:

$$\overline{X}_n \pm u_\gamma \frac{\sigma}{\sqrt{n}},$$

o el que és el mateix:

$$\left[\overline{X}_n - u_\gamma \frac{\sigma}{\sqrt{n}}, \overline{X}_n + u_\gamma \frac{\sigma}{\sqrt{n}}\right].$$

5.2.2. Intervals de confiança per a μ, on σ és desconeguda

Volem construir un interval de confiança per a μ suposant que σ és desconeguda, amb un nivell de confiança γ.

Els passos que seguim per a calcular-lo són:

1) Utilitzarem la desviació estàndard mostral en lloc de la σ; en aquest cas ens queda que

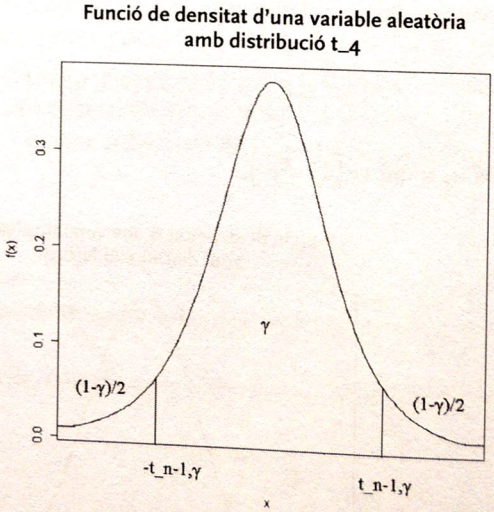
$$T_{n-1} = \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}} \sim t_{n-1},$$

on recordem que $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

2) Calculem el valor crític $t_{n-1,\gamma}$ tal que

$$P\left(-t_{n-1,\gamma} \leq T_{n-1} \leq t_{n-1,\gamma}\right) = \gamma \Rightarrow P(T_{n-1} < t_{n-1,\gamma}) = 1 - \frac{1-\gamma}{2},$$

en aquest cas $t_{n-1,\gamma} = \mathtt{qt}\left(1 - \frac{1-\gamma}{2}, n-1\right)$.



3) L'interval serà:

$$\overline{X}_n \pm t_{n-1,\gamma} \frac{S_n}{\sqrt{n-1}},$$

o el que és el mateix:

$$\left[\overline{X}_n - t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}}, \overline{X}_n + t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}}\right].$$

Observació 24. La t de Student només la farem servir en els casos en què tinguem mostres de mida $n < 30$; en la resta de casos podem aproximar la t de Student per una $N(0,1)$.

5.2.3. Interval de confiança per a la variància σ^2

Volem construir un interval de confiança per a σ^2 amb un nivell de confiança γ.

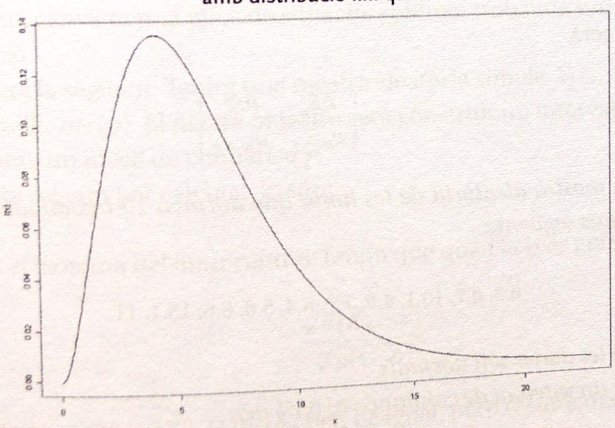
Per construir un interval de confiança per a σ^2, utilitzarem un resultat que ens diu que si $Z_1, \ldots, Z_n$ són variables aleatòries independents idènticament distribuïdes amb distribució normal estàndard, aleshores

$$Z_1^2 + \cdots + Z_n^2 \sim \chi_n^2,$$

on χ_n^2 és una distribució **khi quadrat amb *n* graus de llibertat**. La distribució χ_n^2, **khi quadrat amb *n* graus de llibertat** és una distribució:

- no simètrica,
- que només pren valors positius,
- $E(\chi_n^2) = n$ i $Var(\chi_n^2) = 2n$.

Funció de densitat d'una variable aleatòria amb distribució khi quadrat



Fent servir aquesta nova distribució, podem veure que:

Proposició 11. *Si* $X_1, X_2, \ldots, X_n$ *és una mostra aleatòria d'una distribució normal* $N(\mu, \sigma^2)$ *(això vol dir que* $X_i \sim N(\mu, \sigma^2)$*), aleshores*

$$\frac{nS_n^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Ara, doncs, els passos que seguim per a calcular un interval de confiança per a σ^2 són:

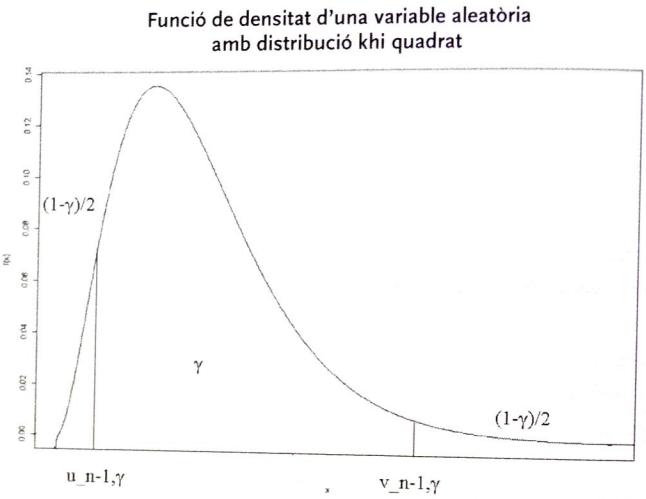
1) Utilitzant el resultat de la proposició anterior tenim

$$\frac{nS_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

2) Calculem els valors crítics $u_{n-1,\gamma}$ i $v_{n-1,\gamma}$ tals que

$$P\left(u_{n-1,\gamma} \leq \chi^2_{n-1} \leq v_{n-1,\gamma}\right) = \gamma;$$

en aquest cas, $u_{n-1,\gamma} = \texttt{qchisq}\left(\frac{1-\gamma}{2}, n-1\right)$ i $v_{n-1,\gamma} = \texttt{qchisq}\left(1-\frac{1-\gamma}{2}, n-1\right)$.



3) L'interval serà

$$\left[\frac{nS_n^2}{v_{n-1,\gamma}}, \frac{nS_n^2}{u_{n-1,\gamma}}\right].$$

Exemple 48. *Una mostra aleatòria de les hores que dormen 10 treballadors d'una empresa ens dóna els resultats següents:*

8.3, 8.7, 10.1, 4.6, 7.7, 5.4, 5.8, 6.6, 13.1, 11.

Suposem que les dades són normals.
Volem trobar un interval de confiança al 95% per:

1) *La mitjana hores que dormen, si es coneix que* $\sigma = 2$.
Com que σ és coneguda, el nostre interval de confiança serà:

$$\left[\overline{x} - u_\gamma \frac{\sigma}{\sqrt{n}}, \overline{x} + u_\gamma \frac{\sigma}{\sqrt{n}}\right].$$

Hem de calcular: $\overline{x} = 8.13$ i $u_{0.95}$=`qnorm(1-0.05/2)`=1.96; per tant,

$$\left[8.13 - 1.96\frac{2}{\sqrt{10}}, 8.13 + 1.96\frac{2}{\sqrt{10}}\right] = [6.89,\ 9.37].$$

2) *La mitjana hores que dormen, si no coneixem el valor de* σ.
Com que σ és desconeguda, el nostre interval de confiança serà:

$$\left[\overline{x} - t_{n-1,\gamma}\frac{s_n}{\sqrt{n-1}},\ \overline{x} + t_{n-1,\gamma}\frac{s_n}{\sqrt{n-1}}\right].$$

Hem de calcular: $\overline{x} = 8.13$, $s_n = 2.55$ i $t_{9,0.95}$ =`qt(1-0.05/2,9)`=2.26; per tant,

$$\left[8.13 - 2.26\frac{2.55}{\sqrt{9}},\ 8.13 + 2.26\frac{2.55}{\sqrt{9}}\right] = [6.209,\ 10.051].$$

3) *La variabilitat de les hores que dormen.*
Com que les dades són normals, l'interval de confiança que volem calcular serà:

$$\left[\frac{n s_n^2}{v_{n-1,\gamma}},\ \frac{n s_n^2}{u_{n-1,\gamma}}\right].$$

Tenim calculada $s_n = 2.55$, i ens falta calcular $u_{9,0.95}$ = `qchisq(0.05/2,9)`=2.7 i $v_{9,0.95}$ = `qchisq(1-0.05/2,9)` = 19.02; per tant,

$$\left[\frac{10 \cdot 2.55^2}{19.02},\ \frac{10 \cdot 2.55^2}{2.7}\right] = [3.42,\ 24.08].$$

5.3. Intervals de confiança per a la proporció

Fins ara hem treballat per trobar intervals de confiança per als paràmetres μ i σ d'una distribució normal. Ara canviem una mica de situació: voldrem trobar un interval de confiança per a una proporció.

La situació serà la següent: Tenim una mostra aleatòria simple $X_1, \ldots, X_n$ que ve d'una distribució Bernouilli, $Ber(p)$. El nostre objectiu serà construir un interval de confiança per a la proporció p amb un nivell de confiança γ.

Els passos que seguim per calcular-lo són:

1) Utilitzant el teorema del límit central, tenim que quan la n és gran

$$\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}}$$

es comporta com una $N(0,1)$ (no és ben bé igual, però recordem que en el cas de n gran s'hi aproxima prou bé).

2) Calculem els valors crítics u_γ tal que

$$P(-u_\gamma \le Z \le u_\gamma) = \gamma \Rightarrow P(Z < u_\gamma) = 1 - \frac{1-\gamma}{2};$$

en aquest cas, $u_\gamma = \texttt{qnorm}\left(1 - \frac{1-\gamma}{2}\right)$.

3) L'interval serà:

$$\overline{X}_n \pm u_\gamma \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}},$$

o el que és el mateix:

$$\left[\overline{X}_n - u_\gamma \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}},\ \overline{X}_n + u_\gamma \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}\right].$$

Observem que $\overline{X}_n$ serà la proporció observada a partir de la mostra. Notem també que hem utilitzat $\overline{X}_n(1-\overline{X}_n)$ en lloc de $\sigma = p(1-p)$, ja que p és el paràmetre que volem estimar.

Exemple 49. *Un radar fix situat en una carretera detecta que 56 cotxes dels 489 que han passat per aquell punt un dia determinat anaven a una velocitat superior a la permesa. Volem trobar un interval de confiança al 90% per a la proporció de cotxes que sobrepassen el límit de velocitat en aquest punt.*

Com que volem calcular un interval de confiança d'una proporció, tindrem en compte que aquest serà de la forma:

$$\left[\overline{x} - u_\gamma \sqrt{\frac{\overline{x}(1-\overline{x})}{n}},\ \overline{x} + u_\gamma \sqrt{\frac{\overline{x}(1-\overline{x})}{n}}\right],$$

on $\overline{x} = \frac{56}{489} = 0.1145$ i $u_{0.9} =$ `qnorm(1-0.1/2)`$=1.645$; per tant,

$$\left[0.1145 - 1.645\sqrt{\frac{0.1145\cdot(1-0.1145)}{489}},\ 0.1145 + 1.645\sqrt{\frac{0.1145\cdot(1-0.1145)}{489}}\right] = [0.0908,\ 0.1382].$$

I si ara volguéssim disminuir l'interval de confiança a la meitat mantenint el nivell de confiança, quina hauria de ser la nova grandària de la mostra?

Observem que la longitud de l'interval és:

$$L = 2u_\gamma \sqrt{\frac{\overline{x}(1-\overline{x})}{n}} = 2\cdot 1.645\sqrt{\frac{0.1145\cdot(1-0.1145)}{489}} = 0.0474.$$

Si ara volem que la longitud sigui la meitat, però variant només la grandària de la mostra, estem demanant que

$$2\cdot 1.645\sqrt{\frac{0.1145\cdot(1-0.1145)}{n}} = \frac{0.0474}{2}$$

i només ens cal aïllar la n de l'equació anterior. Si ho fem, obtenim que $n \approx 1954$.

5.4. Intervals de confiança per a mostres grans

Hem vist fins ara com calcular intervals de confiança per a paràmetres en el cas que la distribució sigui normal i també en el cas d'una proporció. Però ens preguntem què passa si no tenim cap d'aquests casos, sinó que les nostres dades segueixen una distribució qualsevol. Aleshores podrem trobar intervals de confiança per a l'esperança sempre que treballem amb mostres grans ($n > 30$).

Així, doncs, la situació ara és: tenim $X_1, \ldots, X_n$ una mostra aleatòria tal que $n > 30$ d'una variable X amb distribució desconeguda i $E(X) = \mu$. Aleshores podem trobar un interval de confiança per a la μ de la manera següent:

1) Sabem pel teorema del límit central que

$$Z = \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n-1}}}$$

es comporta com una $N(0,1)$.

2) Calculem els valors crítics u_γ tals que

$$P\left(-u_\gamma \leq Z \leq u_\gamma\right) = \gamma \Rightarrow P(Z < u_\gamma) = 1 - \frac{1-\gamma}{2};$$

en aquest cas, $u_\gamma = \texttt{qnorm}\left(1 - \frac{1-\gamma}{2}\right)$.

3) L'interval serà:

$$\overline{X}_n \pm u_\gamma \frac{S_n}{\sqrt{n-1}},$$

o el que és el mateix:

$$\left[\overline{X}_n - u_\gamma \frac{S_n}{\sqrt{n-1}}, \overline{X}_n + u_\gamma \frac{S_n}{\sqrt{n-1}}\right].$$

Exemple 50. *Volem estudiar el temps de vida –el temps que poden funcionar fins a espatllar-se– d'una certa classe de bombetes. Hem observat els temps de vida de 1000 bombetes i hem obtingut unes observacions* $x_1, \ldots, x_{1000}$ *tals que* $\overline{x} = 1840$ *dies i* $s^2 = 96.5$ *dies. A partir d'aquestes dades volem obtenir un interval de confiança al 90% per al temps de vida mitjà de les bombetes.*

En principi no sabem la distribució del temps de vida d'una bombeta. De totes maneres, com que la mostra és molt gran, podem fer un interval de confiança per a l'esperança del temps de vida, que serà:

$$\left[\overline{x} - u_\gamma \frac{s_n}{\sqrt{n-1}}, \overline{x} + u_\gamma \frac{s_n}{\sqrt{n-1}}\right].$$

Només necessitem calcular $u_{0.9} = \texttt{qnorm(1-0.1/2)}$=1.64; per tant, l'interval de confiança per a la mitjana serà:

$$\left[1840-1.64\left(\frac{\sqrt{96.5}}{\sqrt{999}}\right),\ 1840+1.64\left(\frac{\sqrt{96.5}}{\sqrt{999}}\right)\right]=[1839.490,\ 1840.509].$$

Fixeu-vos que com que la mostra és molt gran, l'interval de confiança és bastant petit.

5.5. Intervals de confiança per a dues mostres

Tots els intervals que hem calculat fins ara depenien d'una sola mostra. El problema que se'ns planteja ara és calcular intervals de confiança per a dues mostres. Estudiarem mostres que són observacions de la mateixa variable en condicions diferents.

Quan treballem amb dues mostres, la primera característica en què ens hem de fixar és si les dades són aparellades o no.

Definició 29. *Una* **mostra de dades aparellades** *és un conjunt d'observacions de dues variables observades en un únic conjunt d'individus, de manera que per a cada individu tenim una observació de cada una de les dues variables.*

Aquest factor és important tenir-lo en compte, ja que els càlculs que hem de fer en cada cas són diferents.

5.5.1. Intervals de confiança per a dues mostres de dades aparellades

Per calcular els intervals de confiança, no necessitarem saber res de nou. La manera de procedir és crear una nova variable que serà la diferència entre les dades de la primera mostra i les dades de la segona mostra i que s'obté restant les dues variables. Així, doncs, retornem al cas d'una sola mostra i tot el que hem fet a l'apartat anterior ho podem aplicar per a aquesta nova variable que hem creat.

Exemple 51. *Després d'un règim de 3 mesos, el nivell de glucosa de 6 malalts va presentar la variació següent:*

Abans règim (mg/dl)	130	121	100	110	110	99
Després règim (mg/dl)	105	110	101	82	80	97

Volem estudiar com ha variat el nivell de glucosa. O sigui, volem calcular un interval de confiança al 90% per a la diferència de mitjanes.

Observeu que les dades són aparellades, ja que els individus són els mateixos, de manera que si les restem obtenim una nova variable que serà *el canvi en el nivell de glucosa*:

Abans règim (mg/dl)	130	121	100	110	110	99
Després règim (mg/dl)	105	110	101	82	80	97
Canvi nivell glucosa	−25	−11	1	−28	−30	−2

Podem considerar, per tant, que tenim una única variable. Suposant que tenim normalitat per a les dades inicials, tenim també normalitat per a la nova variable. Així, podem

buscar un interval de confiança per a la mitjana d'aquesta variable –estem en el cas que hem d'obtenir un interval de confiança per a la mitjana quan la variància poblacional és desconeguda i tenim una població normal–. Per tant, l'interval serà:

$$\left[\overline{x} - t_{n-1,\gamma}\frac{s_n}{\sqrt{n-1}}, \overline{x} + t_{n-1,\gamma}\frac{S_n}{\sqrt{n-1}}\right].$$

Necessitem calcular:

$$n = 6 \qquad \overline{x} = -15.833 \quad s_n = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = 12.4554.$$

i $t_{5,0.9} = \texttt{qt(1-0.1/2,5)} = 2.015$. Substituint trobem finalment l'interval per a la mitjana:

$$\left[-15.833 - 2.015\left(\frac{12.4554}{\sqrt{5}}\right), -15.833 + 2.015\left(\frac{12.4554}{\sqrt{5}}\right)\right]$$
$$= [-27.057, -4.6089].$$

Per tant, la mitjana de la variació en el nivell de glucosa està en aquest interval amb un nivell de confiança del 90%, és a dir, hi ha hagut una reducció del nivell de glucosa d'entre 4.6089 i 27.057. Fixeu-vos que es tracta d'un interval bastant gran, ja que tenim una mostra molt petita.

5.5.2. Intervals de confiança per a dues mostres de dades no aparellades

Suposarem a partir d'ara que tenim dues mostres aleatòries $X_1, \ldots, X_n$ i $Y_1, \ldots, Y_m$, independents entre elles, que provenen d'unes distribucions $N(\mu_x, \sigma_x^2)$ i $N(\mu_y, \sigma_y^2)$, respectivament.

Intervals de confiança per a $\mu_x - \mu_y$, on σ_x^2 i σ_y^2 són conegudes

Volem construir un interval de confiança per a $\mu_x - \mu_y$ suposant que σ_x^2 i σ_y^2 són conegudes amb un nivell de confiança γ.

Els passos que seguim per a calcular-lo són:

1) Sabem que

$$\overline{X}_n \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right) \quad \text{i} \quad \overline{Y}_m \sim N\left(\mu_y, \frac{\sigma_y^2}{m}\right)$$

i que $\overline{X}_n$ i $\overline{Y}_m$ són independents; per tant,

$$\overline{X}_n - \overline{Y}_m \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right).$$

Així, doncs,

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_x - \mu_y)}{\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}} \sim N(0,1).$$

2) Calculem el valor crític u_γ tal que

$$P\left(-u_\gamma \leq Z \leq u_\gamma\right) = \gamma \Rightarrow P(Z < u_\gamma) = 1 - \frac{1-\gamma}{2},$$

en aquest cas $u_\gamma = \texttt{qnorm}\left(1 - \frac{1-\gamma}{2}\right)$.

3) L'interval serà:

$$\overline{X}_n - \overline{Y}_m \pm u_\gamma \left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}},$$

o el que és el mateix:

$$\left[\overline{X}_n - \overline{Y}_m - u_\gamma \left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}, \overline{X}_n - \overline{Y}_m + u_\gamma \left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}\right].$$

Intervals de confiança per a $\mu_x - \mu_y$, on σ_x^2 i σ_y^2 són desconegudes però iguals

Volem construir un interval de confiança per a $\mu_x - \mu_y$ suposant que $\sigma^2 = \sigma_x^2 = \sigma_y^2$ són desconegudes amb un nivell de confiança γ.

Els passos que seguim per a calcular-lo són:

1) Sabem com en el cas anterior que

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_x - \mu_y)}{\sigma\left(\frac{1}{n} + \frac{1}{m}\right)^{\frac{1}{2}}} \sim N(0,1).$$

Però no podem utilitzar aquest resultat, ja que no coneixem el valor de σ. El que fem és substituir el valor de σ per les desviacions típiques respectives S_X^2 i S_Y^2. Aleshores, si considerem

$$T_{n+m-2} = \frac{\overline{X}_n - \overline{Y}_m - (\mu_x - \mu_y)}{\left(nS_X^2 + mS_Y^2\right)^{\frac{1}{2}}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

T_{n+m-2} té una distribució t de student amb $n+m-2$ graus de llibertat.

2) Calculem els valors crítics $t_{n+m-2,\gamma}$ tals que

$$P\left(-t_{n+m-2,\gamma} \leq T_{n+m-2} \leq t_{n+m-2,\gamma}\right) = \gamma \Rightarrow P(T_{n+m-2} < t_{n+m-2,\gamma}) = 1 - \frac{1-\gamma}{2},$$

en aquest cas $t_{n+m-2,\gamma} = \texttt{qt}(1 - \frac{1-\gamma}{2}, n+m-2)$.

3) L'interval serà:

$$\overline{X}_n - \overline{Y}_m \pm t_{n+m-2,\gamma} \sqrt{\frac{(n+m)(nS_X^2 + mS_Y^2)}{nm(n+m-2)}},$$

o el que és el mateix:

$$\left[\overline{X}_n - \overline{Y}_m - t_{n+m-2,\gamma}\sqrt{\frac{(n+m)(nS_X^2+mS_Y^2)}{nm(n+m-2)}},\ \overline{X}_n - \overline{Y}_m + t_{n+m-2,\gamma}\sqrt{\frac{(n+m)(nS_X^2+mS_Y^2)}{nm(n+m-2)}}\right].$$

Exemple 52. *És conegut que les notes de l'examen d'Estadística segueixen una distribució normal i les notes de l'examen de Probabilitats també segueixen una distribució normal. Volem estudiar si hi ha diferència entre la mitjana de les notes de les dues assignatures.*

Per fer-ho, disposem de les notes de dos grups diferents d'estudiants:

Estadística	4.17	6	4.67	4.83	5	6.5	4	5.5	5.17	5.33	2.33	7.50	7
Probabilitats	5.1	5.2	4.1	4.9	2.3	4.3	3	3	5.1	4.9	4.1		

Ho farem determinant un interval de confiança al 90% per a la diferència de mitjanes. Observem en primer lloc que les dades no són aparellades, ja que corresponen a diferents estudiants. Calcularem aquest interval en dues situacions diferents.

1) *Suposem que les notes de l'examen d'Estadística tenen una variància poblacional coneguda 1.21 i les notes de l'examen de Probabilitats també tenen una variància poblacional coneguda 1.*

   Com que estem en el cas en què les desviacions són conegudes, l'interval que hem de calcular serà de la forma

$$\left[\overline{x} - \overline{y} - u_\gamma\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}},\ \overline{x} - \overline{y} + u_\gamma\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}\right].$$

   On

$$n = 13, \quad \sigma_x^2 = 1.21, \quad m = 11 \quad \text{i} \quad \sigma_y^2 = 1.$$

   Hem de calcular

$$\overline{x} = 5.23 \quad \text{i} \quad \overline{y} = 4.18,$$

   i finalment, $u_{0.9} = \texttt{qnorm(0.95)}$=1.64.
   Així, obtenim l'interval de confiança per a $\mu_x - \mu_y$

$$\left[5.23 - 4.18 - 1.64\left(\frac{1.21}{13} + \frac{1}{11}\right)^{\frac{1}{2}},\ 5.23 - 4.18 + 1.64\left(\frac{1.21}{13} + \frac{1}{11}\right)^{\frac{1}{2}}\right]$$

$$= [0.3466,\ 1.7534].$$

2) *Suposem que les notes de l'examen d'Estadística i les notes de l'examen de Probabilitats tenen variàncies desconegudes però iguals.*

   Com que estem en el cas de desviacions desconegudes però iguals, l'interval serà de la forma

$$\left[\overline{x}-\overline{y}-t_{n+m-2,\gamma}\sqrt{\frac{(n+m)(ns_x^2+ms_y^2)}{nm(n+m-2)}},\ \overline{x}-\overline{y}+t_{n+m-2,\gamma}\sqrt{\frac{(n+m)(ns_x^2+ms_y^2)}{nm(n+m-2)}}\right].$$

Ens queda calcular $s_x^2 = 1.70$, $s_y^2 = 0.92$ i $t_{22,0.9} = \texttt{qt(0.95,22)} = 1.72$.

Així, obtenim l'interval de confiança per a $\mu_x - \mu_y$:

$$\left[5.23-4.18-1.72\sqrt{\frac{24(13\cdot 1.70+11\cdot 0.92)}{13\cdot 11\cdot 22}},\ 5.23-4.18+1.72\sqrt{\frac{24(13\cdot 1.70+11\cdot 0.92)}{13\cdot 11\cdot 22}}\right]$$

$= [0.1980,\ 1.9019]$.

Observació 25. Fixeu-vos que el zero –és a dir, quan les mitjanes són iguals– no apareix en cap d'aquests intervals, així que no podem dir que la mitjana sigui la mateixa en les dues assignatures.

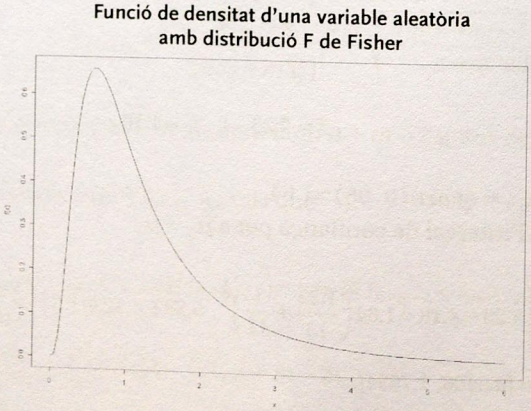
Interval de confiança per a la raó de variàncies

Finalment, volem construir un interval de confiança per a $\frac{\sigma_x^2}{\sigma_y^2}$ amb un nivell de confiança γ. Per construir aquest interval utilitzarem que si $U \sim \chi_n^2$ i $V \sim \chi_m^2$ són dues variables aleatòries independents, aleshores:

$$\frac{\frac{U}{n}}{\frac{V}{m}} \sim F_{n,m},$$

on $F_{n,m}$ té una distribució F **de Fisher amb *n* i *m* graus de llibertat**. La distribució $F_{n,m}$, **de Fisher amb *n* i *m* graus de llibertat**, és una distribució:

- no simètrica,
- que només pren valors positius.



Fent servir aquesta nova distribució, podem veure que:

Proposició 12. *Sigui* $X_1, X_2, \ldots, X_n$ *una mostra aleatòria d'una distribució normal* $N(\mu_x, \sigma^2)$ *i* $Y_1, Y_2, \ldots, Y_m$ *una mostra aleatòria d'una distribució normal* $N(\mu_y, \sigma^2)$. *Suposem a més que*

les dues mostres són independents i que no coneixem els valors poblacionals de μ_X, μ_Y *i* σ^2. *Aleshores,*

$$\frac{nS_X^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}{\sigma^2} \sim \chi^2_{n-1},$$

$$\frac{mS_Y^2}{\sigma^2} = \frac{\sum_{i=1}^{m}\left(Y_i - \overline{Y}_m\right)^2}{\sigma^2} \sim \chi^2_{m-1}.$$

I es compleix que,

$$\frac{\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2}{n-1}}{\frac{\sum_{i=1}^{m}\left(Y_i-\overline{Y}_m\right)^2}{m-1}} = \frac{(m-1)\left(\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2\right)}{(n-1)\left(\sum_{i=1}^{m}\left(Y_i-\overline{Y}_m\right)\right)} \sim F_{n-1,m-1}.$$

Ara, doncs, els passos que seguim per a calcular l'interval de confiança que volíem són:

1) Utilitzant el resultat de la proposició anterior, tenim:

$$\frac{1}{\sigma_x^2}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{nS_X^2}{\sigma_x^2} \sim \chi^2_{n-1} \quad \text{i} \quad \frac{1}{\sigma_y^2}\sum_{i=1}^{m}\left(Y_i - \overline{Y}_m\right)^2 = \frac{mS_Y^2}{\sigma_y^2} \sim \chi^2_{m-1}$$

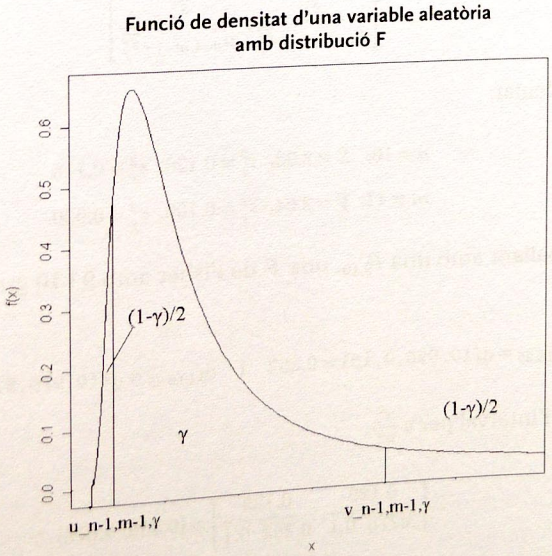
i a més també sabem que són independents; aleshores tenim que

$$F_{n-1,m-1} = \frac{\frac{nS_X^2}{(n-1)\sigma_x^2}}{\frac{mS_Y^2}{(m-1)\sigma_y^2}} = \frac{\sigma_y^2\widetilde{S}_X^2}{\sigma_x^2\widetilde{S}_Y^2},$$

on

$$\widetilde{S}_X^2 = \frac{nS_X^2}{(n-1)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \quad \text{i} \quad \widetilde{S}_Y^2 = \frac{mS_Y^2}{(m-1)} = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}_m\right)^2,$$

i $F_{n-1,m-1}$ té una distribució F **de Fisher amb $n-1$ i $m-1$ graus de llibertat**.



2) Calculem els valors crítics $u_{n-1,m-1,\gamma}$ i $v_{n-1,m-1,\gamma}$ tals que

$$P\left(u_{n-1,m-1,\gamma} \leq F_{n-1,m-1} \leq v_{n-1,m-1,\gamma}\right) = \gamma.$$

En aquest cas com que la distribució de Fisher no és simètrica, agafarem

$$v_{n-1,m-1,\gamma} = \texttt{qf}\left(1 - \frac{1-\gamma}{2}, \texttt{n-1}, \texttt{m-1}\right) \quad \text{i} \quad u_{n-1,m-1,\gamma} = \texttt{qf}\left(\frac{1-\gamma}{2}, \texttt{n-1}, \texttt{m-1}\right).$$

3) L'interval serà:

$$\left[\frac{\widetilde{S}_X^2}{v_{n-1,m-1,\gamma}\widetilde{S}_Y^2}, \frac{\widetilde{S}_X^2}{u_{n-1,m-1,\gamma}\widetilde{S}_Y^2}\right].$$

Exemple 53. *Volem estudiar si la longitud de les ales (en cm) d'una determinada espècie d'insectes depèn de l'altitud. Per fer-ho, agafem 10 individus al nivell del mar i 11 individus a una altitud de 1000 metres. Obtenim els resultats següents:*

Nivell del mar	2.5	2.8	3.1	2.4	3.3	3.6	3	2.9	2.7	3	
1000 metres	2.4	3.2	2.6	2.3	2.9	2.8	2.2	3	2.5	2.4	2.8

Suposarem normalitat.

Volem comparar, per tant, les dues mitjanes de les poblacions. És clar que no són dades aparellades.

Com que les variàncies són desconegudes, abans de calcular l'interval de confiança per a la diferència de mitjanes hem de comprovar que les variàncies són iguals.

El primer pas serà buscar un interval de confiança al 95% per a la raó de variàncies $\frac{\sigma_x^2}{\sigma_y^2}$, on σ_x^2 és la variància al nivell del mar i σ_y^2 és la variància a 1000 metres. Per tant, serà un interval de la forma

$$\left[\frac{\widetilde{s}_x^2}{v_{n-1,m-1,\gamma}\widetilde{s}_y^2}, \frac{\widetilde{s}_x^2}{u_{n-1,m-1,\gamma}\widetilde{s}_y^2}\right].$$

Hem de calcular:

$$n = 10; \ \overline{x} = 2.93, \ \widetilde{s}_x^2 = 0.129, \ s_x^2 = 0.116$$
$$m = 11, \ \overline{y} = 2.64, \ \widetilde{s}_y^2 = 0.100, \ s_y^2 = 0.090.$$

Estem treballant amb una $F_{9,10}$, una F de Fisher amb 9 i 10 graus de llibertat; trobem que

$$u_{9,10,0.95} = \texttt{qf(0.025,9,10)} = 0.252 \quad \text{i} \quad v_{9,10,0.95} = \texttt{qf(0.975,9,10)} = 3.778,$$

i això ens dóna l'interval per a $\frac{\sigma_x^2}{\sigma_y^2}$:

$$\left[\frac{0.129}{3.778 \cdot 0.1}, \frac{0.129}{0.252 \cdot 0.1}\right] = [0.338, 5.076].$$

Com que el valor 1 és a l'interval de confiança per a la raó de variàncies, podrem suposar que $\sigma_x^2 = \sigma_y^2$ –variàncies iguals.

L'interval de confiança per a $\mu_x - \mu_y$ que busquem serà de la forma

$$\left[\overline{x} - \overline{y} - t_{n+m-2,\gamma}\sqrt{\frac{(n+m)(ns_x^2+ms_y^2)}{nm(n+m-2)}},\ \overline{x} - \overline{y} + t_{n+m-2,\gamma}\sqrt{\frac{(n+m)(ns_x^2+ms_y^2)}{nm(n+m-2)}}\right].$$

Utilitzant ara una t de Student amb $10+11-2=19$ graus de llibertat, podem determinar $t_{19,0.95} = \texttt{qt(0.975,19)} = 2.093$ i, substituint, trobem l'interval de confiança per a $\mu_x - \mu_y$. Si anomenem

$$\rho = \sqrt{\frac{21\cdot(10\cdot 0.116 + 11\cdot 0.90)}{10\cdot 11\cdot 19}},$$

podem escriure:

$$\left[2.93 - 2.64 - 2.093\rho,\ 2.93 - 2.64 + 2.093\rho\right] = [-0.024,\ 0.593].$$

Tenim així que els intervals amb un nivell de confiança al 95%:

$$\frac{\sigma_x^2}{\sigma_y^2} \in [0.338,\ 5.076] \quad \text{i} \quad \mu_x - \mu_y \in [-0.024,\ 0.593].$$

Per tant, no podem deduir que hi hagi diferències entre la longitud de les ales a nivell del mar o a 1000 metres, ja que l'interval de confiança de $\mu_x - \mu_y$ conté el 0.

6. CONTRASTOS D'HIPÒTESIS

Recordem que tant en aquest capítol com en l'anterior disposem de les dades provinents d'una observació $x_1, \ldots, x_n$ d'una mostra aleatòria $X_1, \ldots, X_n$ amb una distribució desconeguda. Aquesta distribució depèn d'un paràmetre que és el que volem estimar.

En aquest capítol introduirem una nova tècnica: els contrastos d'hipòtesis. La idea dels contrastos d'hipòtesis consisteix a formular una certa hipòtesi i utilitzar les observacions que tenim, que és l'única informació de què disposem, per decidir si tenim prou evidències per a rebutjar o no la hipòtesi que hem formulat. No ens interessa tant la idea de calcular el valor del paràmetre com decidir si pot ser certa o no alguna hipòtesi sobre ell.

Per exemple, voldrem decidir si tenim prou evidències amb les dades de què disposem per a contestar si la mitjana de l'alçada dels nois d'una classe pot ser 175 cm, o si l'alçada mitjana de les noies és inferior a la dels nois.

Exemple 54. *Tirem 100 vegades una moneda i ens surten 45 cares i 55 creus. Volem saber si la moneda és perfecta. Cada tirada de la moneda, la podem considerar una realització d'una variable* $\mathrm{Ber}(p)$, *on* p *és la probabilitat de treure una cara. Tenim, per tant,* $X_1, \ldots, X_{100}$, *una mostra aleatòria simple d'una distribució* $\mathrm{Ber}(p)$, *i volem saber si* $p = \frac{1}{2}$. *És evident que si ens haguessin sortit 50 cares i 50 creus diríem que sí, però amb la relació* $45-55$ *obtinguda, què hem de fer?*

Exemple 55. *Imagineu que tenim una mostra aleatòria* $X_1, \ldots, X_n$ *d'una distribució* $N(\mu, 16)$. *Una possible hipòtesi és plantejar-nos si a partir de les observacions* $x_1, \ldots, x_n$ *podem acceptar que* $\mu = 23$. *Si, per exemple, les nostres observacions són 40, 42, 37 i 45, ja es veu clarament que la* μ *no serà 23. Malauradament, a la vida, les coses no són tan clares i ens cal una tècnica que ens permeti donar una resposta amb un cert rigor.*

Una qüestió semàntica: moltes vegades s'utilitzen els mots contrast i test indistintament.

6.1. Teoria general del contrast d'hipòtesis

Parlem en primer lloc dels nous conceptes que ens apareixeran quan vulguem fer un contrast d'hipòtesis, i de l'esquema que hem de seguir per a resoldre a un contrast.

1) En primer lloc, hem de plantejar les hipòtesis. Hem de decidir quines seran les hipòtesis nuLla i alternativa que ens plantejarem.

Definició 30. *La* **hipòtesi nuLla**, *que anomenarem* H_0, *és la hipòtesi de sortida. Representa la hipòtesi que mantindrem, llevat que les dades ens indiquin que és falsa. Si no rebutgem* H_0 *vol dir que les coses són com s'esperava que fossin i que la variabilitat observada és només una qüestió de casualitat. L'objectiu del contrast és rebutjar o no aquesta hipòtesi.*

La **hipòtesi alternativa**, *l'anomenarem H_1. Si rebutgem H_0, ho farem a favor d'aquesta hipòtesi; és la que acceptem quan la hipòtesi esperada no és certa. Indica, en aquest cas, que hi ha un canvi i que les observacions reflecteixen clarament aquest canvi.*

2) Dedidir el nivell de significació amb què treballarem.

L'objectiu del contrast és decidir si rebutgem o no la hipòtesi nul·la. Fixeu-vos que quan prenguem aquesta decisió podem estar cometent errors. Vegem-ho amb la taula següent:

	No rebutjar H_0	**Rebutjar H_0**
H_0 certa	Decisió encertada	Error de tipus I
H_0 falsa	Error de tipus II	Decisió encertada

Per tant, quan prenem una decisió, pot ser que cometem dos tipus d'errors diferents: si rebutgem H_0 quan és certa, o si no rebutgem H_0 i aquesta és falsa.

A priori, nosaltres creiem que la hipòtesi nul·la és certa. Si cometem l'error de no rebutjar H_0 quan és falsa (error del tipus II), no és un error tan greu com el de rebutjar H_0 quan resulta que és certa (error del tipus I). Ens interessarà així que l'error de tipus I sigui petit.

Definició 31. *El* **nivell de significació α** *és l'error màxim de tipus I que estem disposats a acceptar.*

Els valors que són més usuals de nivells de significació són: 0.1, 0.05 o 0.01.

Si fixem el nivell de significació $\alpha = 0.05$, estem dient que si decidim rebutjar la hipòtesi nul·la, la probabilitat que ens estiguem equivocant volem que sigui més petita que 0.05. Dit d'una altra manera, que de cada 100 vegades que decidim rebutjar la hipòtesi nul·la només ens equivocarem en 5. No sabem, però, la probabilitat d'equivocar-nos quan decidim acceptar la hipòtesi nul·la.

3) Determinar l'estadístic de contrast que utilitzarem.

Definició 32. *Un* **estadístic de contrast** *és un estadístic del qual coneixem la distribució suposant que la hipòtesi nul·la H_0 és certa.*

4) Calcular el valor observat de l'estadístic de contrast.

A partir de les observacions de què disposem, podrem calcular un valor concret de l'estadístic de contrast. La idea és veure si aquest valor que hem observat és un valor poc o molt probable d'obtenir en el cas que la hipòtesi H_0 sigui certa.

5) Càlcul del p—valor i decisió.

En el cas que la hipòtesi H_0 sigui certa, coneixem la distribució de l'estadístic de contrast, i utilitzarem aquesta informació per a determinar *com és de probable* obtenir el nostre valor observat en el cas que la hipòtesi H_0 sigui certa. El valor que utilitzarem per a determinar-ho és l'anomenat *p—valor.*

Definició 33. *El* p**–valor** *és la probabilitat d'obtenir un valor tan extrem com el que hem obtingut amb la nostra mostra* (EC_{obs}) *si la hipòtesi* H_0 *és certa.*

A partir d'aquesta definició, podem deduir que el fet que el p–valor sigui petit voldrà dir que el valor que hem observat és poc probable, i per tant que la hipòtesi nula és falsa i la decisió que prenem és rebutjar la hipòtesi nula. La probabilitat que ens equivoquem prenent aquesta decisió és exactament el p–valor. La conclusió és que si el p–valor és inferior al nivell de significació podem rebutjar H_0, ja que en el cas que ens equivoquéssim estaríem cometent un error inferior al nivell de significació.

Vegem amb un exemple com aplicar aquest esquema.

Exemple 56. *Hem preguntat a 15 alumnes d'una classe de periodisme quantes vegades han anat al cinema l'últim mes i les dades que hem obtingut són les següents:*

4 2 4 7 5 2 2 2 4 6 4 0 4 3 1

Si creiem que la mitjana és 4, hi ha prou evidències per dir que no és cert? Suposem que les dades són normals i que la desviació típica és 2.5.

1) En primer lloc, hem de plantejar les hipòtesis. En el nostre cas, volem saber si $\mu = 4$ o no, o sigui que les nostres hipòtesis seran:

$$\begin{aligned} H_0&: \quad \mu = 4 \\ H_1&: \quad \mu \neq 4. \end{aligned}$$

2) Decidir el nivell de significació amb què treballarem. Si no ens diuen cap valor, suposarem per defecte que $\alpha = 0.05$.

3) Determinar l'estadístic de contrast que utilitzarem.

En el nostre, cas podem agafar com a estadístic de contrast

$$\frac{\overline{X}_n - 4}{\frac{\sigma}{\sqrt{n}}},$$

recordem que a partir de les nostres dades $n = 15$ i $\sigma = 2.5$. Sota H_0 sabem que aquest estadístic té una distribució $N(0,1)$. Després parlarem en cada cas de quin estadístic de contrast hem d'utilitzar.

4) Calcular el valor observat de l'estadístic de contrast:

$$C_{\text{obs}} = \frac{\overline{x} - 4}{\frac{2.5}{\sqrt{15}}}.$$

Calculem $\overline{x} = 3.333$, i per tant, $E\,C_{\text{obs}} = -1.0328$.

5) Càlcul del $p-$valor i decisió:

$$p-valor = P(N(0,1) < -1.0328) + P(N(0,1) > 1.0328) = 2 \cdot (1 - P(N(0,1) < 1.0328))$$
$$= 2*(1-\texttt{pnorm}(1.0328)) = 0.3017.$$

Com que el $p-$valor és més gran que 0.05, no podem rebutjar la hipòtesi nulla.

Abans d'acabar aquest seguit de definicions i conceptes, encara ens falta afegir-ne un altre. En aquest exemple que acabem de veure hem fet un contrast del tipus

$$\begin{aligned} H_0&: \quad \mu = 4 \\ H_1&: \quad \mu \neq 4. \end{aligned}$$

Aquest tipus de contrast s'anomena **bilateral**. Podríem considerar també unes altres hipòtesis alternatives:

1) Si sospitem que la mitjana pot ser més petita que 4, hem de plantejar el contrast

$$\begin{aligned} H_0&: \quad \mu = 4 \\ H_1&: \quad \mu < 4. \end{aligned}$$

D'aquesta manera, en el cas que tinguem prou evidències per a rebutjar la hipòtesi nulla, acceptarem la hipòtesi alternativa com volíem.

2) Si sospitem que la mitjana pot ser més gran que 4, hem de plantejar el contrast

$$\begin{aligned} H_0&: \quad \mu = 4 \\ H_1&: \quad \mu > 4. \end{aligned}$$

Fixeu-vos que sempre posem a la hipòtesi alternativa el que volem demostrar.

Aquests dos contrastos s'anomenen **unilaterals**.

Observació 26. Remarquem que no és el mateix acceptar H_0 que rebutjar H_0. Quan rebutgem H_0, n'estem segurs; en canvi, quan acceptem H_0 no n'estem.

6.2. Contrastos per a una població normal

Considerem en aquest apartat que tenim una mostra aleatòria $X_1, \ldots, X_n$ que prové d'una distribució $N(\mu, \sigma^2)$. Per als contrastos que presentarem tot seguit seguirem el mateix esquema que hem presentat a l'apartat anterior.

6.2.1. Contrast sobre la mitjana quan la variància és coneguda

La idea d'aquest test és contrastar la hipòtesi nulla que serà el fet que la mitjana sigui un valor μ_0 determinat.

1) Plantegem les hipòtesis nuLla i alternativa. Aquí ens podem plantejar tres tests lleugerament diferents:

a) H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

b) H_0: $\mu = \mu_0$
H_1: $\mu > \mu_0$

c) H_0: $\mu = \mu_0$
H_1: $\mu < \mu_0$.

Independentment del test que triem fer, els tres següents punts de l'esquema seran exactament els mateixos en els tres casos.

2) Fixem el nivell de significació. Anomenarem al nivell de significació α. Recordem que si no es diu res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

Sabem que sota la hipòtesi nuLla H_0 la seva distribució és una $N(0,1)$.

4) Calculem l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

Hem de calcular, per tant, la mitjana mostral de les dades i substituir la σ i la n pels valors que ens faciliti l'enunciat.

5) Calculem el $p-$valor i prenem una decisió. En aquest últim punt sí que és important tenir en compte quines són les hipòtesis que hem plantejat en el primer punt. Així, doncs, tindrem tres casos diferents segons les hipòtesis que haguem triat.

a) En aquest cas, el $p-$valor serà:

$$p\text{—valor} = P(Z > |EC_{\text{obs}}|) + P(Z < -|EC_{\text{obs}}|) = 2 \cdot P(Z > |EC_{\text{obs}}|),$$

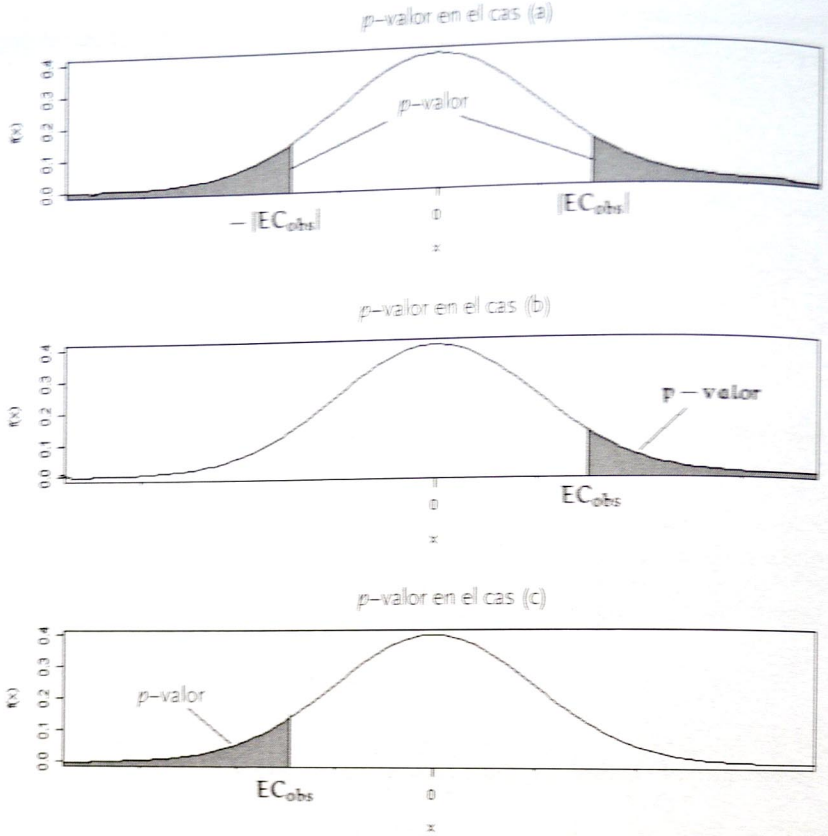
on $Z \sim N(0,1)$.

b) En canvi, en aquest cas,

$$p\text{—valor} = P(Z > EC_{\text{obs}}).$$

c) I finalment, en aquest cas,

$$p\text{—valor} = P(Z < EC_{\text{obs}}).$$



En tots tres casos, però, la decisió que prenem és:

- Si $p-$valor $\leq \alpha$ rebutgem la hipòtesi nula a favor de la hipòtesi alternativa.
- Si $p-$valor $> \alpha$ no tenim prou evidències per a rebutjar la hipòtesi nula.

Exemple 57. *Suposem que l'alçada dels alumnes d'una classe segueix una distribució normal i que sabem segur que la desviació estàndard val 4 cm. Tenim, per tant, una mostra aleatòria simple* $X_1, \ldots, X_n$ *d'una distribució* $N(\mu, 16)$. *Creiem, a més, que la* $\mu = 175$. *Les nostres observacions (en cm) són:*

180 165 168 192 195 187 181 177 175 186.

A partir d'aquestes observacions, volem saber si podem acceptar que la $\mu = 175$ *o si hi ha prou evidències per a dir que no és cert.*

1) Plantegem les hipòtesis nula i alternativa. Com que només ens interessa saber si $\mu = 175$ o no, utilitzem un contrast bilateral:

$$H_0: \quad \mu = 175$$
$$H_1: \quad \mu \neq 175.$$

2) Com que no ens diuen el contrari, suposem que $\alpha = 0.05$.

3) Recordem que l'estadístic de contrast és

$$\frac{\overline{X}_n - 175}{\frac{\sigma}{\sqrt{n}}},$$

que sota la hipòtesi nula H_0 té una distribució normal estàndard.

4) Com que $\overline{x} = 180.6$, l'estadístic de contrast observat serà:

$$EC_{\text{obs}} = \frac{180.6 - 175}{\frac{4}{\sqrt{10}}} = 4.4271.$$

5) Calculem el p–valor:

$$\begin{aligned} p\text{–valor} &= P(Z > 4.4271) + P(Z < -4.4271) = 2 \cdot P(Z > 4.4271) \\ &= \texttt{2*(1-pnorm(4.4271))=0.0000095508}, \end{aligned}$$

on $Z \sim N(0,1)$. Com que p–valor$< \alpha$, tenim prou evidències per a rebutjar H_0 a favor de la hipòtesi H_1; per tant, podem dir que la mitjana de les alçades no és de 175 cm.

6.2.2. Contrast sobre la mitjana quan la variància és desconeguda

La idea d'aquest test, igual que per a l'anterior, és contrastar la hipòtesi nula que la mitjana sigui un valor μ_0 determinat. Observeu que la principal diferència amb el contrast anterior és el canvi en l'estadístic de contrast.

1) Plantegem les hipòtesis nula i alternativa. Aquí, com abans, podem optar entre tres tests lleugerament diferents:

a) H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

b) H_0: $\mu = \mu_0$
H_1: $\mu > \mu_0$

c) H_0: $\mu = \mu_0$
H_1: $\mu < \mu_0$.

Independentment del test que triem fer, els següents punts de l'esquema seran exactament els mateixos en els tres casos, exceptuant l'últim.

2) Fixem el nivell de significació, que anomenarem α. Recordem que si no es diu res, suposarem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n-1}}},$$

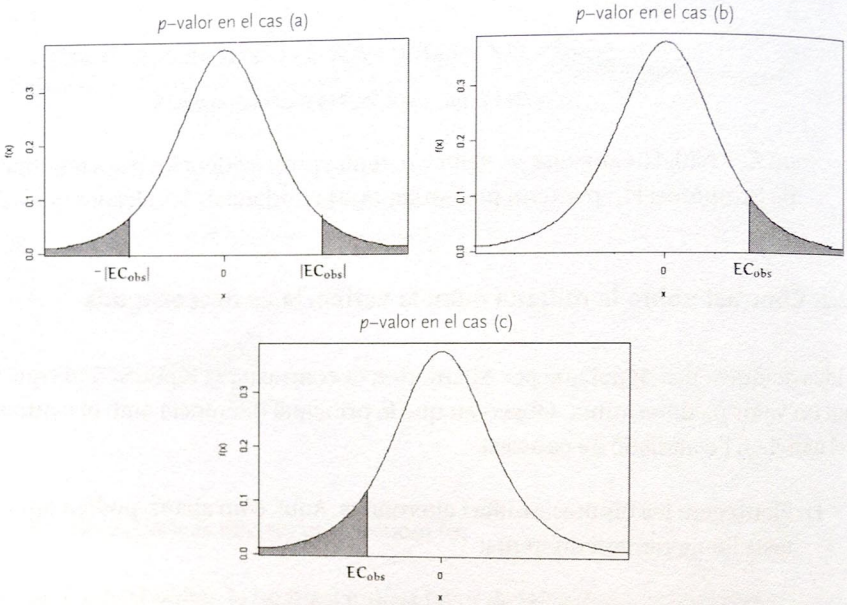
que sabem que sota la hipòtesi nula H_0 té una distribució t de Student amb $n-1$ graus de llibertat. S_n és la desviació mostral.

4) Calculem l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n-1}}}.$$

Hem de calcular, per tant, la mitjana mostral i la desviació mostral a partir de les dades, i substituir la n pel valor que ens faciliti l'enunciat.

5) Calculem el $p-$valor i prenem una decisió. En aquest últim punt sí que és important tenir en compte quines són les hipòtesis que hem plantejat. Així, doncs, tindrem tres casos diferents segons les hipòtesis que haguem triat en el primer punt.



a) En aquest cas, el $p-$valor serà:

$$p\text{–valor} = P(t_{n-1} > |EC_{\text{obs}}|) + P(t_{n-1} < -|EC_{\text{obs}}|) = 2 \cdot P(t_{n-1} > |EC_{\text{obs}}|),$$

on t_{n-1} és una t Student amb $n-1$ graus de llibertat.

b) En canvi, en aquest cas,

$$p\text{–valor} = P(t_{n-1} > EC_{\text{obs}}).$$

c) I finalment, en aquest cas,

$$p\text{–valor} = P(t_{n-1} < EC_{\text{obs}}).$$

En tots tres casos, però, la decisió que prendrem serà:

- Si $p-$valor $< \alpha$, rebutgem la hipòtesi nul·la.
- Si $p-$valor $> \alpha$, no tenim prou evidències per a rebutjar la hipòtesi nul·la.

Exemple 58. *Tornem a l'exemple anterior, amb les mateixes condicions, però suposem ara que no coneixem el valor de la variància. Recordem que les nostres observacions eren:*

180 165 168 192 195 187 181 177 175 186.

Volem saber a partir d'aquestes observacions si podem acceptar que la $\mu = 175$ o si hi ha prou evidències per a dir que no és cert.

1) Plantegem les hipòtesis nulla i alternativa.

$$\begin{aligned} H_0&: \quad \mu = 175 \\ H_1&: \quad \mu \neq 175. \end{aligned}$$

2) Com que no ens diuen el contrari, suposem que $\alpha = 0.05$.

3) Recordem que en aquest cas l'estadístic de contrast és:

$$\frac{\overline{X}_n - 175}{\frac{S_n}{\sqrt{n-1}}},$$

que sota la hipòtesi nulla H_0 té una distribució t de Student amb $n-1$ graus de llibertat.

4) Com que $\overline{x} = 180.6$ i $s = 9.2433$, l'estadístic de contrast observat serà:

$$EC_{\text{obs}} = \frac{180.6 - 175}{\frac{9.2433}{\sqrt{9}}} = 1.8175.$$

5) Calculem el $p-$valor segons la hipòtesi fixada. En aquest cas, el $p-$valor serà:

$$\begin{aligned} p-\text{valor} &= P(t_9 > 1.8175) + P(t_9 < -1.8175) = 2 \cdot P(t_9 > 1.8175) \\ &= \texttt{2*(1-pt(1.8175,9))= 0.1025}, \end{aligned}$$

on t_9 és una t de Student amb 9 graus de llibertat. Observem que com que $p-$valor $> 0.05 = \alpha$, no podem rebutjar H_0.

6.2.3. Contrast sobre la variància

La idea d'aquest test és contrastar la hipòtesi nulla que la variància sigui un valor σ_0^2 determinat.

1) Plantegem les hipòtesis nulla i alternativa. Aquí ens podem plantejar tres tests lleugerament diferents:

a) $$\begin{aligned} H_0&: \quad \sigma^2 = \sigma_0^2 \\ H_1&: \quad \sigma^2 \neq \sigma_0^2 \end{aligned}$$

b) $$\begin{aligned} H_0&: \quad \sigma^2 = \sigma_0^2 \\ H_1&: \quad \sigma^2 > \sigma_0^2 \end{aligned}$$

c) $$\begin{aligned} H_0&: \quad \sigma^2 = \sigma_0^2 \\ H_1&: \quad \sigma^2 < \sigma_0^2. \end{aligned}$$

Independentment del test que triem fer, els següents punts de l'esquema seran exactament els mateixos en els tres casos.

2) Fixem el nivell de significació. Anomenarem α el nivell de significació. Recordem que si no es diu res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{nS_n^2}{\sigma_0^2},$$

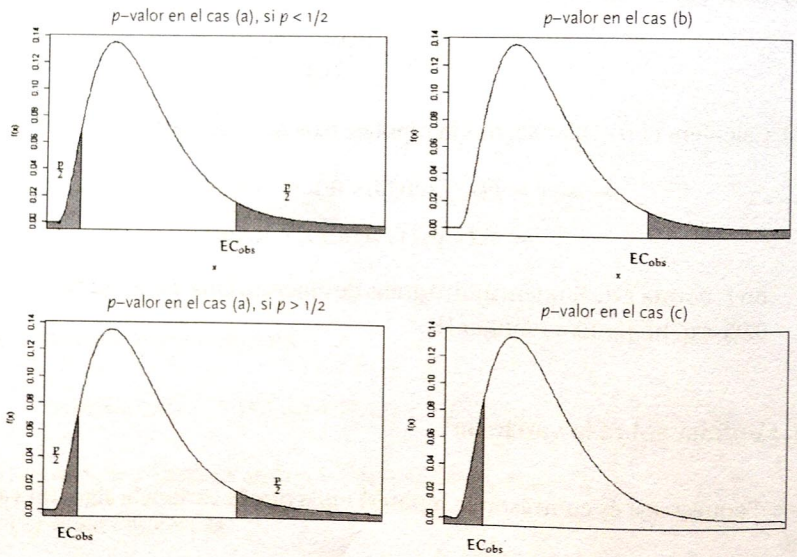
que sabem que sota la hipòtesi nula H_0 segueix una distribució χ^2_{n-1}.

4) Calculem l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Hem de calcular, per tant, la mitjana mostral de les dades, on x_i denota les dades que ens donarà l'enunciat.

5) Calculem el p–valor i prenem una decisió. En aquest últim punt sí que és important tenir en compte quines són les hipòtesis que hem plantejat. Així, doncs, tindrem tres casos diferents segons les hipòtesis que haguem triat en el primer punt.



a) En aquest cas, el p–valor serà una mica complicat de calcular, ja que la distribució

χ^2 no és simètrica:

$$\text{a)} \quad p\text{–valor} = \begin{cases} 2\cdot P(\chi^2_{n-1} > EC_{\text{obs}}) & \text{si} \quad p = P(\chi^2_{n-1} > EC_{\text{obs}}) < \frac{1}{2} \\ 2\cdot P(\chi^2_{n-1} < EC_{\text{obs}}) & \text{si} \quad p = P(\chi^2_{n-1} > EC_{\text{obs}}) > \frac{1}{2}. \end{cases}$$

b) En canvi, en aquest cas,

$$p\text{–valor} = P(\chi^2_{n-1} > EC_{\text{obs}}).$$

c) I finalment, en aquest cas,

$$p\text{—valor} = P(\chi^2_{n-1} < EC_{\text{obs}}).$$

En tots tres casos, però, la decisió que prendrem serà:

- Si p—valor $< \alpha$, rebutgem la hipòtesi nulla.
- Si p—valor $> \alpha$, no tenim prou evidències per a rebutjar la hipòtesi nulla.

Exemple 59. *Tornem a l'exemple anterior, amb les mateixes condicions. Tenim una mostra aleatòria* $X_1, \ldots, X_n$ *d'una distribució* $N(\mu, \sigma^2)$:

180 165 168 192 195 187 181 177 175 186.

A partir d'aquestes observacions, volem saber si podem acceptar que la $\sigma^2 = 36$ *o si hi ha prou evidències per a dir que no és cert.*

1) Plantegem les hipòtesis nulla i alternativa:

$$\begin{aligned} H_0&: \quad \sigma^2 = 36 \\ H_1&: \quad \sigma^2 \neq 36. \end{aligned}$$

2) Com que no ens diuen el contrari, suposem que $\alpha = 0.05$.

3) Recordem que en aquest cas l'estadístic de contrast és:

$$EC = \frac{1}{36}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

que sota la hipòtesi nulla H_0 té una distribució χ^2_{n-1}.

4) Com que $\overline{x} = 180.6$, l'estadístic de contrast observat serà:

$$EC_{\text{obs}} = \frac{1}{36}\sum_{i=1}^{10}(x_i - 180.6)^2 = 23.733.$$

5) Calculem el p—valor segons les hipòtesis fixades. Com que $P(\chi^2_9 > 23.733) = 0.0047 < \frac{1}{2}$,

$$\begin{aligned} p\text{—valor} &= 2 \cdot P(\chi^2_9 > 23.733) = \\ &= \texttt{2*(1-pchisq(23.733,9))} = \texttt{0.009488}. \end{aligned}$$

Com que p—valor $= 0.0094 < 0.05 = \alpha$, tenim prou evidències per a rebutjar H_0, i per tant podem dir que $\sigma^2 \neq 36$.

6.3. Contrast sobre una proporció

De la mateixa manera que quan hem estudiat intervals de confiança hem construït un interval per a la proporció, també podem fer constrastos per a la proporció. Així, doncs, considerem una mostra aleatòria $X_1, \ldots, X_n$ que prové d'una distribució $Ber(p)$. L'esquema que seguirem per a construir aquest constrast és el mateix que en els casos anteriors:

1) Plantegem les hipòtesis nuLla i alternativa. Aquí, com abans, podem optar per tres tests lleugerament diferents:

   a) H_0: $p = p_0$
   H_1: $p \neq p_0$

   b) H_0: $p = p_0$
   H_1: $p > p_0$

   c) H_0: $p = p_0$
   H_1: $p < p_0$

   Independentment del test que triem fer, els següents punts de l'esquema seran exactament els mateixos en els tres casos, exceptuant l'últim.

2) Fixem el nivell de significació, que anomenarem α. Recordem que si no es diu res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\sqrt{n}(\overline{X}_n - p_0)}{\sqrt{p_0(1-p_0)}},$$

   que sabem que sota la hipòtesi nuLla H_0, quan el valor n és gran, té una distribució $N(0,1)$. Observem que la distribució d'aquest estadístic no és ben bé una normal, però quan n és gran, l'aproximació que obtenim és prou bona.

4) Calculem l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{\sqrt{n}(\overline{x} - p_0)}{\sqrt{p_0(1-p_0)}}.$$

   Hem de calcular només la mitjana mostral, que serà de fet la proporció de la mostra, i obtenir el valor de n que ens faciliti l'enunciat.

5) Calculem el $p-$valor i prenem una decisió. En aquest últim punt sí que és important tenir en compte quines són les hipòtesis que hem plantejat. Així, doncs, tindrem tres casos diferents segons les hipòtesis que haguem triat en el primer punt.

   a) En aquest cas, el $p-$valor serà:

$$p\text{–valor} = P(N(0,1) > |EC_{\text{obs}}|) + P(N(0,1) < -|EC_{\text{obs}}|) = 2 \cdot P(N(0,1) > |EC_{\text{obs}}|).$$

b) En canvi, en aquest cas,

$$p\text{−valor} = P(N(0,1) > EC_{\text{obs}}).$$

c) I finalment, en aquest cas,

$$p\text{−valor} = P(N(0,1) < EC_{\text{obs}}).$$

En tots tres casos, però, la decisió que prendrem serà:

- Si p−valor $< \alpha$, rebutgem la hipòtesi nuⱧa.
- Si p−valor $> \alpha$, no tenim prou evidències per a rebutjar la hipòtesi nuⱧa.

Exemple 60. *Volem saber la proporció de joves d'entre 15 i 29 anys que llegeixen algun diari cada dia. Sospitem que hi ha més d'un 25% de joves que llegeixen algun diari cada dia. Agafem 100 joves a l'atzar i en trobem 20 que sí que en llegeixen algun cada dia. Si fixem un nivell de significació* $\alpha = 0.02$*, què podem dir de la nostra hipòtesi?*

Podem pensar que cada jove és una variable aleatòria de Bernoulli $Ber(p)$, on p és la probabilitat que llegeixi algun diari cada dia.

1) Establim les hipòtesis nuⱧa i alternativa. Volem fer el contrast:

$$\mathrm{H}_0\colon\ p = 0.25 \quad \text{contra } \mathrm{H}_1\colon\ p > 0.25.$$

2) Determinem l'estadístic de contrast:

$$EC = \frac{\sqrt{100}(\overline{X}_n - 0.25)}{\sqrt{0.25 \cdot (1-0.25)}},$$

que es comporta aproximadament com una $N(0,1)$.

3) Calculem l'estadístic de contrast observat. Com que $\overline{x} = \frac{20}{100} = 0.2$, ens queda que

$$EC_{\text{obs}} = \frac{\sqrt{100}(0.2-0.25)}{\sqrt{0.25 \cdot 0.75}} = -1.1547.$$

4) Calculem el p−valor i prenem una decisió:

$$p\text{−valor} = P(N(0,1) > -1.1547) = \texttt{1-pnorm(-1.1547)} = 0.87589,$$

i, per tant, no podem rebutjar la hipòtesi nuⱧa. No hi ha prou evidències per a suposar que la proporció és superior al 25%.

6.4. Contrast sobre la mitjana per a mostres grans

El problema que se'ns planteja ara és com hem de fer un contrast quan tenim mostres grans, $n > 30$, però no coneixem la distribució de la nostra mostra. En aquest cas, podem fer contrastos per a l'esperança de qualsevol distribució utilitzant el teorema del límit central.

Considerem una mostra aleatòria $X_1, \ldots, X_n$ $(n > 30)$ d'una variable X amb distribució desconeguda i $E(X) = \mu$ (no cal suposar normalitat).

L'esquema que plantejarem és molt similar a tots els anteriors.

1) Plantegem les hipòtesis nuLla i alternativa. Aquí, com abans, podem optar per tres tests lleugerament diferents:

a) H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

b) H_0: $\mu = \mu_0$
H_1: $\mu > \mu_0$

c) H_0: $\mu = \mu_0$
H_1: $\mu < \mu_0$.

Independentment del test que triem fer, els següents punts de l'esquema seran exactament els mateixos en els tres casos, exceptuant l'últim.

2) Fixem el nivell de significació. Anomenarem α el nivell de significació. Recordem que si no es diu res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n-1}}}.$$

Sabem, utilitzant el teorema del límit central, que sota la hipòtesi nuLla H_0 la seva distribució es comporta com una $N(0,1)$. S_n és la desviació mostral.

4) Calculem l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n-1}}}.$$

Hem de calcular, per tant, la mitjana mostral i la desviació mostral a partir de les dades, i substituir la n pel valor que ens faciliti l'enunciat.

5) Calculem el $p-$valor i prenem una decisió. En aquest últim punt sí que és important tenir en compte quines són les hipòtesis que hem plantejat. Així, doncs, tindrem tres casos diferents segons les hipòtesis que haguem triat en el primer punt.

a) En aquest cas, el $p-$valor serà:

$$p\text{–valor} = P(N(0,1) > |EC_{\text{obs}}|) + P(N(0,1) < -|EC_{\text{obs}}|) = 2 \cdot P(N(0,1) > |EC_{\text{obs}}|).$$

b) En canvi, en aquest cas,

$$p\text{–valor} = P(N(0,1) > EC_{\text{obs}}).$$

c) I finalment, en aquest cas,

$$p\text{–valor} = P(N(0,1) < EC_{\text{obs}}).$$

En tots tres casos, però, la decisió que prendrem serà:

- Si p–valor $< \alpha$, rebutgem la hipòtesi nulla.
- Si p–valor $> \alpha$, no tenim prou evidències per a rebutjar la hipòtesi nulla.

Exemple 61. *Volem estudiar el temps de vida –el temps que poden funcionar fins a espatllar-se– d'una certa classe de bombetes. Hem observat els temps de vida de 1000 bombetes i hem obtingut unes observacions* $x_1, \ldots, x_{1000}$ *tals que* $\bar{x} = 1840$ *dies i* $s^2 = 96.5$. *Sospitem que podem afirmar que la durada mitjana és superior a 1850 amb un nivell de significació de 0.05.*

En principi, no sabem si la distribució del temps de vida d'una bombeta segueix una distribució normal, però com que la mostra és molt gran podem fer un contrast d'hipòtesis. Plantegem el contrast

1) Plantegem les hipòtesis nulla i alternativa:

$$H_0\colon \mu = 1850 \quad \text{contra } H_1\colon \mu > 1850.$$

2) Fixem el nivell de significació: $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\overline{X}_n - \mu_0}{\frac{S_n}{\sqrt{n-1}}}.$$

A partir del teorema del límit central sabem que sota la hipòtesi nulla H_0 la seva distribució es comporta com una $N(0,1)$.

4) Calculem el valor de l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{1840 - 1850}{\frac{\sqrt{96.5}}{\sqrt{999}}} = -32.175.$$

5) Calculem el p–valor i prenem una decisió. El p–valor és:

$$p\text{–valor} = P(N(0,1) > -32.175) = 1\text{-pnorm}(-32.175) = 1.$$

I no podem rebutjar la hipòtesi nulla, és a dir, no tenim prou evidències per a dir que la mitjana no és 1850.

6.5. Contrastos per a dues mostres normals

Ara considerarem dues mostres, i a més suposarem que les dades que tenim no són aparellades. Altrament, si ens trobéssim amb dades aparellades, ho tractaríem com hem fet en el cas dels intervals de confiança, considerant la diferència de les dues mostres i treballant amb una única mostra.

Durant aquest apartat estarem en la situació següent: considerem dues mostres aleatòries independents entre elles: $X_1, \ldots, X_n$ i $Y_1, \ldots, Y_m$ que provenen respectivament de distribucions $N(\mu_x, \sigma_x^2)$ i $N(\mu_y, \sigma_y^2)$.

6.5.1. Contrast sobre la diferència de mitjanes $\mu_x - \mu_y$, cas de variàncies σ_x^2 i σ_y^2 conegudes

Utilitzarem aquest contrast per a veure si les mitjanes de les dues mostres són iguals, en el cas que les variàncies siguin conegudes.

1) Plantegem les hipòtesis nulla i alternativa. Aquí ens podem plantejar tres tests lleugerament diferents:

   a) Cas bilateral:

$$H_0: \mu_x - \mu_y = 0 \quad \text{contra } H_1: \mu_x - \mu_y \neq 0.$$

   b) Cas unilateral:

$$H_0: \mu_x - \mu_y = 0 \quad \text{contra } H_1: \mu_x - \mu_y > 0.$$

   c) Cas unilateral:

$$H_0: \mu_x - \mu_y = 0 \quad \text{contra } H_1: \mu_x - \mu_y < 0.$$

   Observeu que $\mu_x - \mu_y = 0$ és equivalent a escriure $\mu_x = \mu_y$.

2) Fixem el nivell de significació. Recordem que si no es diu res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\overline{X}_n - \overline{Y}_m}{\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}},$$

   que segueix una distribució normal estàndard sota la hipòtesi nulla (recordeu-ho de quan hem estudiat els intervals de confiança).

4) Calculem l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{\overline{x} - \overline{y}}{\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}}.$$

5) Calculem el $p-$valor i prenem una decisió. En aquest últim punt sí que és important tenir en compte quines són les hipòtesis que hem plantejat. Així, doncs, tindrem tres casos diferents segons les hipòtesis que haguem triat en el primer punt:

a) En aquest cas, el $p-$valor serà:

$$p\text{–valor} = P(N(0,1) > |EC_{\text{obs}}|) + P(N(0,1) < -|EC_{\text{obs}}|) = 2 \cdot P(N(0,1) > |EC_{\text{obs}}|).$$

b) En canvi, en aquest cas,

$$p\text{–valor} = P(N(0,1) > EC_{\text{obs}}).$$

c) I finalment, en aquest cas,

$$p\text{–valor} = P(N(0,1) < EC_{\text{obs}}).$$

En tots tres casos, però, la decisió que prendrem serà:

- Si $p-$valor $< \alpha$, rebutgem la hipòtesi nuⱧa.
- Si $p-$valor $> \alpha$, no tenim prou evidències per a rebutjar la hipòtesi nuⱧa.

Exemple 62. *Les notes de l'examen d'Estadística segueixen una distribució normal amb variància 1.21 i les notes de l'examen de Probabilitats també segueixen una distribució normal, però amb variància 1. Creiem que no hi ha diferència entre la mitjana de les notes. Per comprovar-ho, farem un contrast d'hipòtesis amb un nivell de significació de 0.1 per a la diferència de mitjanes.*

Per fer-ho, disposem de les notes de dos grups diferents d'estudiants:

Estadística	4.17	6	4.67	4.83	5	6.5	4	5.5	5.17	5.33	2.33	7.50	7
Probabilitats	5.1	5.2	4.1	4.9	2.3	4.3	3	3	5.1	4.9	4.1		

1) Plantegem el contrast

$$\mathrm{H}_0\colon \mu_x - \mu_y = 0 \quad \text{contra } \mathrm{H}_1\colon \mu_x - \mu_y \neq 0.$$

2) Fixem el nivell de significació. L'enunciat ens diu que $\alpha = 0.1$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\overline{X}_n - \overline{Y}_m}{\left(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)^{\frac{1}{2}}},$$

que segueix una distribució $N(0,1)$.

4) Com que tenim que

$$n = 13, \quad \sigma_x^2 = 1.21, \quad m = 11 \quad \text{i} \quad \sigma_y^2 = 1$$

i podem calcular

$$\overline{x} = 4.86 \quad \text{i} \quad \overline{y} = 4.18,$$

el valor de l'estadístic de contrast observat pren el valor

$$EC_{\text{obs}} = \frac{4.86 - 4.18}{\left(\frac{1.21}{13} + \frac{1}{11}\right)^{\frac{1}{2}}} = 1.5853.$$

5) Calculem el $p-$valor i prenem una decisió:

$$\begin{aligned} p\text{—valor} &= 2 \cdot P(N(0,1) > |EC_{\text{obs}}|) = 2 \cdot P(N(0,1) > 1.5853) \\ &= \texttt{2*(1-pnorm(1.5853))} = 0.1128. \end{aligned}$$

Per tant, com que $p-$valor $> \alpha$, no hi ha prou evidències per a rebutjar la hipòtesi nul·la, i no podem dir que les mitjanes de les notes siguin diferents.

Observem que hem utilitzat el mateix exemple que per a calcular intervals de confiança i hem arribat a la mateixa conclusió.

6.5.2. Contrast sobre la diferència de mitjanes $\mu_x - \mu_y$, cas de variàncies σ_x^2 i σ_y^2 desconegudes però iguals

Igual que en el test anterior, volem fer un contrast per a la diferència de mitjanes, però en aquest cas no coneixem el valor de les variàncies respectives tot i que sabem que són iguals.

Diem ara que $\sigma^2 = \sigma_y^2 = \sigma_x^2$.

1) Plantegem les hipòtesis nul·la i alternativa. Aquí ens podem plantejar tres tests lleugerament diferents, que seran de fet els mateixos que en el cas anterior:

a) Cas bilateral:

$$\text{H}_0\colon \mu_x - \mu_y = 0 \quad \text{contra H}_1\colon \mu_x - \mu_y \neq 0.$$

b) Cas unilateral:

$$\text{H}_0\colon \mu_x - \mu_y = 0 \quad \text{contra H}_1\colon \mu_x - \mu_y > 0.$$

c) Cas unilateral:

$$\text{H}_0\colon \mu_x - \mu_y = 0 \quad \text{contra H}_1\colon \mu_x - \mu_y < 0.$$

Observeu que $\mu_x - \mu_y = 0$ és equivalent a $\mu_x = \mu_y$.

2) Fixem el nivell de significació. Recordem que si no es diu res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast. En aquest cas, no podem utilitzar el mateix estadístic de contrast que hem fet servir en el cas anterior perquè no coneixem el valor de la variància, i, per tant, haurem d'utilitzar les variàncies mostrals:

$$EC = \frac{\overline{X}_n - \overline{Y}_m}{\left(nS_X^2 + mS_Y^2\right)^{\frac{1}{2}}} \sqrt{\frac{nm(n+m-2)}{n+m}}.$$

Sota la hipòtesi nul·la H_0 té una distribució t de Student amb $n+m-2$ graus de llibertat (recordeu que això ja ens apareixia quan fèiem intervals de confiança).

4) Calculem l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{\overline{x} - \overline{y}}{\left(ns_x^2 + ms_y^2\right)^{\frac{1}{2}}} \sqrt{\frac{nm(n+m-2)}{n+m}}.$$

Per això necessitem calcular les mitjanes $\overline{x}$ i $\overline{y}$ i les variàncies mostrals s_x^2 i s_y^2. Els valors n i m ens els donarà l'enunciat.

5) Calculem el p–valor i prenem una decisió. En aquest últim punt sí que és important tenir en compte quines són les hipòtesis que hem plantejat. Així, doncs, tindrem tres casos diferents segons les hipòtesis que haguem triat en el primer punt.

a) Cas bilateral:

$$H_0\colon \mu_x - \mu_y = 0 \quad \text{contra } H_1\colon \mu_x - \mu_y \neq 0.$$

El p–valor és:

$$p\text{–valor} = P(t_{n+m-2} > |EC_{\text{obs}}|) + P(t_{n+m-2} < -|EC_{\text{obs}}|) = 2 \cdot P(t_{n+m-2} > |EC_{\text{obs}}|).$$

b) Cas unilateral:

$$H_0\colon \mu_x - \mu_y = 0 \quad \text{contra } H_1\colon \mu_x - \mu_y > 0.$$

El p–valor és:

$$p\text{–valor} = P(t_{n+m-2} > EC_{\text{obs}}).$$

c) Cas unilateral:

$$H_0\colon \mu_x - \mu_y = 0 \quad \text{contra } H_1\colon \mu_x - \mu_y < 0.$$

El p–valor és:

$$p\text{–valor} = P(t_{n+m-2} < EC_{\text{obs}}).$$

En tots tres casos, però, la decisió que prendrem serà:

- Si p–valor $< \alpha$, rebutgem la hipòtesi nul·la.
- Si p–valor $> \alpha$, no tenim prou evidències per a rebutjar la hipòtesi nul·la.

6.5.3. Contrast sobre la raó de variàncies

Ens plantegem ara si les variàncies de les dues mostres són iguals.

En aquest cas, l'esquema del contrast sobre la raó de variàncies és el següent:

1) Plantegem les hipòtesis nula i alternativa. Aquí ens podem plantejar tres tests lleugerament diferents:

a) Cas bilateral:

$$\mathrm{H}_0\colon \frac{\sigma_x^2}{\sigma_y^2} = 1 \quad \text{contra } \mathrm{H}_1\colon \frac{\sigma_x^2}{\sigma_y^2} \neq 1.$$

b) Cas unilateral:

$$\mathrm{H}_0\colon \frac{\sigma_x^2}{\sigma_y^2} = 1 \quad \text{contra } \mathrm{H}_1\colon \frac{\sigma_x^2}{\sigma_y^2} > 1.$$

c) Cas unilateral:

$$\mathrm{H}_0\colon \frac{\sigma_x^2}{\sigma_y^2} = 1 \quad \text{contra } \mathrm{H}_1\colon \frac{\sigma_x^2}{\sigma_y^2} < 1.$$

Observem que $\frac{\sigma_x^2}{\sigma_y^2} = 1$ és equivalent a escriure que $\sigma_x^2 = \sigma_y^2$.

2) Fixem el nivell de significació. Recordem que si no es diu res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast. Com ja vam veure quan estudiàvem intervals de confiança, tenim que

$$\frac{1}{\sigma_x^2}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{nS_X^2}{\sigma_x^2} \quad \text{i} \quad \frac{1}{\sigma_y^2}\sum_{i=1}^{m}\left(Y_i - \overline{Y}_m\right)^2 = \frac{mS_Y^2}{\sigma_y^2}$$

i segueixen distribucions χ^2_{n-1} i χ^2_{m-1}, respectivament. A més a més, són independents, de manera que

$$\frac{\frac{nS_X^2}{(n-1)\sigma_x^2}}{\frac{mS_Y^2}{(m-1)\sigma_y^2}} = \frac{\sigma_y^2 \widetilde{S}_X^2}{\sigma_x^2 \widetilde{S}_Y^2},$$

i té una distribució F de Fisher amb $n-1$ i $m-1$ graus de llibertat, és a dir, una $F_{n-1,m-1}$. Aleshores, podem agafar com a estadístic de contrast

$$EC = \frac{\widetilde{S}_X^2}{\widetilde{S}_Y^2},$$

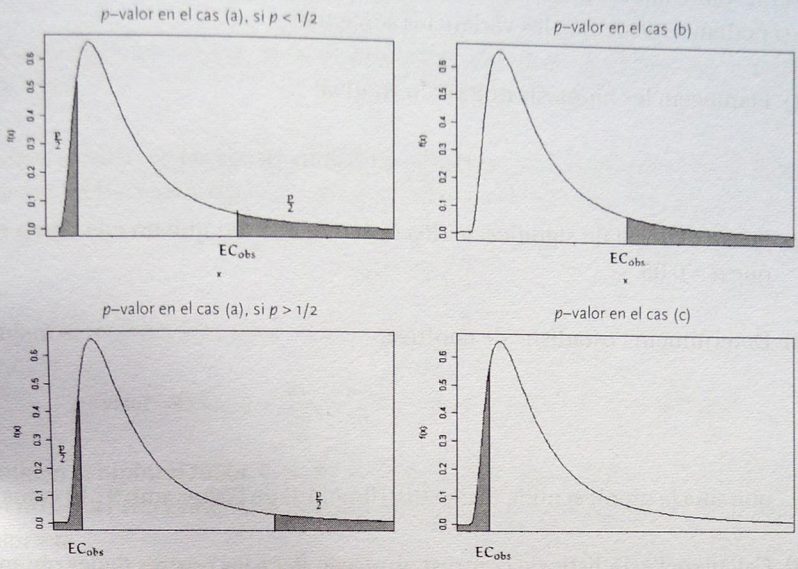
que sota la hipòtesi nula H_0 segueix una distribució F de Fisher amb $n-1$ i $m-1$ graus de llibertat.

4) Calculem l'estadístic de contrast observat:

$$EC_{\text{obs}} = \frac{\widetilde{s}_x^2}{\widetilde{s}_y^2}.$$

Per tant, necessitem calcular les variàncies mostrals corregides.

5) Calculem el p–valor i prenem una decisió. Com que la distribució F no és simètrica, és una mica més difícil donar el p–valor en el cas bilateral. De totes maneres, finalment s'obté:



a) Cas bilateral. Per a la hipòtesi alternativa $H_1: \frac{\sigma_x^2}{\sigma_y^2} \neq 1$,

- Quan $p = P(F_{n-1,m-1} > EC_{\text{obs}}) < \frac{1}{2}$, el p–valor es calcula com

$$p\text{–valor} = 2 \cdot P(F_{n-1,m-1} > EC_{\text{obs}}).$$

- Quan $p = P(F_{n-1,m-1} > EC_{\text{obs}}) > \frac{1}{2}$, el p–valor es calcula com

$$p\text{–valor} = 2 \cdot P(F_{n-1,m-1} < EC_{\text{obs}}).$$

b) Cas unilateral. Per a la hipòtesi alternativa $H_1: \frac{\sigma_x^2}{\sigma_y^2} > 1$,

$$p\text{–valor} = P(F_{n-1,m-1} > EC_{\text{obs}}).$$

c) Cas unilateral. Per a la hipòtesi alternativa $H_1: \frac{\sigma_x^2}{\sigma_y^2} < 1$,

$$p\text{–valor} = P(F_{n-1,m-1} < EC_{\text{obs}}).$$

Exemple 63. Estem estudiant si la longitud de les ales d'una determinada espècie d'insectes depèn de l'altitud. Per fer-ho, agafem 10 individus al nivell del mar i 11 individus a una altitud de 1000 metres. Obtenim els resultats següents:

Nivell del mar	2.5	2.8	3.1	2.4	3.3	3.6	3	2.9	2.7	3	
1000 metres	2.4	3.2	2.6	2.3	2.9	2.8	2.2	3	2.5	2.4	2.8

Suposem que tenim normalitat. Volem comparar les mitjanes de les dues poblacions. Observeu en primer lloc que les dades no són aparellades.

Per resoldre aquest problema hem de fer dos contrastos. Farem primer el contrast per veure si podem suposar que les variàncies són iguals.

1) Plantegem les hipòtesis nuLla i alternativa:

$$\mathrm{H}_0\colon \frac{\sigma_x^2}{\sigma_y^2} = 1 \quad \text{contra } \mathrm{H}_1\colon \frac{\sigma_x^2}{\sigma_y^2} \neq 1.$$

2) Fixem el nivell de significació. Recordem que, com que no ens diuen res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\widetilde{S}_X^2}{\widetilde{S}_Y^2},$$

que sota la hipòtesi nuLla té una distribució F de Fisher amb 9 i 10 graus de llibertat.

4) Calculem l'estadístic de contrast observat. Per a les nostres dades obtenim:

$$n = 10, \ \ \overline{x} = 2.93, \ \ \widetilde{s}_x^2 = 0.129,$$
$$m = 11, \ \ \overline{y} = 2.64, \ \ \widetilde{s}_y^2 = 0.100.$$

Per tant,

$$EC_{\text{obs}} = 1.28.$$

5) Calculem el p–valor i prenem una decisió.

Com que $P(F_{9,10} > 1.28) = \texttt{1-pf(1.28,9,10)} = 0.3512531 < \frac{1}{2}$, el p–valor es calcula com

$$p\text{–valor} = 2P(F_{9,10} > 1.28) = \texttt{2*(1-pf(1.28,9,10))} = 0.7025.$$

Per tant, com que p–valor$> \alpha$, no rebutgem la hipòtesi nuLla, és a dir, acceptem que les variàncies són iguals.

Ara volem comparar les mitjanes dels dos grups. Estem, per tant, en un cas de comparació de mitjanes amb variàncies iguals (ho acabem de comprovar), però desconegudes.

1) Plantegem les hipòtesis nuLla i alternativa:

$$\mathrm{H}_0\colon \mu_x - \mu_y = 0 \quad \text{contra } \mathrm{H}_1\colon \mu_x - \mu_y \neq 0.$$

2) Fixem el nivell de significació. Recordem que, com que no ens diuen res, suposem que $\alpha = 0.05$.

3) Determinem l'estadístic de contrast:

$$EC = \frac{\overline{X}_n - \overline{Y}_m}{\left(nS_X^2 + mS_Y^2\right)^{\frac{1}{2}}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

que sota la hipòtesi nul·la H_0 té una distribució t de Student amb $n+m-2=19$ graus de llibertat.

4) Calculem a partir de les nostres dades el valor que pren l'estadístic de contrast. Tenim les mitjanes i les variàncies mostrals calculades del contrast anterior.

$$EC_{\text{obs}} = 1.92778.$$

5) Calculem el $p-$valor i prenem una decisió:

$$p-\text{valor} = 2P(t_{19} > 1.92778) = \texttt{2*(1-pt(1.92778,19))} = 0.06896557.$$

No rebutgem la hipòtesi nul·la, és a dir, finalment decidim que amb el nivell de significació escollit no hi ha prou evidències que ens diguin que les dues poblacions són diferents, ja que $p-$valor$> \alpha$.

7. EL MODEL DE REGRESSIÓ

En aquest últim tema tractarem el **model de regressió lineal simple**. L'objectiu és donar un model per intentar explicar la relació entre dues variables: una **variable resposta** Y que depèn d'una variable X, que s'anomena **variable predictiva o explicativa**. Aquest model ens servirà també per a fer prediccions.

Observació 27. Hi ha alguns llibres que la variable resposta l'anomenen *variable dependent* i la variable predictora, *variable independent*. Intentarem evitar aquestes expressions, ja que normalment el paper de les dues variables es pot intercanviar.

El cas que ens interessa és quan la variable Y no es pot posar com una funció determinista de la variable X, sinó que tenim, a més, una part aleatòria. Més precisament, estudiarem el cas en què la relació entre Y i X és lineal, és a dir, quan tenim una relació del tipus

$$Y = a + bX + \xi,$$

on ξ indica la part aleatòria.

Podem trobar molts exemples d'aquest tipus de model. Es pot estudiar la relació:

1) entre el pes (Y) i l'alçada dels homes (X),
2) entre l'alçada dels fills (Y) i l'alçada dels pares (X) (aquest és un dels primers exemples estudiats per Galton),
3) entre la velocitat d'una reacció química (Y) i la temperatura (X).

Precisament el nom de *regressió* ve dels treballs de Galton. Galton va demostrar que els fills de pares alts eren alts, però menys que els seus pares, i que els fills de pares baixos eren baixos, però menys també que els seus pares, és a dir, que en l'alçada dels fills es produïa una regressió cap a la mitjana poblacional. Així, Galton va parlar del concepte «regressió cap a la mediocritat».

L'objectiu que ens plantegem, per tant, és, a partir de dues variables X –predictiva– i Y –resposta–, estudiar la relació lineal que hi ha entre aquestes dues variables. Per a aquest objectiu, les dades de què disposem són un conjunt d'observacions bivariants (x_i, y_i), $i = 1, \ldots, n$ corresponents a les variables X i Y.

Exemple 64. En aquest tema utilitzarem com a exemple un dels més típics i tòpics dels cursos de regressió lineal. En una població d'homes, considerarem com a variable predictora X l'alçada i com a variable resposta Y el pes.

Disposem d'una mostra de les alçades i els pesos de 10 homes. Les dades que tenim són següents:

alçada	180	170	165	190	170	181	175	174	167	160	177	180	172	169	174
pes	90	73	67	95	83	80	73	80	62	60	85	86	81	77	91

7.1. Càlcul de la recta de regressió

Abans de començar el càlcul de la recta de regressió, el primer que hem de fer és comprovar si té sentit intentar determinar una relació lineal. Ens en podem fer una idea utilitzant els diagrames de dispersió o calculant el coeficient de correlació.

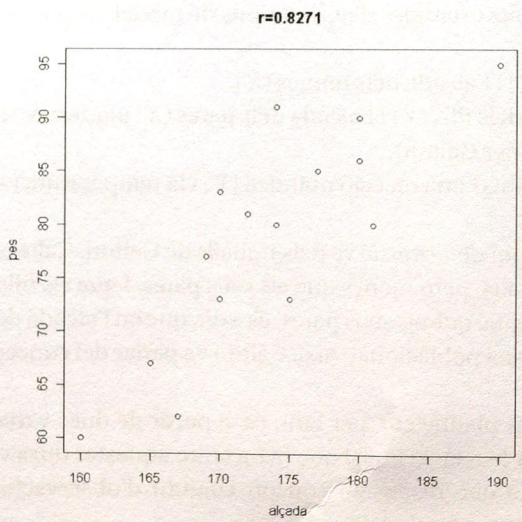
Podem calcular el coeficient de correlació a partir de les dades observades utilitzant la fórmula següent:

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

on $s_{x,y}$ és la covariància mostral i s_x i s_y les desviacions mostrals.

El coeficient de correlació $r_{x,y}$ sempre pren valors entre $[-1, 1]$ i no depèn de les unitats de mesura. Quan pren valors propers als extrems, -1 i 1, ens indicarà que entre les nostres dades x_i i y_i hi ha una relació lineal; en canvi, quan obtinguem un valor proper a 0 ens indicarà que no hi ha cap relació lineal. Mireu els exemples següents:

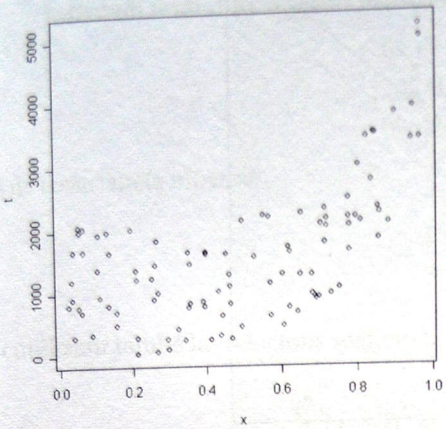
Exemple 65. Vegem què passa amb les dades del nostre exemple:

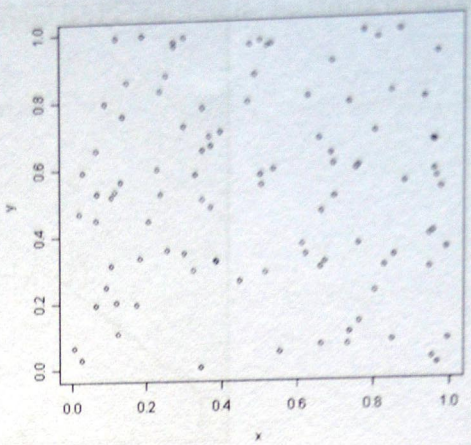


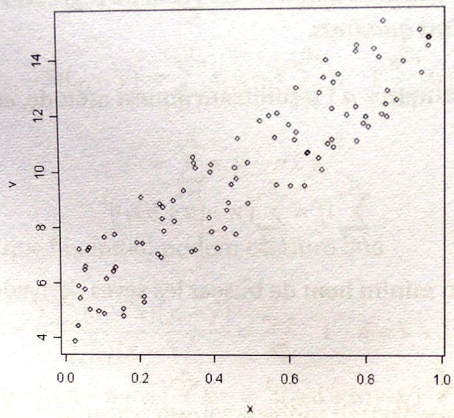
Sembla clar que sí que hi podem aplicar el model lineal.

Exemple 66. Vegem ara altres situacions.

Al primer gràfic ($r_{x,t} = 0.5858$) es veu que hi ha una relació entre les dues variables, però que en tot cas no és lineal. En el segon ($r_{x,y} = -0.047$) hi ha el que es coneix com un *núvol de punts* que indica que no hi ha relació entre les dues variables. Només en el tercer ($r_{x,v} = 0.915$) sembla que hi ha una relació líneal entre les dues variables i que, per tant, podem aplicar el model de regressió lineal.







El pas següent és intentar determinar la recta que s'ajusti bé a les dades. Recordem que una recta té una expressió del tipus

$$y = a + b\,x,$$

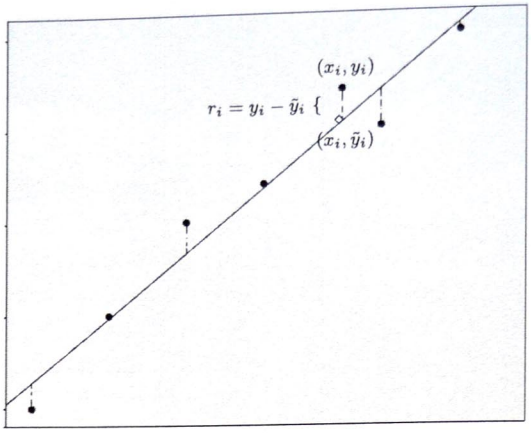
on la a és l'ordenada a l'origen i la b és el pendent de la recta. De rectes possibles, però, n'hi ha moltes, i hem de decidir quina és la que millor s'ajusta a les observacions que tenim. El mètode que utilitzarem per a calcular-la és el que presentem a l'apartat següent:

7.1.1. Mètode dels mínims quadrats

Donada una recta, per a cada punt x_i podem calcular la seva imatge o **predicció** $\widetilde{y}_i = a + b\,x_i$ i el seu corresponent error o **residu**, que serà la distància vertical entre la recta i el punt, és a dir,

$$r_i = y_i - (a + b\,x_i).$$

Considerarem que la recta que millor s'ajusta a les observacions és la que minimitza els residus corresponents. Per tal d'evitar que si sumem els residus es puguin cancel·lar entre ells si són de diferent signe, buscarem la recta que minimitzi la suma dels residus al quadrat.



Definició 34. *El* **mètode dels mínims quadrats** *consistex a determinar els valors a i b que minimitzen la suma dels residus al quadrat. La* **recta de regressió** *és la que s'obté quan s'utilitza el mètode dels mínims quadrats.*

Per a calcular els paràmetres a i b utilitzant aquest mètode, ens proposem minimitzar l'expressió següent:

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \big(y_i - (a + b\,x_i)\big)^2 .$$

Per determinar-ne un mínim hem de buscar les seves derivades parcials i igualar-les a zero:

$$\begin{aligned}
\frac{\partial}{\partial a} \sum_{i=1}^{n} \big(y_i - (a + b\,x_i)\big)^2 &= -2 \sum_{i=1}^{n} \big(y_i - a - b\,x_i\big) = 0, \\
\frac{\partial}{\partial b} \sum_{i=1}^{n} \big(y_i - (a + b\,x_i)\big)^2 &= -2 \sum_{i=1}^{n} x_i \big(y_i - a - b\,x_i\big) = 0.
\end{aligned}$$

Fixeu-vos que ens ha quedat un sistema amb dues equacions i dues incògnites a i b, de manera que ara es tracta tan sols d'aïllar les incògnites. De la primera equació podem deduir fàcilment que

$$a = \frac{1}{n} \left(\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \right) = \overline{y} - b\,\overline{x},$$

i substituint aquest valor a la segona equació tenim:

$$\begin{aligned}
\sum_{i=1}^{n} x_i\,y_i - \big(\overline{y} - b\,\overline{x}\big) \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} x_i^2 &= 0, \\
b \left(n\,\overline{x}^2 - \sum_{i=1}^{n} x_i^2 \right) + \sum_{i=1}^{n} x_i\,y_i - n\,\overline{x}\,\overline{y} &= 0,
\end{aligned}$$

i, per tant, tenim que

$$b = \frac{\sum_{i=1}^{n} x_i\,y_i - n\,\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\,\overline{x}^2}.$$

Intentem reescriure aquesta expressió recuperant les definicions de variància mostral:

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

i de covariància mostral:

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

i utilitzant també les relacions següents:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x})^2 &= \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \\
\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n} x_i y_i - \overline{x}\sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} x_i + \overline{x}\,\overline{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}.
\end{aligned}$$

Fixeu-vos, per tant, que finalment podem escriure que

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{x,y}}{s_x^2} \quad \text{i} \quad \hat{a} = \overline{y} - \hat{b}\,\overline{x}.$$

Per acabar, ens queda comprovar que les segones derivades són positives:

$$\begin{aligned}
\frac{\partial^2}{\partial a^2}\sum_{i=1}^{n}(y_i - (a + b\,x_i))^2 &= 2n > 0, \\
\frac{\partial^2}{\partial b^2}\sum_{i=1}^{n}(y_i - (a + b\,x_i))^2 &= 2\sum_{i=1}^{n} x_i^2 > 0.
\end{aligned}$$

Un cop determinada la recta de regressió

$$y = \hat{a} + \hat{b}\,x,$$

per a cada valor x_i podrem calcular les prediccions corresponents:

$$\widehat{y}_i = \hat{a} + \hat{b}\,x_i.$$

A partir d'ara, quan parlem dels residus ens referirem només als errors corresponents a aquesta recta i utilitzarem la notació:

$$e_i = y_i - \widehat{y}_i = y_i - (\widehat{a} + \widehat{b}\,x_i).$$

Pel fet que a i b minimitzen el mètode dels mínims quadràtics, també s'obté i es dedueix que

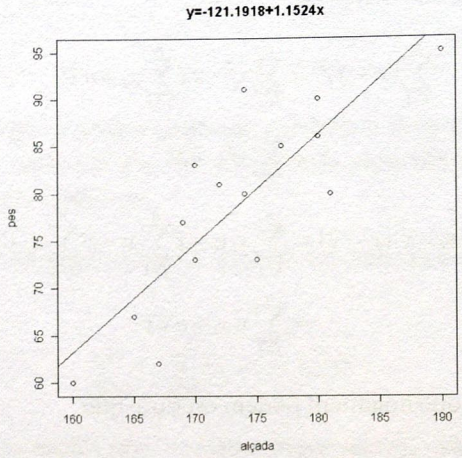
$$\sum_{i=1}^{n} e_i = 0 \quad \text{i} \quad \sum_{i=1}^{n} x_i e_i = 0.$$

Exemple 67. Fent els càlculs corresponents per al nostre exemple, trobem que

$$a = -121.1918 \quad \text{i} \quad b = 1.1524,$$

és a dir, que tenim la recta

$$y = -121.1918 + 1.1524x.$$



Podem utilitzar la recta per a calcular també prediccions i residus. Per exemple, per a la primera observació teníem una alçada de 190 i un pes de 95. Ara bé, a una alçada de 190 li correspon una predicció de

$$-121.1918 + 1.1524 \cdot 190 = 97.7642.$$

Tenim, per tant, un residu de

$$95 - 97.7642 = -2.7642.$$

7.2. Qualitat de l'ajustament

Hem determinat la recta de regressió, que és la recta que, segons el criteri dels mínims quadrats, millor s'ajusta a les dades. Que sigui, però, la que millor s'ajusta no significa que la recta s'ajusti bé a les dades. Per decidir si la recta que hem calculat s'ajusta bé a les dades, calcularem l'anomenat *coeficient de determinació.*

Definició 35. *El* **coeficient de determinació** *es defineix com la proporció de variància explicada per la recta de regressió, és a dir, la variància de les prediccions divida per la variància de les dades inicials:*

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}.$$

Observeu que aquesta fórmula surt del fet que quan calculem les mitjanes de les prediccions s'observa que

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i) + \frac{1}{n}\sum_{i=1}^{n} y_i$$
$$= -\frac{1}{n}\sum_{i=1}^{n} e_i + \overline{y} = \overline{y}.$$

Habitualment, però, el coeficient de determinació s'expressa d'una altra manera. Definim

$$\text{SQT} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \text{ com la suma de quadrats totals.}$$

Es pot escriure:

$$\text{SQT} = \text{SQR} + \text{SQE},$$

on

$$\text{SQE} = \sum_{i=1}^{n} e_i^2 \text{ és la suma dels quadrats dels errors.}$$

$$\text{SQR} = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \text{ és la suma de quadrats de la regressió.}$$

Així, el coeficient de determinació es pot escriure com

$$R^2 = \frac{\text{SQR}}{\text{SQT}} = 1 - \frac{\text{SQE}}{\text{SQT}}.$$

Com l'hem d'interpretar?

Fixeu-vos primer que R^2 serà sempre un valor entre 0 i 1. I, a més, observem:

- $R^2 = 1$ ens diu que SQE = 0 o, el que és el mateix, que tots els residus són zero i que, per tant, la recta s'ajusta perfectament a tots els punts. Hi ha, per tant, una relació lineal entre les dues variables.
- $0 < R^2 < 1$ ens dóna la proporció de variabilitat explicada per la recta de regressió. Evidentment, com més a prop d'1 sigui aquest valor millor serà l'ajust. Per exemple, si tinguéssim $R^2 = 0.6$ en el nostre exemple ens indicaria que l'alçada explicaria el 60% del pes, mentre que el 40% restant seria degut a altres factors.
- $R^2 = 0$ és el cas en què la recta de regressió no explica res. Indica que no hi ha cap relació entre les dues variables.

7.2.1. Relació amb el coeficient de correlació lineal

Abans de fer la recta de regressió hem calculat el coeficient de correlació lineal per veure si tenia sentit calcular la recta de regressió, ja que ens mesura el grau de relació entre les dues variables. És fàcil de comprovar que

$$R^2 = r_{x,y}^2.$$

En efecte, observem primer que

$$\widehat{y}_i - \overline{y} = \widehat{a} + \widehat{b}\,x_i - (\widehat{a} + \widehat{b}\,\overline{x}) = \widehat{b}(x_i - \overline{x});$$

aleshores,

$$R^2 = \frac{\sum_{i=1}^n \left(\widehat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^n \left(y_i - \overline{y}\right)^2} = \frac{\widehat{b}^2 \sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n \left(y_i - \overline{y}\right)^2} = \left(\frac{s_{x,y}}{s_x^2}\right)^2 \frac{s_x^2}{s_y^2} = \frac{s_{x,y}^2}{s_x^2 s_y^2} = r^2.$$

Observació 28. Hi ha, però, una diferència: la r ens dóna una mica més d'informació gràcies al seu signe. Quan aquest és positiu, ens indica que quan la variable X creix la variable Y també ho fa, mentre que quan és negatiu ens indica que quan X creix la variable Y decreix.

7.2.2. Anàlisi dels residus

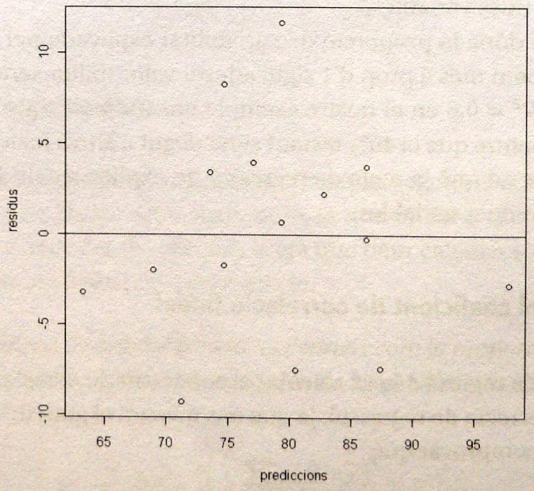
L'anàlisi dels residus és un altre mètode que ens permet comprovar si el model de regressió que hem calculat és bo.

Aquest mètode consisteix a dibuixar un diagrama de dispersió on a l'eix d'abscisses hi posarem les prediccions $\widehat{y}_i$ i a l'eix d'ordenades, els residus e_i. Si el model és bo, esperarem que els residus no depenguin de les prediccions, de manera que quan fem el diagrama esperarem que ens surti un núvol de punts sense cap forma determinada. En cas contrari, ens indica que hi ha una relació que no hem sabut explicar.

Exemple 68. En el nostre exemple podem comprovar que

$$R^2 = 0.68418;$$

és a dir, que la variable alçada explica més del 68.4% de la variable pes. Respecte a l'anàlisi dels residus, obtenim el gràfic següent, que ens dóna un núvol de punts. Tenim, per tant, un bon model de regressió.



7.3. Inferència en el model de regressió

Fins ara hem estudiat el model de regressió des d'un punt de vista descriptiu, a partir de les dades observades i sense tenir en compte les distribucions de les variables implicades. Per tal de poder fer inferència, hem de mirar-nos el model de regressió des d'un punt de vista més probabilista.

Podem considerar que les nostres dades (x_i, y_i), $i = 1, \ldots, n$ són observacions d'un model del tipus

$$Y = \alpha + \beta X + \Xi,$$

on Ξ és una variable aleatòria $N(0, \sigma^2)$ independent de X i de Y.

Ho podem interpretar de la manera següent: per a cada valor x de la variable predictora X, la variable resposta Y és del tipus $\alpha + \beta x + \Xi$ i segueix, per tant, una distribució $N(\alpha + \beta x, \sigma^2)$, on $\alpha + \beta x$ és la part que ens explica la variable X i Ξ es correspon amb el residu.

Tenim, doncs, un model estadístic amb tres paràmetres desconeguts α, β i σ^2, de manera que hi podem aplicar les diverses tècniques que hem vist en aquest curs: intervals de confiança i contrastos d'hipòtesis.

No entrarem en l'estudi d'intervals de confiança i contrastos d'hipòtesis. De totes maneres, convé saber que es poden obtenir intervals de confiança per a α i per a β, o que es poden fer, per exemple, contrastos del tipus

$$\mathrm{H}_0: \alpha = 0 \quad \text{contra} \quad \mathrm{H}_1: \alpha \neq 0$$

o

$$\mathrm{H}_0: \beta = 0 \quad \text{contra} \quad \mathrm{H}_1: \beta \neq 0$$

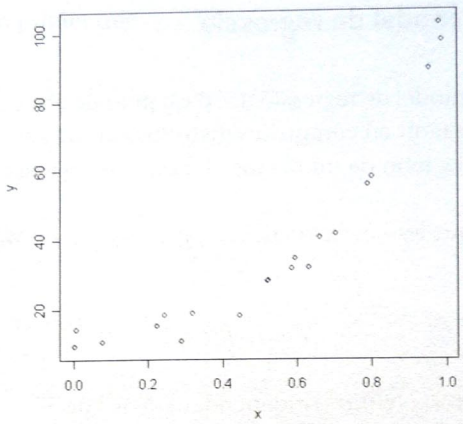
amb qualsevol programari estadístic que utilitzeu. Podeu trobar també les funcions pivotants o els estadístics de contrast corresponents en qualsevol llibre o manual d'estadística.

Penseu només que per aplicar aquestes tècniques cal que el model compleixi algunes condicions. És important, per exemple, que Ξ sigui una variable aleatòria $N(0, \sigma^2)$ independent de X i de Y. Un primer mètode per a comprovar si això és cert és fer l'anàlisi dels residus que ja hem comentat.

7.4. Altres models

Hem estudiat fins ara el model de regressió lineal simple. Existeixen, però, altres tipus de models.

Suposem, de moment, que tenim com abans una variable de resposta Y i una variable predictiva X, però que en fer el diagrama de dispersió ens surt un gràfic del tipus:

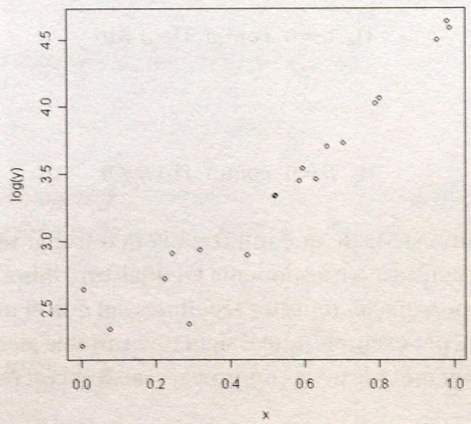


Sembla que en aquest cas un model lineal no és gaire adient, però en canvi podem intentar mirar si una relació del tipus

$$Y = a\,b^X + \xi$$

s'ajusta millor a les mostres dades. Aquest tipus de model s'anomena **model exponencial** i es pot convertir en un model de regressió lineal prenent logaritmes.

Fixeu-vos que si fem el gràfic dels punts $(x_i, \ln(y_i)), i = 1, \ldots, n$, obtenim un gràfic del tipus:



Així, podem considerar el model

$$\ln Y = a + X \ln b + \xi'.$$

Un altre model, entre molts altres, que també podem utilitzar considerant només dues variables és un **model polinomial**:

$$Y = a + b_1 X + b_2 X^2 + \xi.$$

Una altra extensió molt important que cal tenir en compte és el **model de regressió lineal múltiple**. Es tracta ara de considerar el cas en què hi ha una variable resposta Y que

depèn de diverses variables explicatives $X_1, \ldots, X_k$. En aquest model s'estudien relacions del tipus

$$Y = a + b_1 X_1 + \cdots + b_k X_k + \xi,$$

on ξ indica, com sempre, la part aleatòria corresponent als residus. En l'exemple del pes i l'alçada, podem considerar, a més de la variable explicativa alçada, altres variables com el sexe o l'edat.

Bibliografia

BARDINA, X.; FERRÉ, M. *Estadística Descriptiva*. Bellaterra: Servei de Publicacions UAB, 2009.

CANAVOS, C. G. *Probabilidad y Estadística. Aplicaciones y métodos*. Madrid: Mc Graw-Hill, 1988.

CUADRAS, C. M. *Problemas de probabilidad y estadística*. 2a ed. Barcelona: EUB, 1999.

CHUNG, K. L. *Elementary probability theory*. Nova York: Springer, 2003.

FORTIANA, D.; NUALART, D. *Estadística*. Barcelona: Edicions Universitat de Barcelona, 1999.

GIBERGANS, J.; GIL, A.; ROVIRA, C. *Estadística*. Barcelona: EDIUOC, 2002.

GONICK, L.; SMITH, W. *La estadística en comic*. Barcelona: Zendrera Zariquiey, 1999.

IMAN, R. L.; CONOVER, W. J. *A modern approach to statistics*. Nova York: Wiley, 1983.

KINNEY, J. J. *Probability. An introduction with Statistical Applications*. Nova York: Wiley, 1997.

ROVIRA, C. *Estadística bàsica*. Barcelona: Edicions Universitat de Barcelona, 2003.

ZAIATS, V. *Probabilitat i estadística: Exercicis I*. Vic: Eumo, 1998.

Aquest llibre s'adreça tant als alumnes de l'assignatura Probabilitats i Estadística del grau d'Informàtica com a qualsevol estudiant que cursi una assignatura bàsica de probabilitats i d'estadística de titulacions científicotècniques. L'obra s'estructura en dos blocs: el càlcul de probabilitats i la inferència estadística. En la primera part s'inclouen les nocions bàsiques de probabilitat; les variables aleatòries discretes i absolutament contínues, i una versió del Teorema del Límit Central. En la segona part s'estudia l'estimació per intervals de confiança i es presenten els tests d'hipòtesis més habituals.

Universitat de Barcelona
Publicacions i Edicions
Compromís amb el coneixement
www.publicacions.ub.edu
9 788447 536535