ESTADÍSTICA

Notes de Classe

J. M. Aroca

Departament de Matemàtiques,
Universitat Politècnica de Catalunya

e-mail:josep.m.aroca@upc.edu
https://mat-web.upc.edu/people/josep.m.aroca/pie.html

Març de 2018

Índex

Introducció

Aquests apunts desenvolupen l'Estadística assumint un coneixement de la teoria bàsica de la probabilitat i de les variables aleatòries.

Els càlculs i les gràfiques de tipus estadístic es realitzen normalment amb programes informàtics. Aquí s'ha triat el programa *R* que és un programari lliure cada vegada més utilitzat en estadística. El programa es pot descarregar de `http://r-project.org`.

Notació

♦ Final d'exemple.

DEM: Inici de demostració.

♣ Final de demostració.

$P(A)$ Probabilitat de l'esdeveniment A.

$X, Y, Z, \ldots$ Variable aleatòries (En majúscules. Els valors numèrics es posen en minúscula. Així, per exemple, $X < x$ és l'esdeveniment: la variable aleatòria X és menor que el nombre x.).

$F_X(x)$ Funció de distribució de la variable aleatòria X.

$f_X(x)$ Funció de densitat de la variable aleatòria contínua X.

$P_X(x)$ Funció de probabilitat de la variable aleatòria discreta X.

$E[X], \overline{X}, \mu_X$ Esperança de la variable aleatòria X.

$V[X], \sigma_X^2$ Variància de la variable aleatòria X.

σ_X Desviació estàndard de la variable aleatòria X.

$F_{XY}(x, y)$ Funció de distribució conjunta de les variables aleatòries X i Y.

$f_{XY}(x, y)$ Funció de densitat conjunta de les variables aleatòries X i Y.

$f_X(x \,|\, y)$ Funció de densitat de la variable aleatòria X condicionada a $Y = y$. (De manera similar tenim $f_Y(y \,|\, x)$.)

$C[X, Y]$ Covariància de les variables aleatòries X i Y.

ρ Coeficient de correlació.

Bibliografia

1 Spiegel, M., Schiller, J.J., Alu Srinivasan, R. (2009) *Schaum's Outline of Probability and Statistics, Third Edition.* Col·lecció Shaum, McGraw-Hill. ISBN: 978-0-07-154426-9.

Conté resums de teoria, problemes fets amb detall i problemes plantejats, amb les solucions al final. Molt recomanable per complementar aquests apunts.

2 Ross, S. M. (2009) *Introduction to probability and statistics for engineers and scientists.* 4th ed. Amsterdam. Elsevier, cop. 2009. ISBN: 978-0123704832.

Un llibre estàndard d'estadística.

3 Baron R. (2013) *Probability and Statistics for Computer Scientists.* 2 edition. Chapman and Hall/CRC. ISBN: 978-1439875902.

La seva part d'estadística és bastant concisa i orientada a l'enginyeria informàtica i de telecomunicacions.

4 Leon-Garcia, A. (2008) *Probability, statistics and random processes for electrical engineering.* 3rd ed. Upper Saddle River, NY: Pearson Prentice Hall, cop. ISBN: 978-0131471221.

Conté un capítol d'estadística molt enfocat a l'enginyeria. Tracta alguns aspectes que no es troben en llibres més estàndard d'estadística.

5 Teetor, P. (2011) *R Cookbook.* O'Reilly. ISBN: 978-0-596-80915-7.

Una bona visió general de R i les seves aplicacions a l'estadística.

6 Spiegel, M., Liu, J., Abellanas, L. (2005) *Manual de fórmulas y tablas de Matemática aplicada.* 3a ed. Col·lecció Shaum, McGraw-Hill. ISBN: 8448198409.

Les seves taules sobre variables aleatòries ja estan superades per les eines informàtiques. Encara així és útil per fórmules trigonomètriques, de geometria, primitives, transformades de Fourier, etc.

Capítol 1

Paràmetres de les variables aleatòries

A més de l'esperança i la desviació d'una variable aleatòria hi ha altres paràmetres que tenen rellevància quan, a través de mesures estadístiques, hem de trobar les propietats d'una variable de tipus o paràmetres desconeguts.

Recordem que el moment n-èsim de la variable aleatòria X és el nombre $E[X^n]$, i el moment central n-èsim de la variable aleatòria X és el nombre $E[(X-\mu)^n]$, on μ és l'esperança de X. Aquí, $n = 0, 1, 2, \ldots$. El primer moment és l'esperança de X: $\mu = E[X]$. El segon moment central és la variància de X: $\sigma^2 = E[(X-\mu)^2]$. σ és la desviació de X.

μ ens dóna un valor de centralització de X mentre que σ és un valor de dispersió. Altres paràmetres permeten mesurar aspectes de la forma de la distribució de probabilitat de X (funció de densitat f_X si X és contínua o funció de probabilitat P_X si X és discreta) i es descriuen a continuació.

1.1 Asimetria

El coeficient d'asimetria de la variable aleatòria X és el nombre:

$$\kappa_3 = \frac{E[(X-\mu)^3]}{\sigma^3}. \tag{1.1}$$

Notem que si la densitat de X és simètrica al voltant de X aquest coeficient val zero.

Exemple 1.1 Si $X \sim N(\mu, \sigma)$ el coeficient d'asimetria es zero per simetria de la campana de Gauss al voltant de μ.

Si $X \sim \text{Exp}(\lambda)$, recordem que $\mu = \sigma = \lambda^{-1}$ i els moments valen $E[X^n] = \frac{n!}{\lambda^n}$. Els moments centrals es poden anar calculant a partir d'aquests, desenvolupant la potència dins de l'esperança:

$$E[(X-\mu)^3] = E[X^3 - 3\mu X^2 + 3\mu^2 X - \mu^3] = E[X^3] - 3\mu E[X^2] + 3\mu^2 E[X] - \mu^3 =$$
$$\frac{6}{\lambda^3} - 3\lambda^{-1}\frac{2}{\lambda^2} + 3\lambda^{-2}\lambda^{-1} - \lambda^{-3} = 2\lambda^{-3},$$

d'on el coeficient d'asimetria és 2. ♦

1.2 Curtosi

El coeficient de curtosi de la variable aleatòria X és el nombre:

$$\kappa_4 = \frac{E[(X-\mu)^4]}{\sigma^4}. \tag{1.2}$$

Se sol interpretar com una mesura del grau d'aplanament de la densitat al voltant de μ.

Exemple 1.2 Si $X \sim \mathrm{N}(\mu, \sigma)$ recordem que els moments centrals valen zero per n imparella i $\sigma^n \dfrac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!}$ per n parella. Llavors $E[(X-\mu)^4] = \sigma^4 \frac{4!}{2^2 \cdot 2!} = 3\sigma^4$ d'on la curtosi val 3. Aquest valor se sol prendre com a referència de manera que en ocasions es defineix la curtosi com $\dfrac{E[(X-\mu)^4]}{\sigma^4} - 3$ (aquest valor s'anomena també *excés de curtosi*).

Si $X \sim$ Uniforme$(-L, L)$, $\mu = 0$ amb el que els moments centrals coincideixen amb els moments ordinaris. Aquests valen 0 per n imparella mentre que per n parella:

$$E[X^n] = \int_{-L}^{L} x^n \frac{1}{2L} dx = \frac{1}{L} \int_0^L x^n dx = \frac{L^n}{n+1}.$$

Llavors, amb $n = 2$, $\sigma^2 = \dfrac{L^2}{3}$. El quart moment val $\dfrac{L^4}{5}$ i la curtosi $\dfrac{9}{5}$. Resulta ser menor que 3, de manera consistent amb el fet que la densitat té forma aplanada en comparació amb la gaussiana.

Si X és una variable de Laplace amb densitat $f(x) = \dfrac{\lambda}{2} e^{-\lambda|x|}$ per $-\infty < x < \infty$, per simetria trobem de nou $\mu = 0$. Els moments imparells són nuls i els parells valen:

$$E[X^n] = \int_{-\infty}^{\infty} x^n \frac{1}{2} e^{-\lambda|x|} dx = \int_0^{\infty} x^n e^{-\lambda|x|} dx = \frac{n!}{\lambda^n} \qquad (n \text{ parella}).$$

Així, $\sigma^2 = \dfrac{2}{\lambda^2}$ i $E[X^4] = \dfrac{24}{\lambda^4}$. La curtosi val 6, més gran que 3 consistent amb la forma punxeguda de la densitat.

La gràfica mostra les tres densitats fent $\mu = 0$, $\sigma = 1$ en la normal (vermell) i prenent $L = \sqrt{3}$ en la uniforme (blau) i $\lambda = \sqrt{2}$ en la de Laplace (verd) per a que la desviació sigui la mateixa en les tres variables.

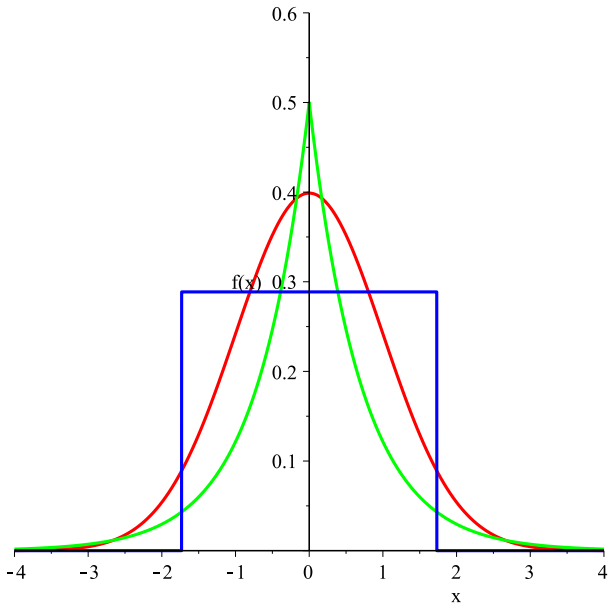

Figura 1.1: Comparació de curtosis.

♦

1.3 Modes

Una moda és un valor de x on la densitat (o funció de probabilitat en el cas discret) de la variable aleatòria X té un màxim local. En algunes variables bàsiques hi ha una sola moda (variables

normals, exponencials, de Cauchy,...) però una variable pot tenir vàries modes. La gràfica 1.2 mostra una variable amb dues modes.

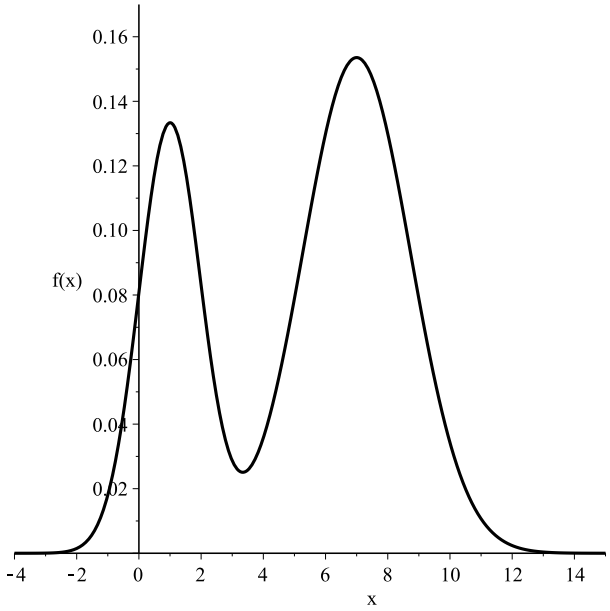

Figura 1.2: Exemple de distribució bimodal.

Notem que algunes variables aleatòries, com la de Cauchy, no tenen moments definits ja que les integrals que els defineixen no són convergents. Els paràmetres que definim a continuació no presenten aquest possible problema i tenen alguns avantatges en estadística.

1.4 Mediana

La mediana d'una variable aleatòria contínua X és un nombre m tal que $P(X \leq m) = P(X \geq m) = \frac{1}{2}$. Notem que $P(X \leq m)$ és $F_X(m)$. L'equació $F_X(m) = \frac{1}{2}$ té solucions ja que, en aquest cas, F_X és contínua i creixent de 0 a 1. La solució normalment és un únic punt. Si resulta ser un interval, per conveni prenem el punt mig com a mediana.

En el cas de variables discretes, si prenen valors x_k, ordenats:

- Si hi ha algun valor x_s tal que $F_X(x_s) = \frac{1}{2}$, prenem $m = \frac{x_s + x_{s+1}}{2}$
- En cas contrari prenem m igual a la mínima x_k tal que $F_X(x_k) > \frac{1}{2}$.

Exemple 1.3 Si $X \sim \mathrm{N}(\mu, \sigma)$, $m = \mu$. Això passa amb les variables contínues amb densitat simètrica al voltant de μ.

Si $X \sim \mathrm{Exp}(\lambda)$, $F_X(x) = 1 - e^{-\lambda x} = \frac{1}{2}$ té solució $m = \frac{\ln 2}{\lambda}$. Notem que en aquest cas $\mu = \lambda^{-1}$ i $m \neq \mu$, d'acord amb el caràcter esbiaixat de la variable.

Si X és la variable que dóna el resultat en la tirada d'un dau, qualsevol valor $3 < m < 4$ té $\frac{1}{2}$ de probabilitat a dreta i esquerra. Prenem $m = 3.5$.

Si X és de Cauchy amb paràmetre α, $F_X(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg} \frac{x}{\alpha} = \frac{1}{2}$ té solució $m = 0$. Així, la mediana val zero mentre que l'esperança μ no està definida.

Si $\Omega_X = \{1,2,3,4\}$ i $P_X(1) = 0.2$, $P_X(2) = 0.3$, $P_X(3) = 0.4$, $P_X(4) = 0.1$, la mediana val $m = 2.5$.

Si $\Omega_X = \{1,2,3,4\}$ i $P_X(1) = 0.4$, $P_X(2) = 0.3$, $P_X(3) = 0.2$, $P_X(4) = 0.1$, la mediana val $m = 2$. ♦

1.5 Quantils

Donat $0 < p < 1$ diem que x_p és el p-quantil de la variable aleatòria X si $P(X \leq x_p) = p$. Notem que el p-quantil es troba resolent l'equació $F_X(x_p) = p$. Si la funció de distribució és estrictament creixent, la solució és única i podem dir que el p-quantil és $x_p = F_X^{-1}(p)$. La funció $F_X^{-1} : (0,1) \to \mathbb{R}$ s'anomena **funció de quantils**.

1.6 Quartils

De manera estàndard es fixa l'atenció en els p-quantils pels casos $p = \dfrac{1}{4}$, $p = \dfrac{1}{2}$ i $p = \dfrac{3}{4}$ i se'ls anomena Q_1, Q_2, i Q_3, respectivament. El primer quartil té a la seva esquerra el 25% de la probabilitat. El segon quartil coincideix amb la mediana i té a la seva esquerra el 50% de la probabilitat. El tercer quartil té a la seva esquerra el 75% de la probabilitat. Els tres quartils divideixen la recta en quatre regions, cadascuna amb probabilitat igual a $\dfrac{1}{4}$.

Exemple 1.4 Considerem X exponencial amb $\lambda = 1$. L'equació pels quantils és $F(x) = 1 - e^{-x} = p$, d'on la funció de quantils és $x = -\ln(1-p)$. Llavors els quartils valen: $Q_1 = -\ln\left(1 - \dfrac{1}{4}\right) = \ln\dfrac{4}{3} = 0{,}288$, $m = Q_2 = -\ln\left(1 - \dfrac{1}{2}\right) = \ln 2 = 0{,}693$, $Q_3 = -\ln\left(1 - \dfrac{3}{4}\right) = \ln 4 = 1{,}386$. ♦

1.7 Recorregut interquartílic

Una mesura de la dispersió, alternativa a la desviació típica σ, és el recorregut interquartílic:

$$\mathrm{IQR} = Q_3 - Q_1. \tag{1.3}$$

Exemple 1.5 Si X és exponencial amb $\lambda = 1$, $\mathrm{IQR} = \ln 4 - \ln\frac{4}{3} = \ln 3 = 1{,}1$. Notem que, per aquesta variable, $\sigma = \lambda^{-1} = 1$.

Si X és de Cauchy amb paràmetre $\alpha = 1$, la seva desviació no està definida. En canvi, podem resoldre $F_X(x) = \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg} x = p$ d'on $x = \operatorname{tg}\left(\pi\left(p - \frac{1}{2}\right)\right)$. Ara podem obtenir $Q_1 = -1$, $Q_2 = 0$, $Q_3 = 1$ i $\mathrm{IQR} = 2$. ♦

1.8 Relació entre variables aleatòries

En aquest apartat considerem variables aleatòries contínues amb funció de distribució estrictament creixent.

Hem vist anteriorment la funció dels p-quantils, F_X^{-1} que aplica l'interval $(0,1)$ sobre la recta real (o l'interval de valors que pren la variable X). Notem ara el següent: si X és una variable aleatòria qualsevol, la nova variable $U = F_X(X)$ (és a dir, transformen X aplicant com a funció de transformació F_X) pren valors en $(0,1)$ (recorregut de F_X) i té densitat:

$$f_U(u) = f_X(x)\frac{1}{|du/dx|} = \frac{f_X(x)}{F_X'(x)} = 1.$$

Llavors U és una variable uniforme en l'interval $(0,1)$.

Així podem obtenir valors d'una variable amb la densitat que vulguem a partir de valors d'una variable U, uniforme en $(0,1)$, fent $X = F_X^{-1}(U)$, és a dir, transformant la uniforme amb la funció de quantils.

Una altra aplicació d'aquest fet és que donades dues densitats f_X i f_Y hi ha una transformació $Y = g(X)$ que implica la transformació de la densitat de X en la de Y. Com $X = F_X^{-1}(U)$ i $Y = F_Y^{-1}(U)$, $g = F_Y^{-1} \circ F_X$. Si anem donant valors al paràmetre $p \in (0,1)$, anem obtenint punts (x,y) on $x = F_X^{-1}(p)$ i $y = F_Y^{-1}(p)$ (notem que x i y són els p-quantils de les variables X i Y respectivament). Aquest punts descriuen una corba en el pla corresponent a la relació $y = g(x)$. Aquest procediment s'anomena **comparació de quantils**. En estadística es comparen els quantils d'una distribució teòrica (generalment la $N(0,1)$) amb els quantils obtinguts per mostreig estadístic. Si la distribució empírica és normal, la gràfica ha de mostrar una recta.

Capítol 2

Variables aleatòries importants en estadística

En aquest capítol tractem algunes variables aleatòries que juguen un paper rellevant en l'Estadística. Abans discutirem el concepte de funció característica que és útil en alguns càlculs i demostracions.

2.1 Funció característica

La funció característica d'una variable aleatòria és:

$$C_X(\omega) = E[e^{j\omega X}]. \tag{2.1}$$

Pel teorema de l'esperança, $C_X(\omega) = \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)dx$. Així, $C_X(\omega)$ és la transformada de Fourier[1] de la funció de densitat. La transformada existeix ja que f_X és una funció de mòdul integrable a $\mathbb{R}$.

En el cas d'una variable n-dimensional:

$$C_{X_1,X_2,\ldots,X_n}(\omega_1,\omega_2,\ldots,\omega_n) = E[e^{j(\omega_1 X_1+\omega_2 X_2+\cdots+\omega_n X_n)}]. \tag{2.2}$$

Propietats:

(1) La funció característica determina la funció de densitat (fent la transformada inversa).

(2) $C_X(0) = 1$. En el cas multidimensional, obtenim la funció característica d'un subconjunt de les variables (marginal, per tant) fent $\omega_i = 0$ per les variables que no són del subconjunt. Per exemple $C_{X_1}(\omega) = C_{X_1,X_2}(\omega, 0)$, etc.

(3) Les variables $X_1, X_2, \ldots, X_n$ són independents si i només si $C_{X_1,X_2,\ldots,X_n}(\omega_1,\omega_2,\ldots,\omega_n) = C_{X_1}(\omega_1)C_{X_2}(\omega_2)\cdots C_{X_n}(\omega_n)$.

(4) El moment k-èsim d'una variable es pot obtenir derivant la funció característica k vegades i posant $\omega = 0$: $E[X^k] = (-j)^k \dfrac{d^k C_X}{d\omega^k}(0)$.

(5) Si $Z = X_1 + X_2 + \ldots + X_n$, on les X_i son variables independents:

$$C_Z(\omega) = C_{X_1}(\omega)C_{X_2}(\omega)\cdots C_{X_n}(\omega).$$

Aquest resultat és el teorema de convolució.

[1] Habitualment la transformada de Fourier es defineix amb un signe menys a l'exponent i en termes de la variable f on $\omega = 2\pi f$. No és difícil relacionar les dues versions.

El següent exemple és rellevant:

Exemple 2.1 Si $X \sim \mathrm{N}(\mu, \sigma)$:

$$C_X(\omega) = e^{j\mu\omega - \frac{1}{2}\sigma^2\omega^2}. \tag{2.3}$$

DEM:

$$C_X(\omega) = \int_{-\infty}^{\infty} e^{j\omega x} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx = \frac{e^{j\mu\omega}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{j\omega\sigma t - \frac{t^2}{2}} dt = e^{j\mu\omega - \frac{1}{2}\sigma^2\omega^2}.$$

(amb el canvi $x = \mu + \sigma t$ i les fórmules d'integració gaussiana.) ♣

Notem que $C_X(\omega)$ ha resultat també una funció gaussiana (exponencial d'un polinomi de segon grau). Així podem definir una variable normal com aquella que té densitat igual a l'exponencial d'un polinomi de segon grau o, alternativament, com aquella que té funció característica igual a l'exponencial d'un polinomi de segon grau.

Tenim, ara, una demostració fàcil del següent resultat:

Donades les dues variables independents, $X \sim \mathrm{N}(\mu_1, \sigma_1)$ i $Y \sim \mathrm{N}(\mu_2, \sigma_2)$, la seva suma $Z = X + Y$ és $Z \sim \mathrm{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

DEM: Per la propietat (5) de la funció característica: $C_Z(\omega) = C_X(\omega)C_Y(\omega) = e^{j\mu_1\omega - \frac{1}{2}\sigma_1^2\omega^2} e^{j\mu_2\omega - \frac{1}{2}\sigma_2^2\omega^2} = e^{j(\mu_1+\mu_2)\omega - \frac{1}{2}(\sigma_1^2+\sigma_2^2)\omega^2}$. ♣ ♦

2.2 Distribució gamma

Aquesta variable és la base per alguns raonaments que farem amb les variables normal i khi quadrat.

La variable aleatòria de tipus gamma amb paràmetres $\alpha > 0$ (forma) i $\beta > 0$ (escala) és la que té densitat:

$$f_X(x) = Kx^{\alpha-1}e^{-\frac{x}{\beta}}, \quad x > 0. \tag{2.4}$$

Escriurem $X \sim \mathrm{Gamma}(\alpha, \beta)$. K és una constant que es determina:

$$1 = K \int_0^{\infty} x^{\alpha-1}e^{-\frac{x}{\beta}} dx = K\beta^{\alpha} \int_0^{\infty} t^{\alpha-1}e^{-t} dt = K\beta^{\alpha}\Gamma(\alpha).$$

(canvi $x = \beta t$). Llavors $K = \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}$. Amb un càlcul similar trobem que:

$$C_X(\omega) = \frac{1}{(1 - j\beta\omega)^{\alpha}}. \tag{2.5}$$

DEM:

$$C_X(\omega) = \int_0^{\infty} e^{j\omega x} K x^{\alpha-1} e^{-\frac{x}{\beta}} dx = K \int_0^{\infty} x^{\alpha-1} e^{-x(\frac{1}{\beta} - j\omega)} dx = K \left(\frac{\beta}{1 - j\beta\omega}\right)^{\alpha} \Gamma(\alpha) = \frac{1}{(1 - j\beta\omega)^{\alpha}}. \clubsuit$$

Els moments valen:

$$E[X^k] = \beta^k \frac{\Gamma(k + \alpha)}{\Gamma(\alpha)} = \beta^k \alpha(\alpha + 1) \cdots (\alpha + k - 1). \tag{2.6}$$

En particular, l'esperança i la variància valen:

$$E[X] = \alpha\beta, \qquad V[X] = \alpha\beta^2. \tag{2.7}$$

DEM:

$$E[X^k] = \int_0^{\infty} x^k K x^{\alpha-1} e^{-\frac{x}{\beta}} dx = K \int_0^{\infty} x^{\alpha+k-1} e^{-\frac{x}{\beta}} dx = K\beta^{\alpha+k}\Gamma(\alpha + k) = \beta^k \frac{\Gamma(k + \alpha)}{\Gamma(\alpha)}.$$

$$E[X] = \beta \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha)} = \alpha\beta, \quad V[X] = E[X^2] - E[X]^2 = \beta^2 \frac{\Gamma(\alpha + 2)}{\Gamma(\alpha)} - (\alpha\beta)^2 = \beta^2(\alpha + 1)\alpha - \alpha^2\beta^2 = \alpha\beta^2. \clubsuit$$

Propietats:

(1) Donades les dues variables independents, $X \sim \text{Gamma}(\alpha_1, \beta)$ i $Y \sim \text{Gamma}(\alpha_2, \beta)$, la seva suma $Z = X + Y$ és $Z \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta)$.

DEM: $C_Z(\omega) = C_X(\omega)C_Y(\omega) = \dfrac{1}{(1-j\beta\omega)^{\alpha_1}} \dfrac{1}{(1-j\beta\omega)^{\alpha_2}} = \dfrac{1}{(1-j\beta\omega)^{\alpha_1+\alpha_2}}$. ♣

(2) Suposem que tenim tres variables X, Y, Z, verificant $Z = X + Y$ amb X i Y independents. Si $Z \sim \text{Gamma}(\alpha_Z, \beta)$ i $Y \sim \text{Gamma}(\alpha_Y, \beta)$ llavors és $X \sim \text{Gamma}(\alpha_Z - \alpha_Y, \beta)$.

DEM: Com abans, $C_Z(\omega) = C_X(\omega)C_Y(\omega)$ d'on $C_X(\omega) = \dfrac{C_Z(\omega)}{C_Y(\omega)} = \dfrac{1}{(1-j\beta\omega)^{\alpha_Z-\alpha_Y}}$. ♣

(3) Si $X \sim \text{N}(\mu, \sigma)$ llavors $Y = \dfrac{(X-\mu)^2}{\sigma^2}$ és de tipus Gamma $\left(\dfrac{1}{2}, 2\right)$.

DEM: Y pren valors en $(0, \infty)$. Hi ha dos valors de x corresponents a un valor de y donat i els dos donen la mateixa contribució a la densitat:

$$f_Y(y) = 2f_X(x)\frac{1}{|dy/dx|} = 2\frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\frac{1}{2|x-\mu|/\sigma^2} = \frac{1}{\sqrt{2\pi}}y^{-\frac{1}{2}}e^{-\frac{y}{2}}.$$

corresponent a la gamma amb $\alpha = \frac{1}{2}$ i $\beta = 2$. ♣

2.3 Variables normals

La variable normal unidimensional, $\text{N}(\mu, \sigma)$, és generalitza a dimensió arbitrària n prenent per a $(X_1, X_2, \ldots, X_n)$ una densitat que sigui l'exponencial d'un polinomi de segon grau en les variables $x_1, x_2, \ldots, x_n$. Ara ens concentrarem en la següent propietat:

$X_1, X_2, \ldots, X_n$ són conjuntament normals $\Longleftrightarrow$ la variable $\displaystyle\sum_{i=1}^{n} \alpha_i X_i$ és normal per a qualsevol valors de les constants α_i.

DEM: L'aspecte clau és que si sabem la funció característica d'un dels costats llavors queda determinada la de l'altre costat. Anomenem $Z = \displaystyle\sum_{i=1}^{n} \alpha_i X_i$. Llavors

$$C_Z(\omega) = E[e^{j\omega\sum_{i=1}^{n}\alpha_i X_i}].$$

Si fem $\omega = 1$ i $\alpha_i = \omega_i$, $i = 1, \ldots, n$, l'anterior expressió ens dóna $C_{X_1,X_2,\ldots,X_n}(\omega_1, \omega_2, \ldots, \omega_n)$. Inversament, donada

$$C_{X_1,X_2,\ldots,X_n}(\omega_1, \omega_2, \ldots, \omega_n) = E[e^{j(\omega_1 X_1 + \omega_2 X_2 + \cdots + \omega_n X_n)}],$$

fent $\omega_i = \omega\alpha_i$ obtenim $C_Z(\omega)$.

Ara, si en un dels costats tenim comportament normal, la funció característica és l'exponencial d'un polinomi de segon grau i, per tant, això passa també amb la funció característica de l'altre costat.

Per a més detall, si $\mu_i = E[X_i]$ i $C_{ij} = C[X_i, X_j]$:

$$C_{X_1,X_2,\ldots,X_n}(\omega_1, \omega_2, \ldots, \omega_n) = e^{j\sum_i \mu_i\omega_i - \frac{1}{2}\sum_{i,j} C_{ij}\omega_i\omega_j}, \qquad C_Z(\omega) = e^{j\omega\sum_i \mu_i\alpha_i - \frac{\omega^2}{2}\sum_{i,j} C_{ij}\alpha_i\alpha_j}. \clubsuit$$

2.4 Teorema del límit central

En l'apartat anterior hem vist que amb variables conjuntament normals, una combinació lineal de les variables també és normal. En particular, si $X_1, X_2, X_3, \cdots, X_n$ són variables independents, totes de tipus $\mathrm{N}(\mu, \sigma)$, llavors $Y = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ és normal amb valor mitjà $E[Y] = \mu$ i $V[Y] = \dfrac{\sigma^2}{n}$. Així, $Y \sim \mathrm{N}\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

Si les variables X_i no són de tipus normal, la densitat de Y depèn d'una forma complicada de n i la densitat de les X_i. El Teorema del Límit Central diu que Y és comporta com una variable normal quan n és gran. Més precisament:

Siguin $X_1, X_2, X_3, \cdots$ mesures independents corresponents a una variable aleatòria X d'esperança μ i variància σ^2. La mitjana aritmètica

$$Y_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

és tal que $\dfrac{Y_n - \mu}{\sigma/\sqrt{n}}$ tendeix, per a $n \to \infty$, a una variable normal d'esperança 0 i variància 1.

DEM: Donem la idea de la demostració deixant al marge els detalls fins. Expressem $C_{X_i}(\omega) = e^{\Phi(\omega)}$. Desenvolupant per Taylor, $\Phi(\omega) = a_0 + a_1\omega + a_2\omega^2 + \cdots$. Atès que $C_{X_i}(0) = 1$, ha de ser $a_0 = 0$. Amb la propietat (4) de la funció característica relacionarem els moments amb les derivades de C_{X_i}:

$$\frac{dC_{X_i}(\omega)}{d\omega} = e^{\Phi}\Phi', \qquad \frac{d^2C_{X_i}(\omega)}{d\omega^2} = e^{\Phi}(\Phi'' + \Phi'^2).$$

Substituint $\omega = 0$

$$j\mu = a_1, \qquad -E[X_i^2] = 2a_2 + a_1^2.$$

Llavors, $a_1 = j\mu$, $a_2 = -\dfrac{\sigma^2}{2}$. Així,

$$C_{X_i}(\omega) = e^{j\mu\omega - \frac{\sigma^2}{2}\omega^2 + \cdots}.$$

Ara calculem:

$$C_{Y_n}(\omega) = C_{X_1}\left(\frac{\omega}{n}\right)\cdots C_{X_n}\left(\frac{\omega}{n}\right) = e^{n(j\mu\frac{\omega}{n} - \frac{\sigma^2}{2}(\frac{\omega}{n})^2 + \cdots)} = e^{j\mu\omega - \frac{\sigma^2}{2n}\omega^2 + \cdots}$$

Els termes a l'exponent després dels dos primers decauen a zero més ràpid que n^{-1}, així que per n gran la funció característica es redueix a $e^{j\mu\omega - \frac{\sigma^2}{2n}\omega^2}$ que és la d'una variable $\mathrm{N}\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$. ♣

2.5 La variable khi quadrat

La variable khi quadrat, χ^2, es construeix a partir de n variables $N_i \sim \mathrm{N}(0,1)$, $i = 1, \ldots, n$, independents:

$$\chi^2 = \sum_{i=1}^{n} N_i^2. \tag{2.8}$$

Diem que tenim una variable de tipus khi quadrat amb n graus de llibertat. En ocasions escriurem χ^2_n.

Per la propietat (3) de la variable gamma, cada N_i^2 és $\mathrm{Gamma}\left(\dfrac{1}{2}, 2\right)$. Per la propietat (2) resulta

$$\chi^2 \sim \mathrm{Gamma}\left(\frac{n}{2}, 2\right). \tag{2.9}$$

Amb la fórmula (2.4) trobem la seva densitat:

$$f_X(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \qquad x > 0. \tag{2.10}$$

Amb la fórmula (2.6) els seu moment k-èsim val:

$$E[(\chi^2)^k] = 2^k \frac{\Gamma(k+\frac{n}{2})}{\Gamma(\frac{n}{2})} = 2^k \frac{n}{2}\left(\frac{n}{2}+1\right)\cdots\left(\frac{n}{2}+k-1\right) = n(n+2)(n+4)\cdots(n+2k-2). \quad (2.11)$$

En particular:

$$E[\chi^2] = n, \qquad V[\chi^2] = 2n. \quad (2.12)$$

Donada la definició (2.8), pel Teorema del Límit Central, quan n és gran, χ^2 és comporta com una variable normal $\mathrm{N}(n, \sqrt{2n})$.

Notem també que si $X_1, X_2, \ldots, X_n$ són variables de tipus $\mathrm{N}(\mu, \sigma)$, independents, tenim:

$$\chi^2_n = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \quad (2.13)$$

ja que cada $\dfrac{X_i - \mu}{\sigma}$ és $\mathrm{N}(0,1)$.

2.6 La variable t d'Student

Donades dues variables independents, $Y \sim \mathrm{N}(0,1)$ i $X \sim \chi^2_n$ (khi quadrat amb n graus de llibertat) la variable t d'Student amb n graus de llibertat és $T = \dfrac{Y}{\sqrt{\frac{X}{n}}}$:

$$T = \frac{N(0,1)}{\chi_n/\sqrt{n}}. \quad (2.14)$$

La seva funció de densitat és:

$$f_T(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}, \qquad -\infty < t < \infty. \quad (2.15)$$

DEM: La funció de distribució de T és:

$$F_T(t) = P(T \le t) = P\left(\frac{Y}{\sqrt{\frac{X}{n}}} \le t\right) = P\left(Y \le t\sqrt{\frac{X}{n}}\right) = \int_0^\infty dx f_X(x) \int_{-\infty}^{t\sqrt{\frac{x}{n}}} dy f_Y(y).$$

Derivant respecte a t:

$$f_T(t) = \int_0^\infty dx f_X(x) f_Y\left(t\sqrt{\frac{x}{n}}\right)\sqrt{\frac{x}{n}} = \int_0^\infty dx \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \frac{e^{-\frac{t^2 x}{2n}}}{\sqrt{2\pi}}\sqrt{\frac{x}{n}}$$

$$= \frac{1}{2\sqrt{\pi n}\Gamma(\frac{n}{2})} \int_0^\infty dx \left(\frac{x}{2}\right)^{\frac{n+1}{2}-1} e^{-(1+\frac{t^2}{n})\frac{x}{2}}$$

(canvi $z = \left(1+\dfrac{t^2}{n}\right)\dfrac{x}{2}$)

$$= \frac{1}{\sqrt{\pi n}\Gamma(\frac{n}{2})(1+\frac{t^2}{n})^{\frac{n+1}{2}}} \int_0^\infty dz z^{\frac{n+1}{2}-1} e^{-z} = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})} \frac{1}{\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}}. \clubsuit$$

Notem que:

- La densitat de la t d'Student decau en forma potencial $t^{-(n+1)}$. Aquesta caiguda lenta implica que no existeixen els moments d'ordre n o superior.

- Per $n = 1$ la densitat (2.15) és la d'una variable de Cauchy de paràmetre $\alpha = 1$. En aquest cas el denominador en T és $\chi_1 = N_1$ (veure (2.8)), una $N(0,1)$. Llavors podem interpretar la variable de Cauchy com el quocient de dues variables $N(0,1)$ independents.

- $\lim_{n\to\infty} \left(1 + \frac{t^2}{n}\right)^{\frac{n+1}{2}} = e^{\frac{t^2}{2}}$. Això mostra que per n gran la t amb n graus de llibertat tendeix a una variable N$(0,1)$.

2.7 La variable F de Fisher

Donades dues variables independents, $U \sim \chi^2_m$ (khi quadrat amb m graus de llibertat) i $V \sim \chi^2_n$ (khi quadrat amb n graus de llibertat) la variable F de Fisher amb m i n graus de llibertat és $F = \frac{\frac{U}{m}}{\frac{V}{n}}$:

$$F = \frac{n}{m}\frac{\chi^2_m}{\chi^2_n}. \tag{2.16}$$

La seva funció de densitat és:

$$f_F(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{m+n}{2}}}, \qquad x > 0. \tag{2.17}$$

DEM: La funció de distribució de F és:

$$F_F(x) = P(F \le x) = P\left(\frac{n}{m}\frac{U}{V} \le x\right) = P\left(U \le x\frac{m}{n}V\right) = \int_0^\infty dv f_V(v) \int_0^{x\frac{m}{n}v} du f_U(u).$$

Derivant respecte a x:

$$f_F(x) = \int_0^\infty dv f_V(v) f_U\left(x\frac{m}{n}v\right)\frac{m}{n}v = \int_0^\infty dv \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} v^{\frac{n}{2}-1} e^{-\frac{v}{2}} \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} \left(x\frac{m}{n}v\right)^{\frac{m}{2}-1} e^{-x\frac{m}{n}\frac{v}{2}} \frac{m}{n}v$$

$$= \frac{x^{\frac{m}{2}-1}}{2\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} \int_0^\infty dv \left(\frac{v}{2}\right)^{\frac{m+n}{2}-1} e^{-\left(1+\frac{m}{n}x\right)\frac{v}{2}}$$

(canvi $t = \left(1 + \frac{m}{n}x\right)\frac{v}{2}$, etc.)

$$= \frac{x^{\frac{m}{2}-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{\Gamma(\frac{m+n}{2})}{(1+\frac{m}{n}x)^{\frac{m+n}{2}}} = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} \frac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{m+n}{2}}}. \clubsuit$$

Els seus moments valen

$$E[F^k] = \left(\frac{n}{m}\right)^k \frac{\Gamma(\frac{m}{2}+k)\Gamma(\frac{n}{2}-k)}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} = \left(\frac{n}{m}\right)^k \frac{m(m+2)(m+4)\cdots(m+2k-2)}{(n-2)(n-4)\cdots(n-2k)} \tag{2.18}$$

(són finits per $k < 2n$.)

DEM: $E[F^k] = E\left[\left(\frac{n}{m}\frac{\chi^2_m}{\chi^2_n}\right)^k\right] = \left(\frac{n}{m}\right)^k E[(\chi^2_m)^k]E[(\chi^2_n)^{-k}] = \left(\frac{n}{m}\right)^k 2^k \frac{\Gamma(k+\frac{m}{2})}{\Gamma(\frac{m}{2})} 2^{-k} \frac{\Gamma(-k+\frac{n}{2})}{\Gamma(\frac{n}{2})}$
$= \left(\frac{n}{m}\right)^k \frac{\Gamma(k+\frac{m}{2})}{\Gamma(\frac{m}{2})}\frac{\Gamma(-k+\frac{n}{2})}{\Gamma(\frac{n}{2})}$, utilitzant la independència de les khi quadrat i la fórmula (2.11). $\clubsuit$

En particular

$$E[F] = \frac{n}{n-2}, \qquad V[F] = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}. \tag{2.19}$$

2.8 Variables aleatòries en R

El programa R incorpora les principals variables aleatòries i facilita l'avaluació numèrica de les seves propietats. Cada variable té un nom assignat. Les diferents comandes es fan afegint una lletra al principi del nom de la variable. Els arguments són el propi de la comanda i els propis de la variable.

Tipus de variable	**Nom en R**
Binomial	`binom`
Geomètrica	`geom`
Poisson	`pois`
Uniforme	`unif`
Exponencial	`exp`

Tipus de variable	**Nom en R**
Gamma	`gamma`
Normal	`norm`
χ^2	`chisq`
t d'Student	`t`
F de Fisher	`f`

Comanda	**Lletra a afegir**
Generació de valors	`r`
Funció de densitat/Funció de probabilitat	`d`
Funció de distribució	`p`
Funció de quantils (inversa de la distribució)	`q`

Exemple 2.2 Generem una mostra amb 50 valors d'una variable uniforme en l'interval $(0, 3)$:

```
> n<-50
> x<-runif(n,0,3)
> x
```

```
 [1] 2.38309980 0.74861131 2.89682274 0.60481460 2.01743236 0.02084415
 [7] 1.18053551 2.92315242 1.97499782 1.20123443 1.81997422 0.67652020
[13] 1.59074254 0.84658351 1.00487482 1.36487412 1.66956933 2.39416931
[19] 1.27730177 0.50178552 1.71505815 1.56214194 0.13479990 1.38078812
[25] 2.43185411 2.76256236 1.47029883 0.41934069 0.60633500 1.70006436
[31] 0.97390668 0.64311903 0.12160505 2.63463491 2.87228253 2.12644035
[37] 2.95597755 2.42482648 0.54006501 1.28774907 0.60741486 2.45704248
[43] 2.99175715 2.01196351 2.14646363 2.34612276 0.74035371 2.56189058
[49] 2.90210715 2.67455176
```

♦

Exemple 2.3 Calculem la probabilitat que una variable normal amb $\mu = 2$ i $\sigma = 1.5$ es trobi entre 1 i 5. És a dir, per $X \sim \mathrm{N}(1, 1.5)$ calcular $P(1 < X < 5)$. El resultat és $F_X(5) - F_X(1)$ on F_X és la funció de distribució de X.

```
> pnorm(5,2,1.5)-pnorm(1,2,1.5)
[1] 0.7247573
```

♦

Exemple 2.4 X és χ^2 amb 10 graus de llibertat. Volem el valor x tal que $P(X > x) = 0.01$. Això implica $F_X(x) = 0.99$ amb el que x és el corresponent quantil.

```
> qchisq(0.99,10)
[1] 23.20925
```

♦

Exemple 2.5 La t d'Student amb n graus de llibertat tendeix a una $\mathrm{N}(0, 1)$ quan $n \to \infty$. Comparem la densitat de les dues variables en $x = 1$ per a diferents valors de n.

```
> # Densitat de la N(0,1) en x=1 (per defecte, els parametres de
> # la normal són 0,1).
> dnorm(1)
[1] 0.2419707
> # Densitat de la t amb n=1 en x=1
> dt(1,1)
[1] 0.1591549
> # Densitat de la t amb n=5 en x=1
> dt(1,5)
[1] 0.2196798
> # Densitat de la t amb n=20 en x=1
> dt(1,20)
[1] 0.2360456
> # Densitat de la t amb n=100 en x=1
> dt(1,100)
[1] 0.2407659
> # Densitat de la t amb n=1000 en x=1
> dt(1,1000)
[1] 0.2418498
> # Densitat de la t amb n=10000 en x=1
> dt(1,10000)
[1] 0.2419586
```

♦

Exemple 2.6 Dibuixem la densitat de la χ^2 amb 5 graus de llibertat.

```
> # Farem la gràfica per 0<x<20. Generem una seqüència de 100 valors
> # en aquest interval
> x<-seq(from=0,to=20,length=100)
> # Ara generem la seqüència dels valors de la densitat en aquests punts
> f<-dchisq(x,5)
> # La gràfica:
> plot(x,f,type="l",main="Densitat de la khi quadrat amb 5 graus de libertat")
```

La sortida es mostra a la figura 2.1. ♦

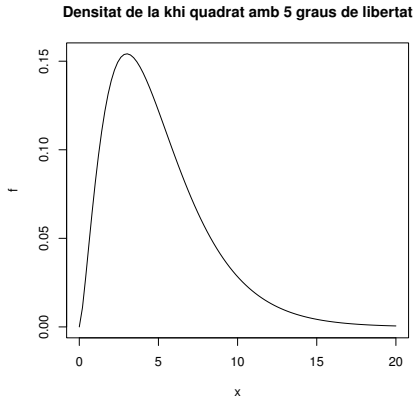

Figura 2.1: Densitat de la variable χ^2_5.

Capítol 3

Poblacions i mostres

3.1 Població

Una població és un fragment de la realitat format per un conjunt de molts elements (individus), les característiques dels quals admeten una descripció probabilística. La població pot ser finita (mida N) o infinita. Les característiques dels individus de la població es representen per variables numèriques. El fet de tenir un model probabilístic vol dir que aquestes variables són aleatòries.

En la pràctica utilitzem de manera ambigua el terme "població" per referir-nos tant al conjunt total d'individus que trobem en la realitat com al model teòric que els descriu.

Per exemple, si la característica d'interès és l'alçada de les persones de certa regió, la variable X és aquesta alçada. Llavors podem tenir un model que digui que X és normal amb certs valors de μ i σ però també tenim el conjunt d'alçades $a_1, a_2, \ldots, a_N$, corresponents a les persones existents. Quan triem una persona a l'atzar i obtenim la seva alçada podem contemplar-ho com la tria d'un dels nombres a_i o com l'obtenció d'un valor d'una variable $X \sim \mathrm{N}(\mu, \sigma)$. El valor mitjà poblacional és, per un costat la constant μ del model i per un altre la mitjana $\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} a_i$. Si N és gran els dos han de coincidir.

Els objectius de l'Estadística són: trobar el model probabilístic que descriu una població donada, determinar els paràmetres propis d'un model teòric que expliqui aquesta població i verificar si aquests models proposats són correctes.

3.2 Mostra

Una mostra d'una població és un subconjunt finit d'individus de la població, triats a l'atzar. El nombre d'elements en la mostra és la seva mida, n. La mostra és la informació que coneixem ja que sol resultar impossible examinar tota la població. La mostra és un conjunt de valors de la variable que estiguem considerant: $X_1, X_2, \ldots, X_n$. També anomenem *dades* a aquest valors.

Suposarem que les dades són variables independents i totes segueixen la distribució que modela la població. En aquest cas diem que les dades són **iid** (independents, idènticament distribuïdes). Notem que al triar elements diferents d'una població finita es perd la independència. Suposarem que N és prou gran per a que aquest efecte no es noti, o que fem el *mostreig amb reemplaçament*. Es possible calcular alguns d'aquests efectes i disposar de fórmules pel que s'anomena *mostreig sense reemplaçament* però aquí no les considerarem.

Si hi ha més d'una variable associada a cada individu, la mostra és una col·lecció de valors d'una variable multidimensional. Per exemple, amb dues variables seria: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

L'Estadística fa deduccions sobre les propietats globals de la població a partir de mostres. Com veurem, la qualitat d'aquestes deduccions és millor com més gran sigui n, encara que per motius pràctics no sempre podrem triar n tant gran com vulguem.

Exemple 3.1 La població està formada per la producció de xips de memòria de cert fabricant. Alguns xips produïts són defectuosos de manera que la variable d'interès és un indicador X que val 1 pels xips defectuosos i 0 pels correctes. El model aleatori és redueix al valor del paràmetre p de la variable X, és a dir la probabilitat que un xip triat a l'atzar sigui defectuós o la proporció de xips defectuosos en la producció total.

La mida de la població és un valor N gran, de l'ordre de milions. Una mostra de mida $n = 50$ consisteix en triar 50 xips a l'atzar dins de la producció i observar quants d'ells són defectuosos. Els mètodes de l'estadística permeten estimar el valor de p a partir del que s'observa a la mostra. ♦

Exemple 3.2 Disposem d'un dau. El model probabilístic consisteix en donar les probabilitats d'obtenir cada cara del dau. En els problemes de probabilitat bàsica es consideren daus ideals pels que cada cara té probabilitat $\frac{1}{6}$ de sortir. Ara tenim un dau físic on volem comprovar si aquestes probabilitats són iguals, per controlar la qualitat de fabricació de daus o perquè sospitem que podria estar carregat.

En aquest exemple, la població són les infinites tirades del dau. Ara $N = \infty$. Una mostra consisteix en tirar el dau, per exemple, $n = 100$ vegades. Les dades estadístiques són els 100 resultats obtinguts en aquestes tirades.

Podem plantejar el problema com el càlcul empíric de les probabilitats d'obtenir cada cara, però en estadística també és habitual plantejar-ho com el test d'una hipòtesi sobre la distribució de probabilitat de la població. En el nostre cas, la hipòtesi seria que totes les cares tenen la mateixa probabilitat i hem de veure si els resultats obtinguts són consistents amb la hipòtesi o, pel contrari, indiquen que aquesta és falsa. ♦

Exemple 3.3 La població són tots els estudiants del Campus Nord (alguns milers). Ens interessa estudiar el grau de satisfacció amb els restaurants del campus. Per això faríem una enquesta a una mostra de 200 estudiants. Com el nivell de satisfacció és una propietat més complexa demanaríem que valoressin de 1 a 5 aquest grau de satisfacció. Així tenim un resultat numèric que ens posa en el terreny de les variables aleatòries.

Algunes propietats no es poden reduir a variables aleatòries. Per exemple, a la vora d'unes eleccions polítiques es fan enquestes d'intenció de vot. Llavors cada dada és un partit polític, cosa que no podem situar simplement sobre una escala numèrica. En qualsevol cas segueix havent un model probabilístic on els paràmetres són la probabilitat que té una persona triada a l'atzar de votar per cadascun dels partits. ♦

Exemple 3.4 Un proveïdor d'accés a internet ens assegura una velocitat mitjana de 100 Mb/s. Per a verificar que això és cert prenem una mostra mirant la velocitat de connexió en 20 instants diferents. A partir d'aquesta mostra hauríem de comprovar si és certa la hipòtesi que el valor mitjà és el que assegura el proveïdor, o pel contrari aquest valor mitjà és menor. ♦

3.3 Procediments descriptius

Quan es té una mostra, el primer pas en Estadística és visualitzar les dades. Això ens dóna una primera idea sobre el tipus de distribució que tenim.

3.3.1 Histograma

Es divideix la recta real en intervals de mida fixada i es dibuixa una barra sobre cada interval, amb alçada igual al nombre de dades que hi cauen a dins.

Exemple 3.5 Generem en R una mostra de 500 valors d'una variable exponencial de paràmetre $\lambda = \frac{1}{2}$. La mostra es guarda en la variable `x` i es genera un histograma a partir d'aquesta variable. R tria una mida d'interval adequada:

```
> x<-rexp(100,0.5)
> hist(x)
```

La sortida es mostra a la figura 3.1.

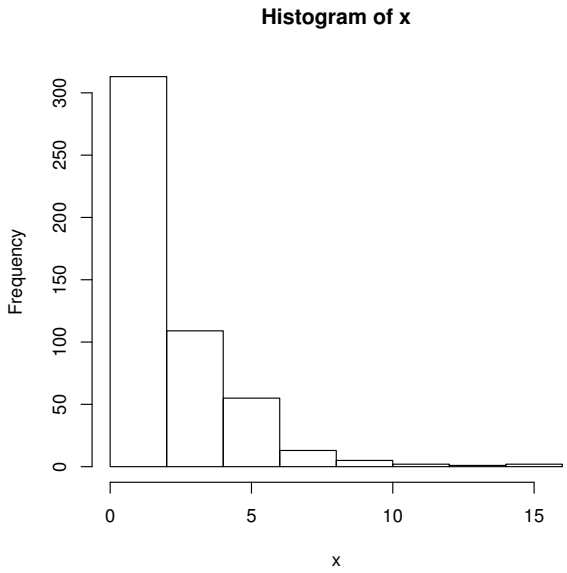

Figura 3.1: Histograma d'una mostra exponencial.

Ho fem ara amb una variable $\mathrm{N}(3, 1)$.

```
> y<-rnorm(500,3,1)
> hist(y)
```

La sortida es mostra a la figura 3.2. ♦

3.3.2 Gràfic de caixa (boxplot)

En la recta real es marca la posició dels tres quartils mostrals (Q_1 s'obté amb un valor que té a l'esquerra un 25% de les dades, etc). Gràficament es dibuixa una caixa que comença en Q_1 i acaba en Q_3. Una línia dins la caixa marca la posició de Q_2 (mediana). A més es dibuixen una línia a cada costat de la caixa fins als valors mostrals més extrems entre $Q_1 - 1.5 \cdot \mathrm{IQR}$ i $Q_3 + 1.5 \cdot \mathrm{IQR}$. Les dades que queden fora de d'aquest interval es dibuixen explícitament i s'anomenen *dades atípiques* (*outliers*). Aquests valors són molt poc probables i poden representar algun error present quan s'ha pres la dada corresponent. A vegades es suprimeixen de la mostra.

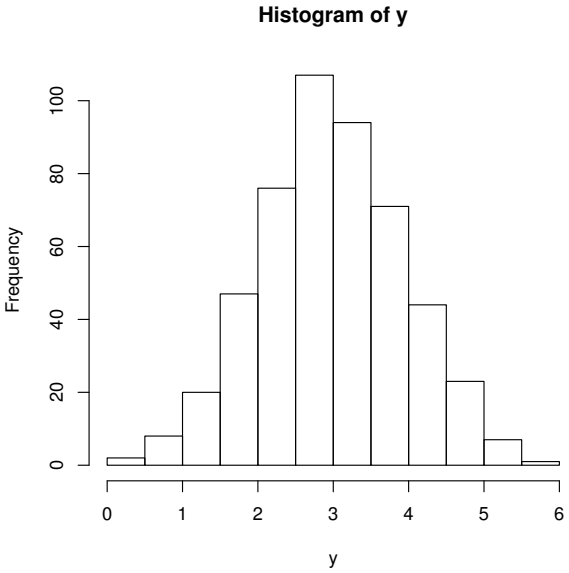

Figura 3.2: Histograma d'una mostra normal.

Exemple 3.6 Generem una mostra amb $n = 200$ d'una variable $\mathrm{N}(2, 1)$.

```
> x<-rnorm(200,2,1)
> boxplot(x)
```

La sortida es mostra a la figura 3.3. ♦

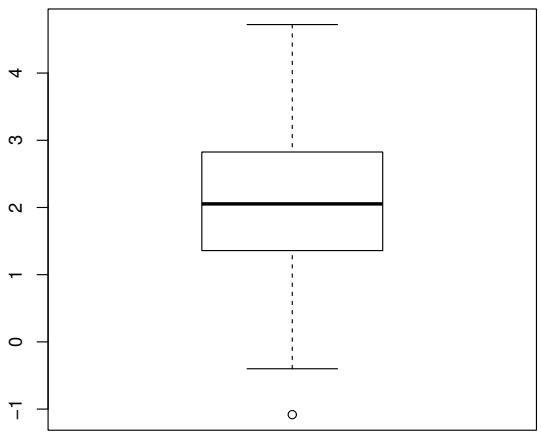

Figura 3.3: Boxplot.

Els boxplots són útils quan volem comparar la distribució de dues poblacions diferents. El boxplot visualitza de forma ràpida les principals magnituds.

3.3.3 Diagrama de dispersió (scatter plot)

Amb dades bidimensionals podem representar al pla el conjunt de punts obtinguts. Aquesta gràfica visualitza el grau de correlació entre les dues variables.

Exemple 3.7 Generem mostres normals amb $n = 50$. Notem que `x` i `y` són independents fet que es manifesta en la falta d'estructura en la corresponent gràfica. Pel contrari `x` i `z` estan correlades i el corresponent núvol mostra un cert pendent positiu.

```
> x<-rnorm(50,3,1)
> y<-rnorm(50,1,2)
> z<-2*x-y
> plot(x,y)
> plot(x,z)
```

La sortida es mostra a la figura 3.4. ♦

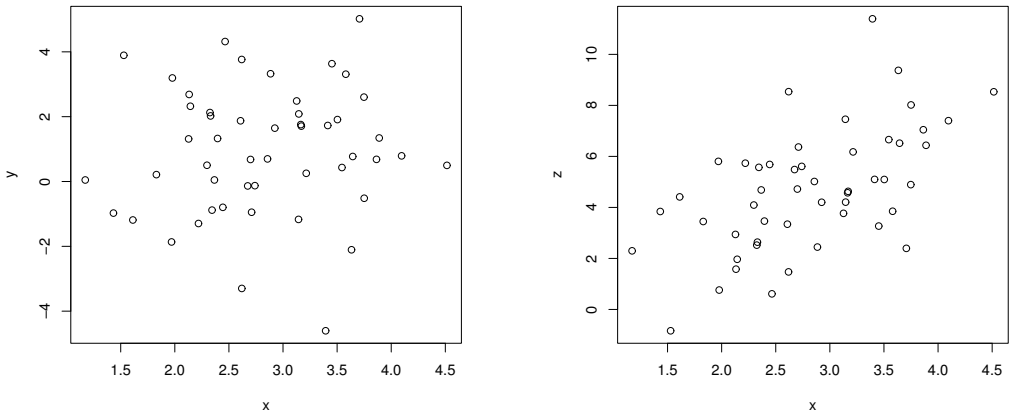

Figura 3.4: Scatter plots.

3.4 Estadístics

Un estadístic és una magnitud que es calcula a partir dels valors de la mostra. Els estadístics són funcions de les variables $X_1, X_2, \ldots, X_n$. Aquestes funcions es trien de manera que el valor que donen contingui informació significativa sobre la població. En la base d'aquesta idea està la llei dels grans nombres i el fet que l'esperança d'una variable X es correspon amb el límit de $\frac{1}{n}(X_1+\cdots+X_n)$ quan $n \to \infty$. Com els valors de la mostra són variables aleatòries, l'estadístic és una nova variable aleatòria. Les variables X_i són independents i segueixen totes la distribució poblacional. Llavors un estadístic és una nova variable funció de les X_i: $Z = \varphi(X_1, X_2, \ldots, X_n)$

Donat un estadístic voldrem conèixer els seus paràmetres i la seva distribució de probabilitat:

- El seu valor mitjà $E[Z] = E[\varphi(X_1, X_2, \ldots, X_n)]$ és el valor a on "apunta" l'estadístic Z. Normalment serà un paràmetre important de la població.
- La seva variància $V[Z] = V[\varphi(X_1, X_2, \ldots, X_n)]$ ens dóna la seva dispersió i, per tant, és una mesura de l'error quan prenem Z com a valor experimental de $E[Z]$.

- La seva distribució de probabilitat permet controlar els errors amb precisió, en el sentit de poder calcular les probabilitats que Z difereixi de $E[Z]$ una quantitat donada qualsevol.

A continuació definim els dos estadístics més importants i calculem els seus paràmetres. També discutim la seva distribució de probabilitat. Aquests resultats s'utilitzaran i s'ampliaran en capítols posteriors.

3.5 Mitjana mostral

La mitjana mostral és

$$\widehat{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{3.1}$$

Si la població té esperança μ i desviació σ:

$$E[\widehat{X}] = \mu. \tag{3.2}$$

$$V[\widehat{X}] = \frac{\sigma^2}{n}. \tag{3.3}$$

DEM: Utilitzant la linealitat de l'esperança i el fet que $E[X_i] = \mu$, $\forall i$:

$$E[\widehat{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\cdot n\mu = \mu.$$

Com els valors de la mostra són independents, tenim que la variància de la suma és suma de variàncies. A més, $V[X_i] = \sigma^2$, $\forall i$:

$$V[\widehat{X}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}V\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. \clubsuit$$

Sobre el tipus de variable que és $\widehat{X}$, tenim:

- Si la població és normal (és a dir $X_i \sim \mathrm{N}(\mu, \sigma)$), $\widehat{X}$ és una combinació lineal de les X_i i per tant també és normal: $\widehat{X} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
- Si la població no és normal no podem dir res en general, però si n és gran el Teorema del Límit Central ens diu que $\widehat{X}$ es comporta com una variable normal, per tant $\widehat{X} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ per n gran. En la pràctica és considera n gran quan $n > 30$ (mostres grans). En cas contrari es parla de mostres petites.

Exemple 3.8

```
> # Generem una mostra de mida 10 d'una població normal amb mitjana 2
> # i desviació 1
> x<-rnorm(10,2,1)
> x
 [1] 1.167175 4.776473 2.569489 1.315357 3.843460 1.317970 2.677908 3.433255
 [9] 2.217958 1.162673
> # Calculem la seva mitjana mostral
> m<-mean(x)
> m
```

```
[1] 2.448172
> # Calculem la probabilitat que la mitjana mostral hagués estat més
> # gran que 2.5
> 1-pnorm(2.5,2,1/sqrt(10))
[1] 0.05692315
```

En aquest exemple la mostra era petita però la població era normal i hem pogut utilitzar la distribució normal per calcular la probabilitat demanada. ♦

Exemple 3.9

```
> # Generem una mostra de mida 50 d'una població uniforme en l'interval (0,2)
> # Notem que la mitjana de la població és 1 i la desviació 1/sqrt(3)
> x<-runif(50,0,2)
> x
 [1] 0.99386261 0.29061111 0.02002708 0.46850911 1.44413622 1.19664701
 [7] 1.38891852 0.24445703 1.85100703 0.68878492 1.57925576 1.33059741
[13] 1.53233081 1.21844994 0.82486995 0.08133076 0.86009917 1.84328973
[19] 1.15549972 0.79231910 0.59965196 1.42794773 1.60942037 1.25429776
[25] 1.39764902 0.11801187 1.00873980 0.49148397 1.79892327 1.22691438
[31] 0.01782683 1.03104929 0.96950695 1.09832800 0.34292595 0.72722582
[37] 0.24333880 0.64746859 1.99358323 0.12001050 1.67822195 0.37218527
[43] 1.23117716 1.65668958 0.17362957 1.86235164 0.29814011 1.06373692
[49] 0.13943438 0.73254976
> # Calculem la seva mitjana mostral
> m<-mean(x)
> m
[1] 0.9427485
> # Calculem la probabilitat que la mitjana mostral hagués estat més
> # petita que 0.8
> pnorm(0.8,1,1/sqrt(3*50))
[1] 0.007152939
```

En aquest exemple la població no era normal però la mostra era gran i hem pogut utilitzar la distribució normal per calcular la probabilitat demanada. ♦

3.6 Variància mostral

La variància d'una variable aleatòria X és $E[(X-\mu)^2]$. L'esperança la representarem com a mitjana aritmètica sobre els valors mostrals però hi ha un fet addicional que és que no coneixem el valor de μ així que el substituirem per la mitjana mostral. Això ens porta a considerar el següent estadístic:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{X})^2. \tag{3.4}$$

Si la població té esperança μ, desviació σ i curtosi κ_4:

$$E[S^2] = \frac{n-1}{n}\sigma^2. \tag{3.5}$$

$$V[S^2] = \sigma^4 \frac{n-1}{n^2}\left(\kappa_4 - 1 + \frac{3-\kappa_4}{n}\right). \tag{3.6}$$

DEM:

Definim les variables $Z_i = \dfrac{X_i - \mu}{\sigma}$, $i = 1, \ldots, n$. Aquestes variables són independents, de mitjana zero i variància 1. Com $X_i = \mu + \sigma Z_i$:

$$S^2 = \frac{\sigma^2}{n} \sum_{i=1}^{n} (Z_i - \widehat{Z})^2.$$

Ara, notem que $\displaystyle\sum_{i=1}^{n} (Z_i - \widehat{Z})^2 = \sum_{i=1}^{n} (Z_i^2 - 2Z_i\widehat{Z} + \widehat{Z}^2) = \sum_{i=1}^{n} Z_i^2 - 2n\widehat{Z}^2 + n\widehat{Z}^2 = \sum_{i=1}^{n} Z_i^2 - n\widehat{Z}^2$. També és $\widehat{Z}^2 = \dfrac{1}{n^2}\left(\displaystyle\sum_{i=1}^{n} Z_i^2 + 2\sum_{i<j}^{n} Z_i Z_j\right)$. Llavors

$$S^2 = \frac{\sigma^2}{n^2}\left((n-1)\sum_{i=1}^{n} Z_i^2 - 2\sum_{i<j}^{n} Z_i Z_j\right).$$

Ara, notem que $E[Z_i^2] = V[Z_i] = 1$, $E[Z_i Z_j] = E[Z_i]E[Z_j] = 0$, per $i \neq j$. Llavors

$$E[S^2] = \frac{\sigma^2}{n^2}\left((n-1)\sum_{i=1}^{n} E[Z_i^2] - 2\sum_{i<j}^{n} E[Z_i Z_j]\right) = \frac{\sigma^2}{n^2}(n-1)n = \sigma^2\frac{n-1}{n}.$$

Per calcular la variància de S^2 avaluarem la suma de variàncies del termes dels sumatoris ja que els termes de covariància dos a dos són sempre nuls (per exemple, si $i \neq j$, $C[Z_i^2, Z_j^2] = 0$ per independència, $C[Z_i^2, Z_j Z_k] = 0$, ja que en les esperances implicades sempre hi haurà alguna Z_i sola i $E[Z_i] = 0$, etc). També tindrem en compte que $V[Z_i^2] = E[Z_i^4] - E[Z_i^2]^2 = \kappa_4 - 1$ i $V[Z_i Z_j] = E[Z_i^2 Z_j^2] = E[Z_i^2]E[Z_j^2] = 1$.

$$V[S^2] = \frac{\sigma^4}{n^4}\left((n-1)^2\sum_{i=1}^{n} V[Z_i^2] + 4\sum_{i<j}^{n} V[Z_i Z_j]\right)$$

$$= \frac{\sigma^4}{n^4}\left((n-1)^2 n(\kappa_4 - 1) + 4\frac{n(n-1)}{2}\right) = \sigma^4\frac{n-1}{n^2}\left(\kappa_4 - 1 + \frac{3-\kappa_4}{n}\right). \clubsuit$$

Sobre el tipus de variable que és S^2:

Si la població és normal, $\dfrac{n}{\sigma^2}S^2$ és khi quadrat amb $n-1$ graus de llibertat.

DEM:

$$\frac{n}{\sigma^2}S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \widehat{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu - (\widehat{X} - \mu))^2 = \sum_{i=1}^{n}\frac{(X_i - \mu)^2}{\sigma^2} - n\frac{(\widehat{X} - \mu)^2}{\sigma^2}.$$

Llavors:

$$\sum_{i=1}^{n}\frac{(X_i - \mu)^2}{\sigma^2} = \frac{n}{\sigma^2}S^2 + \frac{(\widehat{X} - \mu)^2}{(\sigma/\sqrt{n})^2}.$$

El terme de l'esquerra és χ^2 amb n graus de llibertat. El terme $\dfrac{(\widehat{X} - \mu)^2}{(\sigma/\sqrt{n})^2}$ és χ^2 amb 1 grau de llibertat. A més, els dos sumands de la dreta són independents ja que $C[X_i - \widehat{X}, \widehat{X}] = E[(X_i - \widehat{X})\widehat{X}] = \dfrac{\sigma^2}{n} - \dfrac{\sigma^2}{n} = 0$ i al ser variables normals incorrelades, també són independents. Llavors, per la propietat (2) de la variable Gamma, $\dfrac{n}{\sigma^2}S^2$ ha de ser χ^2 amb $n-1$ graus de llibertat. $\clubsuit$

S^2 vol ser una versió mostral de la variància però hauria de tenir esperança igual a σ^2 en lloc del que senyala la fórmula 3.5. Degut a això s'utilitza com a variància mostral una versió modificada:

$$\widehat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{X})^2. \tag{3.7}$$

S^2 és l'estimador esbiaixat de la variància. $\widehat{S}^2$ és l'estimador no esbiaixat.

Si la població té esperança μ, desviació σ i curtosi κ_4:

$$E[\widehat{S}^2] = \sigma^2. \tag{3.8}$$

$$V[\widehat{S}^2] = \frac{\sigma^4}{n-1}\left(\kappa_4 - 1 + \frac{3-\kappa_4}{n}\right). \tag{3.9}$$

Si la població és normal, $\dfrac{(n-1)}{\sigma^2}\widehat{S}^2$ és khi quadrat amb $n-1$ graus de llibertat.

Com s'ha vist en la última demostració, el fet que $\widehat{X}$ i $X_i - \widehat{X}$ siguin incorrelades $\forall i$ implica que $\widehat{X}$ i $\widehat{S}$ són variables independents.

Exemple 3.10 En R, la variància mostral no esbiaixada (fórmula 3.7) s'obté amb `var()`. `sd()` dóna la desviació mostral (arrel de la variància mostral).

```
> # Generem una mostra de mida 10 d'una població normal amb mitjana 2
> # i desviació 1
> n<-10
> mu<-2
> sigma<-1
> x<-rnorm(n,mu,sigma)
> x
 [1] 2.1806683 0.3038982 2.8801761 1.9751809 1.1658977 1.5452250 0.2554891
 [8] 3.7215836 4.1787763 1.0847175
> # Calculem la seva mitjana mostral
> m<-mean(x)
> m
[1] 1.929161
> # Calculem la seva desviació mostral
> s<-sd(x)
> s
[1] 1.3387
> # Comprovem que la variància mostral és el quadrat de la desviació
> var(x)
[1] 1.792119
> s^2
[1] 1.792119
> # Calculem la versió esbiaixada de la desviació
> sqrt(n/(n-1))*s
[1] 1.411114
> # Calculem la probabilitat que la desviació mostral s'hagués trobat
> # entre 0.8 i 1.2. Això implica que 0.8^2<S^2<1.2^2. Multiplicant
> # la desigualtat per (n-1) i dividint per sigma^2 tenim que una
> # khi quadrat amb 9 graus de llibertat s'ha de trobar entre 9*0.8^2
> # i 9*1.2^2
> pchisq((n-1)*1.2^2,n-1)-pchisq((n-1)*0.8^2,n-1)
[1] 0.5992528
> # Si fem el mateix amb n=100
> n2<-100
> x2<-rnorm(n2,mu,sigma)
> x2
  [1]  1.1671747  4.7764727  2.5694891  1.3153569  3.8434598  1.3179705
  [7]  2.6779078  3.4332551  2.2179578  1.1626725  3.3616413  2.0730781
```

```
 [13]  0.4836763  1.6589754  0.7597645  1.9593215  2.7212383  1.1504056
 [19]  2.1900378  1.4157593  2.6793501  1.9390617  3.2052603  1.0860113
 [25]  1.6767099  3.4988141  1.5515720  2.7694855  1.5480920  2.8903841
 [31]  2.2469422  1.1383061  3.4111899  0.9888514  1.4377808  3.0049521
 [37]  1.6766066  0.7006759  1.9226094  3.9168314 -2.0974478  1.5927127
 [43]  1.8062696  1.7538546  2.0626961  0.1375675  0.5640650  2.3938297
 [49]  1.7550085  4.4074778  2.8526576  2.2479107  2.8713139  1.8316857
 [55]  0.4484005  1.4038809  1.0980704  2.5472478  3.5491025  4.1728302
 [61]  1.9923078 -0.3258402  2.5889963  2.5085303  3.4430976  2.8051307
 [67]  2.7262766  1.7787143  1.8237521  2.1961404  1.4750989  2.8585667
 [73]  2.5210227  2.0109539  3.2784900 -0.3691954  1.9617733  1.0516002
 [79]  0.8333180  4.7256870  2.9908106  2.2939150  0.6393666  0.9595699
 [85]  0.5220994  1.5347008  2.1147677  1.7076363  1.5096469  1.3697448
 [91]  1.0908466  2.2200410  0.1214833  2.5947809  2.6863174  1.8471319
 [97]  2.4692626  1.7490047  0.3400313  1.7446092
> m2<-mean(x2)
> m2
[1] 1.973344
> s2<-sd(x2)
> s2
[1] 1.125807
> pchisq((n2-1)*1.2^2,n2-1)-pchisq((n2-1)*0.8^2,n2-1)
[1] 0.9952305
```

♦

3.7 Altres estadístics

Pel moment k-èsim, $\mu_k = E[X^k]$, podem utilitzar $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ ja que $E[M_k] = \mu_k$ i $V[M_k] = \frac{\mu_{2k} - \mu_k^2}{n}$.

Els quantils mostrals es determinen trobant un punt que tingui a l'esquerra el percentatge corresponent de les dades. Cal que n sigui una mica gran. Per exemple, si $n = 100$, ordenem les dades i la mitjana entre les dades número 25 i número 26 ens dóna un estadístic que estima el primer quartil, etc.

També es poden considerar el mínim i el màxim de les dades.

Exemple 3.11 En R donada una mostra `x`, `summary(x)` dóna els principals estadístics. Fem-ho amb 100 valors d'una variable de Cauchy ($\alpha = 1$) i una $\mathrm{N}(0, 1)$:

```
> n<-100
> x<-rcauchy(n)
> summary(x)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-72.84000  -0.89640   0.02974   0.24560   1.19700 103.70000
> y<-rnorm(n)
> summary(y)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-2.43000 -0.73710 -0.01146 -0.02757  0.71420  2.01100
```

Notem que la mitjana mostral s'aproxima a zero de manera similar a la mediana mostral en el cas normal mentre que pel cas de Cauchy la mitjana mostral difereix de zero en comparació amb la mediana mostral. Això es deu a que la variable de Cauchy té mediana 0 però la seva esperança no està definida. ♦

Exemple 3.12 En l'exemple anterior s'observa la utilitat dels quantils ja que algunes variables no tenen moments definits. Una altra utilitat és que els quantils són més robustos davant la presència de dades anòmales. Comprovem-ho comparant la mitjana i la mediana mostrals abans i després de "pertorbar" la mostra:

```
> x<-rnorm(5)
> x
[1]  0.88676361  0.09958062 -0.35997765  0.97799721 -0.12152694
> mean(x)
[1] 0.2965674
> median(x)
[1] 0.09958062
> x1<-x+c(10,0,0,0,0)
> x1
[1] 10.88676361  0.09958062 -0.35997765  0.97799721 -0.12152694
> mean(x1)
[1] 2.296567
> median(x1)
[1] 0.09958062
```

Per a aquesta variable, $\mu = m = 0$. La mitjana mostral i la mediana mostral ens aproximen al valor de μ. La variació introduïda en la primera dada (valor molt improbable per una $\mathrm{N}(0,1)$, més fàcil que es degui a un error en la recol·lecció de dades) afecta el valor mitjà però no la mediana. ♦

Capítol 4

Teoria de l'estimació. Intervals de confiança

En aquest capítol considerem el problema de determinar el valor que té algun paràmetre característic de la població. La primera part consisteix en obtenir una estimació del valor d'aquest paràmetre a partir de les dades estadístiques. La segona part tracta de donar informació més precisa donant no només un valor estimat sinó també un interval que contingui el valor real amb certa probabilitat prefixada.

4.1 Estimadors

Considerem cert paràmetre θ propi de la població. Com les característiques dels elements de la població es representen amb variables aleatòries, aquest paràmetre és alguna propietat de la variable X que modela la població. Habitualment es tracta de l'esperança o la variància de X, però també hi altres casos; per exemple, si X és una variable uniforme podríem voler conèixer els extrems del corresponent interval.

Un estimador és un estadístic, una funció de les dades $\widehat{\theta}(X_1, X_2, \ldots, X_n)$ tal que els seus valors són bones aproximacions del valor de θ. Notem que $\widehat{\theta}$ és una variable aleatòria. Perquè aquesta variable aproximi la constant θ són desitjables les següents propietats:

Estimador no esbiaixat (o centrat):

$$E[\widehat{\theta}] = \theta. \tag{4.1}$$

En general, es defineix el **biaix** de l'estimador com:

$$B_{\widehat{\theta}}(\theta) = E[\widehat{\theta}] - \theta. \tag{4.2}$$

Estimador consistent:

$$\lim_{n\to\infty} E[(\widehat{\theta} - \theta)^2] = 0. \tag{4.3}$$

Notem que $E[(\widehat{\theta} - \theta)^2] = E[(\widehat{\theta} - E[\widehat{\theta}] + E[\widehat{\theta}] - \theta)^2] = E[(\widehat{\theta} - E[\widehat{\theta}])^2] + (E[\widehat{\theta}] - \theta)^2 = V[\widehat{\theta}] + B^2_{\widehat{\theta}}$. Com els dos termes són positius, (4.3) equival a $\lim_{n\to\infty} V[\widehat{\theta}] = 0$ i $\lim_{n\to\infty} B_{\widehat{\theta}} = 0$

Per estimadors centrats la condició de consistència és, per tant, $\lim_{n\to\infty} V[\widehat{\theta}] = 0$. La variància de l'estimador dóna una mesura de l'error mitjà que cometem al prendre $\widehat{\theta}$ com a valor de θ.

(4.3) implica que $\widehat{\theta}$ tendeix a θ en probabilitat. És a dir, $\forall \epsilon > 0$, $\lim_{n\to\infty} P(|\widehat{\theta} - \theta| < \epsilon) = 1$.

DEM: Igual que en la demostració de la desigualtat de Txebixev, trobem $E[(\widehat{\theta}-\theta)^2] \geq \epsilon^2 P(|\widehat{\theta}-\theta| \geq \epsilon)$ d'on

$$P(|\widehat{\theta}-\theta| < \epsilon) = 1 - P(|\widehat{\theta}-\theta| \geq \epsilon) \geq 1 - \frac{E[(\widehat{\theta}-\theta)^2]}{\epsilon^2}.$$

Quan $n \to \infty$ la fita de la dreta tendeix a 1. ♣

Donats dos estimadors per al mateix paràmetre diem que és més **eficient** el que té menor valor de $E[(\widehat{\theta}-\theta)^2]$. En el cas d'estimadors centrats, és el que té menor variància.

4.2 Estimadors per al valor mitjà i per a la variància

Com es dedueix de les fórmules (3.2) i (3.3), la mitjana mostral $\widehat{X}$ és un estimador no esbiaixat i consistent de l'esperança μ de la població.

Com es dedueix de les fórmules (3.8) i (3.9), la variància mostral no esbiaixada $\widehat{S}^2$ és un estimador no esbiaixat i consistent de la variància σ^2 de la població. Notem que la primera versió, S^2, és consistent però és esbiaixada. Cambiant el denominador n per $n-1$ obteníem la versió no esbiaixada, $\widehat{S}^2$. Per n gran la diferència entre les dues versions és petita.

Com estimació de la desviació de la població, σ, prendrem l'arrel de la variància mostral, $\widehat{S}$.

Per estimar paràmetres més generals tenim dos mètodes alternatius: el mètode dels moments i el mètode de màxima versemblança. En general produeixen estimadors diferents encara que en ocasions coincideixen.

4.3 Mètode dels moments

Si la distribució de la variable poblacional conté k paràmetres desconeguts, determinem la estimació d'aquests paràmetres a partir de les k equacions que s'obtenen igualant els k primers moments de la població amb els moments mostrals. Per a $l = 1, 2, \ldots, k$:

$$E[X^l] = \frac{1}{n}\sum_{i=1}^{n} X_i^l. \tag{4.4}$$

Exemple 4.1 Donada una població on $X \sim \text{Exp}(\lambda)$, estimar el paràmetre λ. Només cal una equació i recordar que $E[X] = \dfrac{1}{\lambda}$:

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Llavors el nostre estimador de λ és $\dfrac{1}{\widehat{X}}$. Resultat que es podia esperar, ja que $\lambda = \dfrac{1}{\mu}$ i la millor estimació de μ és $\widehat{X}$. ♦

Exemple 4.2 Donada una població on $X \sim \text{Uniforme}(a, b)$, estimar els paràmetres a, b.

$$\begin{cases} E[X] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i, \\ E[X^2] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2. \end{cases}$$

Restant-li a la segona equació el quadrat de la primera:

$$\begin{cases} E[X] = \widehat{X}, \\ V[X] = S^2. \end{cases}$$

És a dir:

$$\begin{cases} \dfrac{a+b}{2} = \widehat{X}, \\ \dfrac{(b-a)^2}{12} = S^2. \end{cases}$$

La solució és $a = \widehat{X} - \sqrt{3}S$ i $b = \widehat{X} + \sqrt{3}S$. ♦

4.4 Mètode de la màxima versemblança

Com abans, considerem que la variable poblacional té una densitat que depèn de k paràmetres desconeguts $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$: $f(x, \theta)$. Quan prenem una mostra $X_1, X_2, \ldots, X_n$, el resultat és una variable n-dimensional amb densitat

$$f(x_1, x_2, \ldots, x_n) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta). \tag{4.5}$$

El mètode de la màxima versemblança es basa en el fet que els valors obtinguts en la mostra han de tenir una probabilitat alta d'aparèixer. Llavors, determinarem $\theta_1, \theta_2, \ldots, \theta_k$ demanant que la funció $f(x_1, x_2, \ldots, x_n)$, on les x_i corresponen als valors obtinguts a la mostra, sigui màxima respecte a aquests paràmetres.

El màxim es determina generalment imposant que les derivades respecte a cada paràmetres s'anul·lin:

$$\frac{\partial}{\partial \theta_l} f(x_1, x_2, \ldots, x_n) = 0, \qquad l = 1, 2, \ldots, k. \tag{4.6}$$

Donada la forma (4.5) és més pràctic treballar amb la funció:

$$L = \ln f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i, \theta). \tag{4.7}$$

Com el logaritme és una funció monòtona, el màxim de f el trobem també anul·lant les derivades de L:

$$\frac{\partial}{\partial \theta_l} L = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_l} \ln f(x_i, \theta) = 0, \qquad l = 1, 2, \ldots, k. \tag{4.8}$$

Exemple 4.3 Considerem de nou el cas de l'exemple 4.1. Tenim $f(x, \lambda) = \lambda e^{-\lambda x}$.

$$L = \sum_{i=1}^{n} \ln(\lambda e^{-\lambda x_i}) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i,$$

$$\frac{\partial}{\partial \lambda} L = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0,$$

d'on $\lambda = \dfrac{1}{\widehat{X}}$, resultat que coincideix amb el del mètode dels moments. ♦

Exemple 4.4 Considerem de nou el cas de l'exemple 4.2. Tenim $f(x, a, b) = \frac{1}{b-a}$, $a < x < b$.

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(b-a)^n}, \qquad a < x_1, x_2, \ldots, x_n < b.$$

En aquest cas el màxim no el trobem derivant sinó observant que $b - a$ hauria de ser mínim. Com els valors x_i es troben dins de l'interval, hem de prendre $a = \min(x_1, x_2, \ldots, x_n)$ i $b = \max(x_1, x_2, \ldots, x_n)$. El resultat ha estat diferent que amb el mètode dels moments.

Comparem les estimacions dels dos mètodes en R. Treballarem amb una mostra de 100 valors uniformes en $(0, 1)$.

```
> # Generem una mostra de mida 100 d'una població uniforme en (0,1)
> n<-100
> x<-runif(n,0,1)
> x
  [1] 0.794366601 0.249537104 0.965607580 0.201604866 0.672477453 0.006948049
  [7] 0.393511838 0.974384139 0.658332608 0.400411478 0.606658072 0.225506735
 [13] 0.530247513 0.282194505 0.334958273 0.454958041 0.556523109 0.798056436
 [19] 0.425767257 0.167261839 0.571686050 0.520713978 0.044933298 0.460262707
 [25] 0.810618038 0.920854119 0.490099611 0.139780229 0.202111666 0.566688120
 [31] 0.324635559 0.214373011 0.040535015 0.878211638 0.957427510 0.708813448
 [37] 0.985325851 0.808275494 0.180021670 0.429249691 0.202471619 0.819014160
 [43] 0.997252384 0.670654502 0.715487878 0.782040921 0.246784570 0.853963528
 [49] 0.967369049 0.891517254 0.247610142 0.649046289 0.751084918 0.876208848
 [55] 0.924107527 0.599459686 0.586269009 0.637726795 0.201204257 0.203626079
 [61] 0.913344438 0.231121304 0.529128030 0.006477683 0.064718766 0.701575590
 [67] 0.366542549 0.095309511 0.107444154 0.079870506 0.483776086 0.437357777
 [73] 0.764618549 0.932305347 0.197775305 0.775394852 0.575360233 0.679456696
 [79] 0.279529197 0.905328171 0.751541973 0.858447138 0.475704177 0.250473954
 [85] 0.885948595 0.185388057 0.180361393 0.021855230 0.373237765 0.302902663
 [91] 0.933039076 0.328497919 0.326922172 0.466812163 0.779197415 0.261723089
 [97] 0.325667643 0.673710683 0.813370168 0.074034197
> # Calculem la seva mitjana mostral
> m<-mean(x)
> m
[1] 0.511721
> # Calculem la seva desviació esbiaixada
> s<-sqrt((n-1)/n)*sd(x)
> s
[1] 0.2929871
> # Segons el mètode dels moments la millor estimació de a i b és
> am<-m-sqrt(3)*s
> am
[1] 0.004252499
> bm<-m+sqrt(3)*s
> bm
[1] 1.019189
> # Segons el mètode de màxima versemblança la millor estimació de a i b és
> av<-min(x)
> av
[1] 0.006477683
> bv<-max(x)
> bv
[1] 0.9972524
```

Els dos mètodes donen bones aproximacions als valors reals de a i b. En aquest exemple no s'observa cap diferència apreciable en favor d'un d'ells. ♦

4.5 Informació de Fisher i fita de Cramér-Rao

En aquesta secció suposarem sense fer-les explícites que les densitats verifiquen condicions que permeten intercanviar la integració amb la derivació respecte als paràmetres. Considerem una

variable X amb densitat $f(x,\theta)$ on θ és cert paràmetre. $X_1, X_2, \ldots, X_n$ és una mostra d'aquesta variable, és a dir, n variables independents totes amb la mateixa distribució que X.

Ja hem definit la funció $L(x_1, x_2, \ldots, x_n; \theta) = \ln f(x_1, x_2, \ldots x_n; \theta)$ on $f(x_1, x_2, \ldots x_n; \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$. Considerem la funció $\dfrac{\partial L}{\partial \theta}$. En tant que funció de les variables $X_1, X_2, \ldots, X_n$, $\dfrac{\partial L}{\partial \theta}$ és una variable aleatòria que verifica la següent propietat: per a tota funció $G = g(X_1, X_2, \ldots, X_n; \theta)$:

$$E\left[G\frac{\partial L}{\partial \theta}\right] = \frac{\partial}{\partial \theta} E[G] - E\left[\frac{\partial G}{\partial \theta}\right]. \tag{4.9}$$

DEM: Denotant $f(x;\theta) = f(x_1, x_2, \ldots, x_n; \theta)$, $\int d^n x = \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_n$, etc:

$$\frac{\partial}{\partial \theta} E[G] = \frac{\partial}{\partial \theta} \int d^n x \; g(x;\theta) f(x;\theta) = \int d^n x \; \frac{\partial g}{\partial \theta}(x;\theta) f(x;\theta) + \int d^n x \; g(x;\theta) \frac{\partial f}{\partial \theta}(x;\theta)$$

$$= E\left[\frac{\partial G}{\partial \theta}\right] + \int d^n x \; g(x;\theta) \frac{1}{f(x;\theta)} \frac{\partial f}{\partial \theta}(x;\theta) f(x;\theta)$$

$$= E\left[\frac{\partial G}{\partial \theta}\right] + \int d^n x \; g(x;\theta) \frac{\partial L}{\partial \theta}(x;\theta) f(x;\theta) = E\left[\frac{\partial G}{\partial \theta}\right] + E\left[G\frac{\partial L}{\partial \theta}\right]. \clubsuit$$

Ara, considerant el cas $G = 1$ obtenim:

$$E\left[\frac{\partial L}{\partial \theta}\right] = 0. \tag{4.10}$$

Prenent $G = \dfrac{\partial L}{\partial \theta}$ obtenim:

$$E\left[\left(\frac{\partial L}{\partial \theta}\right)^2\right] = E\left[-\frac{\partial^2 L}{\partial \theta^2}\right]. \tag{4.11}$$

Aquesta derivada segona ens dóna la curvatura de L. Com més concentrada sigui la probabilitat, més gran serà la curvatura en el màxim. Així prenem aquest valor com una mesura de la incertesa i definim la **informació de Fisher**:

$$I_F = E\left[\left(\frac{\partial L}{\partial \theta}\right)^2\right]. \tag{4.12}$$

Utilitzant (4.10) tenim que:

$$V\left[\frac{\partial L}{\partial \theta}\right] = I_F. \tag{4.13}$$

Notem que I_F és funció de θ i de n.

Si G és independent de θ,

$$V[G] \geq \frac{\left(\frac{\partial}{\partial \theta} E[G]\right)^2}{I_F} \tag{4.14}$$

DEM: (4.9) i (4.10) ens donen $C\left[G, \dfrac{\partial L}{\partial \theta}\right] = E\left[G\dfrac{\partial L}{\partial \theta}\right] = \dfrac{\partial}{\partial \theta} E[G]$.

La desigualtat de Cauchy-Schwartz és $C[A,B]^2 \leq V[A]V[B]$. Per tant, $\left(\dfrac{\partial}{\partial \theta} E[G]\right)^2 \leq V[G] I_F$ d'on s'obté la desigualtat (4.14). $\clubsuit$

Ara, donat un estimador $\widehat{\theta}$ per al paràmetre θ, es verifica la **fita de Cramér-Rao**:

$$V[\widehat{\theta}] \geq \frac{(1+B'_{\widehat{\theta}})^2}{I_F(\theta)}. \tag{4.15}$$

DEM: Apliquem (4.14) i tenim en compte que, de l'equació (4.2), $E[\widehat{\theta}] = \theta + B_{\widehat{\theta}}(\theta)$. ♣

Pel cas d'estimadors centrats:

$$V[\widehat{\theta}] \geq \frac{1}{I_F(\theta)}. \tag{4.16}$$

Si volem estimadors centrats, amb la màxima eficiència (mínima variància), hem de saturar la desigualtat. Això passa quan les dues variables $\widehat{\theta}$ i $\dfrac{\partial L}{\partial \theta}$ tenen una relació lineal, $\dfrac{\partial L}{\partial \theta} = a(\theta)\widehat{\theta} + b(\theta)$. Imposant (4.10), ha de ser $b(\theta) = -a(\theta)E[\widehat{\theta}]$ i així: Pel cas d'estimadors centrats:

$$\frac{\partial L}{\partial \theta} = a(\theta)(\widehat{\theta} - \theta). \tag{4.17}$$

L'estimador de màxima versemblança $\widehat{\theta} = \theta_{MV}$ verifica precisament aquesta condició:

$$0 = \frac{\partial L}{\partial \theta}(\theta_{MV}) = a(\theta_{MV})(\widehat{\theta} - \theta_{MV}).$$

És a dir, si l'estimador de màxima versemblança és centrat, aquest té la màxima eficiència possible.

4.6 Intervals de confiança

Considerem un paràmetre poblacional θ. Un interval de confiança amb nivell de confiança c és un parell de valors γ_1 i γ_2, depenents de les dades mostrals, tals que

$$P(\gamma_1 < \theta < \gamma_2) = c. \tag{4.18}$$

En l'anterior relació θ és el valor del paràmetre poblacional i, per tant, no és una variable aleatòria, però γ_1 i γ_2 depenen de $X_1, X_2, \ldots, X_n$ i és a través de la distribució de probabilitat d'aquestes variables que calculem l'anterior probabilitat. El nivell de confiança se sol expressar en tant per cent. Valors típics són el 95% ($c = 0{,}95$) i el 99% ($c = 0{,}99$). L'interval de confiança és (γ_1, γ_2). també diem que γ_1 i γ_2 són els límits de confiança. Per una c donada, l'interval que verifica (4.18) no és únic. Idealment, hem de triar el més petit possible encara que en la practica solem utilitzar criteris de simetria (situar a cada costat de l'interval la mateixa probabilitat $\dfrac{1-c}{2}$).

El procediment general consisteix en prendre un estimador $\widehat{\theta}$ del paràmetre θ, funció d'una mostra $X_1, \ldots, X_n$ de mida n, i construir una variable $W = \varphi(\widehat{\theta}, \theta, n)$ tal que la seva densitat sigui totalment coneguda. Així, podem trobar valors w_1, w_2 tals que $P(w_1 < W < w_2) = c$ (típicament $w_1 = F_W^{-1}(\frac{1-c}{2})$ i $w_2 = F_W^{-1}(\frac{1+c}{2})$). Finalment, de l'interval $w_1 < W < w_2$ "s'ailla" θ per arribar a $\gamma_1 < \theta < \gamma_2$ on γ_i depenen de $\widehat{\theta}$ i de n.

A continuació determinem intervals per a diversos paràmetres. En molts casos només tenim resultats quan les poblacions són normals o quan la mida mostral n és gran (Teorema del Límit Central).

4.6.1 Intervals per al valor mitjà

Volem intervals de confiança pel valor mitjà poblacional μ.

- **Cas en que coneixem la variància poblacional σ^2.**

  Si la **població és normal**:

  Ja sabem que $\widehat{X} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Per tant, la variable

  $$Z = \frac{\widehat{X} - \mu}{\frac{\sigma}{\sqrt{n}}}. \tag{4.19}$$

  és $Z \sim \mathrm{N}(0,1)$. Si determinem z_c tal que

  $$P(-z_c < Z < z_c) = c, \tag{4.20}$$

  l'esdeveniment en l'anterior probabilitat és $|Z| < z_c$, és a dir, $|\widehat{X} - \mu| < z_c \frac{\sigma}{\sqrt{n}}$. Això és equivalent a $\widehat{X} - z_c \frac{\sigma}{\sqrt{n}} < \mu < \widehat{X} + z_c \frac{\sigma}{\sqrt{n}}$. Per tant, l'interval de confiança per μ correspon als límits de confiança:

  $$\widehat{X} \pm z_c \frac{\sigma}{\sqrt{n}}. \tag{4.21}$$

  Si la **població no és normal**:

  Per **mostres grans** ($n > 30$), pel Teorema del Límit Central $\widehat{X}$ es comporta com una variable $\mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. En aquest cas és vàlid el resultat anterior (4.21).

- **Cas en que no coneixem la variància poblacional σ^2.**

  Si la **població és normal**:

  En aquest cas estimem σ^2 amb la variància mostral $\widehat{S}^2$. La variable a considerar és:

  $$T = \frac{\widehat{X} - \mu}{\frac{\widehat{S}}{\sqrt{n}}}. \tag{4.22}$$

  Podem expressar T com

  $$T = \frac{\frac{\widehat{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\widehat{S}^2}{\sigma^2}}}.$$

  El numerador és una variable $\mathrm{N}(0,1)$. El denominador és $\sqrt{\frac{\chi^2_{n-1}}{n-1}}$. A més, numerador i denominador són independents ja que, com es va veure a l'apartat 3.6, $\widehat{X}$ i S^2 són independents. Llavors, T és una variable t d'Student amb $n-1$ graus de llibertat.

  Similarment al que feiem abans, determinem t_c tal que

  $$P(-t_c < T < t_c) = c, \qquad (T \sim t \text{ d'Student amb } n-1 \text{ graus de llibertat}). \tag{4.23}$$

  Recordem que per mostres grans T tendeix a una variable $\mathrm{N}(0,1)$. Treballant amb el programa R, el càlcul de t_c no té cap dificultat així que no és necessari utilitzar aquesta aproximació.

  Els límits de confiança per μ són:

  $$\widehat{X} \pm t_c \frac{\widehat{S}}{\sqrt{n}}. \tag{4.24}$$

Si la **població no és normal**:

Per mostres petites no disposem d'un resultat general. En aquest cas la forma de la densitat tindrà un efecte significatiu.

Per mostres grans podem utilitzar l'interval (4.24) (podem substituir t_c per z_c, si n és prou gran) encara que si la població difereix molt de la forma normal el resultat no està justificat. Una manera més correcta de procedir és utilitzar **agregació de dades**. Dividim la mostra de mida n en n_1 blocs de mida b propera a 30 (llavors, $n = bn_1$) i passem a una nova mostra de mida n_1 on cada element és la mitjana aritmètica d'un bloc de mida b. Aquestes mitjanes tenen la mateixa esperança que la variable original i, pel Teorema del Límit Central, tenen comportament gaussià, amb el que podem aplicar justificadament la fórmula

Exemple 4.5 Treballarem amb mostres generades amb R. D'aquesta manera coneixem els paràmetres poblacionals i podem jutjar la bondat dels resultats.

En aquest exemple treballem amb una població normal amb valor mitjà 1 i desviació coneguda $\sigma = 2$. Calcularem també l'interval en el cas que no coneguéssim la desviació. La mostra serà de mida 20.

```
> # Generem una mostra de mida 20 d'una població normal amb mitjana 1
> # i desviació 2:
> n<-20
> mu<-1
> sigma<-2
> x<-rnorm(n,mu,sigma)
> x
 [1] -0.6656507  6.5529454  2.1389783 -0.3692862  4.6869197 -0.3640590
 [7]  2.3558156  3.8665102  1.4359156 -0.6746550  3.7232825  1.1461563
[13] -2.0326473  0.3179509 -1.4804710  0.9186429  2.4424767 -0.6991888
[19]  1.3800756 -0.1684814
> # Calculem la seva mitjana mostral:
> m<-mean(x)
> m
[1] 1.225562
> # Fixem un 95\% de confiança i calculem zc:
> c<-0.95
> zc<-qnorm((1+c)/2)
> zc
[1] 1.959964
> # El límit inferior de confiança és:
> m1<-m-zc*sigma/sqrt(n)
> m1
[1] 0.349039
> # El límit superior de confiança és:
> m2<-m+zc*sigma/sqrt(n)
> m2
[1] 2.102084
> #
> # Si no coneguéssim la desviació:
> s<-sd(x)
> s
[1] 2.224325
> # Ara fem servir la t d'Student amb n-1 graus de llibertat:
> tc<-qt((1+c)/2,n-1)
```

```
> tc
[1] 2.093024
> # El límit inferior de confiança és ara:
> mm1<-m-tc*s/sqrt(n)
> mm1
[1] 0.1845454
> # El límit superior de confiança és ara:
> mm2<-m+tc*s/sqrt(n)
> mm2
[1] 2.266578
```

Coneixent σ, l'interval ha estat $(0.349, 2.102)$. Si no coneixem el seu valor, l'interval és més gran: $(0.184, 2.266)$. Amb mostres més grans la diferència es redueix. ♦

Exemple 4.6 Considerem una població de tipus exponencial amb $\mu = 2$. Llavors $\lambda = \dfrac{1}{2}$. Obtindrem una mostra de mida $n = 100$. Encara que la població no és normal, atès que la mostra és gran, utilitzarem el resultat (4.24) estimant σ amb $\widehat{S}$. Trobarem un interval amb un 99% de confiança i el compararem amb el del 95%. Finalment, obtindrem un resultat més correcte aplicant agregació de dades.

```
> # Generem una mostra de mida 200 d'una població exponencial de
> # paràmetre 1/2 (en aquest cas no escriurem els valors obtinguts):
> n<-200
> lambda<-1/2
> x<-rexp(n,lambda)
> # Calculem la seva mitjana i desviació mostrals:
> m<-mean(x)
> m
[1] 2.177696
> s<-sd(x)
> s
[1] 2.409393
> # Fixem un 99% de confiança i calculem zc:
> c<-0.99
> zc<-qnorm((1+c)/2)
> zc
[1] 2.575829
> # El límit inferior de confiança és:
> m1<-m-zc*s/sqrt(n)
> m1
[1] 1.738853
> # El límit superior de confiança és:
> m2<-m+zc*s/sqrt(n)
> m2
[1] 2.61654
> # Ara canviem a un 95% de confiança:
> c<-0.95
> zc<-qnorm((1+c)/2)
> zc
[1] 1.959964
> # El nou límit inferior de confiança és:
> m1<-m-zc*s/sqrt(n)
```

```
> m1
[1] 1.843777
> # El nou límit superior de confiança és:
> m2<-m+zc*s/sqrt(n)
> m2
[1] 2.511615
> # Utilitzem ara el mètode més correcte d'agregació de dades
> # Per a obtenir les mitjanes dels blocs disposem x en forma
> # de matriu amb b files i n1 columnes i sumem les columnes
> b<-20
> n1<-10
> xb<-colSums(matrix(x,b,n1))/b
> mb<-mean(xb)
> mb
[1] 2.177696
> sb<-sd(xb)
> sb
[1] 0.4344453
> tc<-qt((1+c)/2,n1-1)
> # El límit inferior de confiança és:
> mb1<-mb-tc*sb/sqrt(n1)
> mb1
[1] 1.866913
> # El límit superior de confiança és:
> mb2<-mb+tc*sb/sqrt(n1)
> mb2
[1] 2.488479
```

Al 99% l'interval ha estat $(1.74, 2.62)$. Al 95% de confiança, l'interval és més petit: $(1.84, 2.51)$. Com més confiança tinguem que l'interval contingui el valor correcte més gran resulta l'interval. Amb agregació de dades, al 95% de confiança, l'interval és $(1.87, 2.49)$. ♦

4.6.2 Intervals per a proporcions

Suposem que hi ha una propietat que cada element de la població té o no té. La variable X seria una variable de Bernoulli, indicadora de la presència de la propietat. El paràmetre rellevant és p, probabilitat que un element de la població triat a l'atzar tingui la propietat. p és la proporció d'aparició de la propietat sobre el total de la població. Com $p = E[X]$, el nostre estimador és $\widehat{p}$ (és a dir, $\widehat{X}$ o la proporció d'aparició de la propietat en la mostra).

El nombre d'elements en la mostra amb la propietat és $F = X_1 + X_2 + \cdots + X_n$, variable de tipus Binomial(n, p). Llavors $\widehat{p} = \frac{F}{n}$.

Per mostres grans ($n > 30$) podem aproximar $F \sim \mathrm{N}(np, \sqrt{np(1-p)})$ i, per tant, $\widehat{p} \sim \mathrm{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$. En aquest cas, la variable

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}. \tag{4.25}$$

és $Z \sim \mathrm{N}(0, 1)$. A partir del valor z_c tal que $P(-z_c < Z < z_c) = c$ tenim els límits de confiança:

$$\widehat{p} \pm z_c \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}. \tag{4.26}$$

DEM: En principi tenim que el nivell de confiança c correspon a $-z_c < Z < z_c$:

$$-z_c < \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_c.$$

Si considerem Z en funció de p és fàcil veure que és una funció decreixent de ∞ a $-\infty$ per $0 < p < 1$. Llavors la desigualtat anterior equival a $p_1 < p < p_2$ on p_1, p_2 corresponen a les solucions de $\frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \pm z_c$. Elevant al quadrat, arribem a l'equació de segon grau $\left(1 + \frac{z_c^2}{n}\right) p^2 - \left(2\widehat{p} + \frac{z_c^2}{n}\right) p + \widehat{p}^2 = 0$, que té solucions:

$$\frac{1}{\left(1 + \frac{z_c^2}{n}\right)} \left(\widehat{p} + \frac{z_c^2}{2n} \pm z_c \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n} + \frac{z_c^2}{4n^2}} \right).$$

Si n és gran podem despreciar termes d'ordre $\frac{1}{n}$ i quedar-nos només amb els termes fins a ordre $\frac{1}{\sqrt{n}}$, obtenint (4.26). ♣

Una manera més directa de veure el pas de (4.25) a (4.26) és notar que la diferència entre $\widehat{p}$ i p és d'ordre $\frac{1}{\sqrt{n}}$ de manera que poden substituir p per $\widehat{p}$ en el denominador de (4.25).

Exemple 4.7 En tirar un dau 200 vegades, l'1 ha aparegut 46 vegades. Trobar un interval que contingui, amb confiança del 95%, la probabilitat de treure l'1.

```
> n<-200
> p<-46/n
> c<-0.95
> zc<-qnorm((1+c)/2)
> # El límit inferior de confiança és:
> p-zc*sqrt(p*(1-p)/n)
[1] 0.1716767
> # El límit superior de confiança és:
> p+zc*sqrt(p*(1-p)/n)
[1] 0.2883233
```

L'interval $(0.171, 0.288)$ no conté el valor que tindria p en un dau equilibrat ($\frac{1}{6} = 0.166$). Això ens faria sospitar que el nostre dau no és simètric, encara que amb un nivell superior de confiança l'interval creixeria i passaria a contenir el valor $\frac{1}{6}$. ♦

4.6.3 Intervals per a la diferència de paràmetres

En ocasions voldrem conèixer la diferència entre dos paràmetres corresponents a poblacions diferents, típicament la diferència entre els seus valors mitjans. Si tenim dos estadístics, $\widehat{\theta}_1 \sim \mathrm{N}(\theta_1, \sigma_1)$ i $\widehat{\theta}_2 \sim \mathrm{N}(\theta_2, \sigma_2)$, independents, la seva diferència és una variable normal:

$$\widehat{\theta}_1 - \widehat{\theta}_2 \sim \mathrm{N}\left(\theta_1 - \theta_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right) \tag{4.27}$$

Llavors, podem obtenir l'interval a partir de la variable $\mathrm{N}(0, 1)$

$$Z = \frac{\widehat{\theta}_1 - \widehat{\theta}_2 - (\theta_1 - \theta_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

resultant els límits de confiança:

$$\widehat{\theta}_1 - \widehat{\theta}_2 \pm z_c \sqrt{\sigma_1^2 + \sigma_2^2} \tag{4.28}$$

Un cas particular n'és la diferència de proporcions a partir de mostres grans. Si en una població, una mostra de mida n_1 ens dóna una proporció per certa propietat $\widehat{p}_1$ i en una altra població, una mostra de mida n_2 dóna proporcio $\widehat{p}_2$, un interval per la diferència $p_1 - p_2$ és

$$\widehat{p}_1 - \widehat{p}_2 \pm z_c \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \tag{4.29}$$

(En aquest cas $\sigma_i = \dfrac{p_i(1-p_i)}{n_i}$, però, com s'ha comentat anteriorment, en la variància podem substituir p_i per $\widehat{p}_i$.)

Exemple 4.8 En el cas de certa infecció vírica s'ha trobat que en la regió nord una mostra de 300 persones en presentava 25 infectades mentre que en la regió sud una mostra de 200 persones en presentava 19 infectades. Calcular un interval al 95% de confiança per la diferència de les prevalències de la infecció en les dues regions.

```
> n1<-300
> p1<-25/n1
> n2<-200
> p2<-19/n2
> c<-0.95
> zc<-qnorm((1+c)/2)
> # El límit inferior de confiança és:
> p1-p2-zc*sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
[1] -0.06294528
> # El límit superior de confiança és:
> p1-p2+zc*sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
[1] 0.03961195
> p1
[1] 0.08333333
> p2
[1] 0.095
> p1-p2
[1] -0.01166667
```

L'interval $(-0.062, 0.039)$ conté el zero, així que la informació és compatible amb que no hi hagi diferència entre les regions, si bé el valor estimat $\widehat{p}_1 - \widehat{p}_2 = -0.011$ sugereix que p_1 pot ser una mica inferior a p_2. ♦

La fórmula (4.28) requereix conèixer la variància de les poblacions indicades. Si **no coneixem les variàncies poblacionals** les hem d'estimar a partir de les variàncies mostrals. L'estadístic a utilitzar per estimar $\mu_1 - \mu_2$ seria

$$T = \frac{\widehat{X}_1 - \widehat{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\widehat{S_1}^2}{n_1} + \frac{\widehat{S_2}^2}{n_2}}} \tag{4.30}$$

L'anterior variable ja no és de tipus t degut als dos termes diferents en l'arrel del denominador. Hi ha dues maneres de tractar aquesta situació:

- Si sabem que **les dues poblacions tenen la mateixa variància** estimem el valor de la variància comuna amb el que s'anomena **pooling**.

$$\widehat{S_p}^2 = \frac{(n_1-1)\widehat{S_1}^2 + (n_2-1)\widehat{S_2}^2}{n_1 + n_2 - 2} \tag{4.31}$$

L'estadístic és

$$T = \frac{\widehat{X}_1 - \widehat{X}_2 - (\mu_1 - \mu_2)}{\widehat{S_p}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.32}$$

t d'Student amb $n_1 + n_2 - 2$ graus de llibertat. L'interval és:

$$\widehat{X}_1 - \widehat{X}_2 \pm t_c \widehat{S_p}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{4.33}$$

on t_c és tal que $P(-t_c < T < t_c) = c$ amb T Student amb $n_1 + n_2 - 2$ graus de llibertat.

- Si les **variàncies poblacionals són diferents** utilitzem l'**aproximació de Satterwaite** que diu que (4.30) s'aproxima a una t d'Student amb ν graus de llibertat on:

$$\nu = \frac{\left(\frac{\widehat{S_1}^2}{n_1} + \frac{\widehat{S_2}^2}{n_2}\right)^2}{\frac{\widehat{S_1}^4}{n_1^2(n_1-1)} + \frac{\widehat{S_2}^4}{n_2^2(n_2-1)}} \tag{4.34}$$

(es pot utilitzar el seu valor encara que no sigui enter). L'interval és:

$$\widehat{X}_1 - \widehat{X}_2 \pm t_c \sqrt{\frac{\widehat{S_1}^2}{n_1} + \frac{\widehat{S_2}^2}{n_2}} \tag{4.35}$$

on t_c és tal que $P(-t_c < T < t_c) = c$ amb T Student amb ν graus de llibertat.

4.6.4 Intervals per a la variància

Considerem poblacions normals.

Estimem la desviació poblacional σ mitjançant la desviació mostral $\widehat{S}$. La variable amb que treballem per a determinar intervals de confiança és la khi quadrat:

$$\chi^2 = \frac{(n-1)\widehat{S}^2}{\sigma^2} \qquad (n-1 \text{ graus de llibertat}). \tag{4.36}$$

Com la distribució no és simètrica, triarem l'interval (x_1, x_2) per la χ^2 fent que a l'esquerra de x_1 hi hagi la mateixa probabilitat que a la dreta de x_2. Com la probabilitat a l'interior de l'interval és c, obtenim x_1 i x_2 imposant $F_{\chi^2}(x_1) = \frac{1-c}{2}$ i $F_{\chi^2}(x_2) = 1 - \frac{1-c}{2} = \frac{1+c}{2}$. Els valors venen donats per la funció de quantils: $x_1 = F^{-1}_{\chi^2}\left(\frac{1-c}{2}\right)$ i $x_2 = F^{-1}_{\chi^2}\left(\frac{1+c}{2}\right)$. Així, la probabilitat c queda associada a l'esdeveniment:

$$x_1 < \frac{(n-1)\widehat{S}^2}{\sigma^2} < x_2$$

que equival a

$$\frac{\sqrt{n-1}}{\sqrt{x_2}}\widehat{S} < \sigma < \frac{\sqrt{n-1}}{\sqrt{x_1}}\widehat{S}. \tag{4.37}$$

Per mostres grans podem utilitzar l'aproximació de $\widehat{S}^2$ per una normal, si bé fent els càlculs en R aquesta aproximació no suposa cap avantatge.

Exemple 4.9 Generem una mostra $\mathrm{N}(1, 2)$ amb $n = 20$ i obtenim un interval amb un 95% de confiança per a σ.

```
> # Generem una mostra de mida 20 d'una població normal amb mitjana 1
> # i desviació 2:
> n<-20
> mu<-1
> sigma<-2
> x<-rnorm(n,mu,sigma)
> x
 [1]  1.39725863 -1.43982050 -0.94162739  1.38532940  1.82695481 -3.95561485
 [7]  4.29430147  2.11034743  0.32692624  3.22104581  0.07937882  1.79759051
[13] -0.07143744  2.98179141  2.10276450 -0.63204842  3.70798218  3.01051621
[19]  1.11387999  1.65526346
> # Calculem la seva desviació mostral:
> s<-sd(x)
> s
[1] 1.976313
> # Fixem un 95% de confiança i calculem x1, x2:
> c<-0.95
> x1<-qchisq((1-c)/2,n-1)
> x1
[1] 8.906516
> x2<-qchisq((1+c)/2,n-1)
> x2
[1] 32.85233
> # El límit inferior de confiança és:
> s1<-sqrt((n-1)/x2)*s
> s1
[1] 1.502967
> # El límit superior de confiança és:
> s2<-sqrt((n-1)/x1)*s
> s2
[1] 2.886547
```

L'interval per a σ és $(1.502, 2.886)$. ♦

4.6.5 Intervals per al quocient de variàncies

Considerem poblacions normals.

D'una població amb variància σ_1^2 s'obté una mostra de mida n_1 i variància mostral $\widehat{S}_1^2$. D'una altra població amb variància σ_2 s'obté una mostra de mida n_2 i variància mostral $\widehat{S}_2^2$. Volem obtenir intervals per al quocient de variàncies $\dfrac{\sigma_2^2}{\sigma_1^2}$. Considerem l'estadístic:

$$F = \frac{\frac{\widehat{S}_1^2}{\sigma_1^2}}{\frac{\widehat{S}_2^2}{\sigma_2^2}}. \tag{4.38}$$

F és una variable tipus F de Fisher amb $n_1 - 1$, $n_2 - 1$ graus de llibertat.

DEM: Per poblacions normals, $\dfrac{(n_1-1)\widehat{S}_1^2}{\sigma_1^2}$ és $\chi^2_{n_1-1}$ i $\dfrac{(n_2-1)\widehat{S}_2^2}{\sigma_2^2}$ és $\chi^2_{n_2-1}$ i les dues khi quadrat són independents. Llavors

$$\frac{\frac{\widehat{S}_1^2}{\sigma_1^2}}{\frac{\widehat{S}_2^2}{\sigma_2^2}} = \frac{\frac{\chi^2_{n_1-1}}{n_1-1}}{\frac{\chi^2_{n_2-1}}{n_2-1}}$$

que correspon a la definició de la F de Fisher (2.16). ♣

Per determinar un interval amb nivell de confiança c, cerquem valors x_1, x_2 tals que $P(x_1 < F < x_2) = c$ (F de Fisher amb $n_1 - 1$, $n_2 - 1$ graus de llibertat). Com la distribució de Fisher no és simètrica, posem la mateixa probabilitat, $\frac{1-c}{2}$ a cada costat de l'interval i, per tant, $P(F < x_1) = \frac{1-c}{2}$, $P(F < x_2) = \frac{1+c}{2}$. Els valors x_1, x_2 es troben amb la funció de quantils de la variable F. Ara tenim

$$x_1 < \frac{\sigma_2^2 \widehat{S}_1^2}{\sigma_1^2 \widehat{S}_2^2} < x_2,$$

d'on

$$x_1 \frac{\widehat{S}_2^2}{\widehat{S}_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < x_2 \frac{\widehat{S}_2^2}{\widehat{S}_1^2}, \tag{4.39}$$

o

$$\sqrt{x_1} \frac{\widehat{S}_2}{\widehat{S}_1} < \frac{\sigma_2}{\sigma_1} < \sqrt{x_2} \frac{\widehat{S}_2}{\widehat{S}_1}. \tag{4.40}$$

Exemple 4.10 Generem en R dues mostres independents: una de mida 150 en una població $\mathrm{N}(3, 2)$ i una de mida 230 en una població $\mathrm{N}(7, 3)$. A partir de les variàncies mostrals obtindrem un interval al 95% de confiança pel quocient de desviacions (en aquest cas coneixem el seu valor, $\frac{3}{2} = 1.5$).

```
> n1<-150
> mu1<-3
> sigma1<-2
> x<-rnorm(n1,mu1,sigma1)
> n2<-230
> mu2<-7
> sigma2<-3
> y<-rnorm(n2,mu2,sigma2)
> s1<-sd(x)
> s1
[1] 2.139269
> s2<-sd(y)
> s2
[1] 3.06583
> c<-0.95
> x1<-qf((1-c)/2,n1-1,n2-1)
> x1
[1] 0.7428324
> x2<-qf((1+c)/2,n1-1,n2-1)
> x2
[1] 1.333897
> # El límit inferior de confiança per sigma2/sigma1 és:
> sqrt(x1)*s2/s1
[1] 1.235174
> # El límit superior de confiança per sigma2/sigma1 és:
> sqrt(x2)*s2/s1
[1] 1.655175
```

♦

Capítol 5

Test d'hipòtesis

5.1 Hipòtesis estadístiques

En moltes situacions, més que avaluar el valor d'un paràmetre poblacional desconegut el que volem és comprovar si aquest verifica certa suposició com saber si es troba en certa regió o si pren realment un valor donat. Per exemple:

- Saber si la mitjana μ de la població és més petita que un cert valor donat μ_0. Un proveidor d'accés a internet ens assegura una velocitat mitjana μ_0 de conexió. A través de mesures d'aquesta velocitat podem sospitar que la mitjana és realment menor.
- Saber si la mitjana d'una població és diferent de la mitjana d'una altra població. Analitzar si la proporció de articles defectuosos és la mateixa en dues plantes de producció.
- Saber si un paràmetre pren un valor donat. Tenim una moneda i solem suposar que la probabilitat de treure cara és $p = \frac{1}{2}$. Tirant moltes vegades la moneda podem posar a proba aquesta hipòtesi.
- Analitzar si la variable poblacional és de cert tipus. Comprovar que el senyal de soroll en un canal de comunicació és de tipus normal. Aquest cas és més complex que les hipòtesis sobre paràmetres.

Ens centrarem en les hipòtesis sobre paràmetres, donant per coneguts els tipus de variables implicades. Una **hipòtesi simple** és aquella on queda fixat el tipus de variable i els seus paràmetres. És del tipus $\theta = \theta_0$ on θ és un paràmetre de la població. Una **hipòtesi composta** no fixa el valor del paràmetre i li assigna cert conjunt de valors possibles. La hipòtesi **unilateral** pren la forma $\theta > \theta_0$ (**cua dreta**) o $\theta < \theta_0$ (**cua esquerra**). La hipòtesi **bilateral** pren la forma $\theta \neq \theta_0$.

En un problema de test d'hipòtesis partim d'una hipòtesi simple H_0 que s'anomena **hipòtesi nul·la**. La hipòtesi nul·la correspon a l'estat d'informació que tenim a priori (que la moneda és justa, que la conexió a internet té la velocitat que diu el proveidor, etc). Moltes vegades les dades estadístiques poden fer sospitar que aquesta hipòtesi no sigui certa. En aquest cas volem comprovar si les desviacions que observem respecte al que podríem esperar atesa la hipòtesi nul·la són prou significatives. Si ho són, abandonem la hipòtesi nul·la en favor d'una **hipòtesi alternativa** H_A que sol ser de tipus compost. Generalment la hipòtesi alternativa implica que hem de fer canvis amb un possible cost associat així que cal prou evidència per acceptar H_A en lloc de H_0.

Exemple 5.1 Tenim una moneda que suposem justa. Si p és la probabilitat de cara, $H_0 = \{p = \frac{1}{2}\}$. Per a comprovar que és així faríem el test de H_0 contra la hipòtesi alternativa $H_A = \{p \neq \frac{1}{2}\}$. ♦

Exemple 5.2 El temps de transmissió (en ms) de paquets de dades en certa línia és una variable normal amb $\mu = 10$. En cert moment apareixen interferències elèctriques i sospitem que el temps

mitjà és més gran. Llavors $H_0 = \{\mu = 10\}$ i $H_A = \{\mu > 10\}$. L'objectiu del test és veure si les interferències han afectat la transmissió. Cal prou evidència en contra de H_0 ja que en cas contrari haurem de reparar la línia. ♦

5.2 Significació i potència dels tests

Una vegada fet el test haurem arribat a una conclusió que podria ser falsa. En aquest cas tenim un error que pot ser de dos tipus:

- **Error de tipus I:** Rebutgem H_0 quan H_0 és certa.
- **Error de tipus II:** Acceptem H_0 quan H_0 és falsa.

La **significació** α del test és la probabilitat d'error de tipus I:

$$\alpha = P(\text{Error de tipus I}) = P(\text{Rebutjar } H_0 | H_0 \text{ certa}). \tag{5.1}$$

La **potència** del test és la probabilitat de rebutjar H_0 quan H_0 és falsa. És, per tant, $1 - \beta$ on:

$$\beta = P(\text{Error de tipus II}) = P(\text{Acceptar } H_0 | H_0 \text{ falsa}). \tag{5.2}$$

La potència és més difícil d'analitzar ja que la condició no fixa el tipus de variable. La significació s'utilitza per definir el test. El procediment general és considerar un estadístic $\widehat{\theta}$ la distribució del qual queda fixada sota H_0. Aquest estadístic s'obté a partir d'un estimador del paràmetre objecte de la hipòtesi. Llavors dividim el conjunt de valors que pot prendre $\widehat{\theta}$ (típicament, la recta real o el semieix positiu) en dues parts. Una part, $\mathcal{A}$, conté la zona central de probabilitat gran i l'anomenem **regió d'acceptació**. Si $\widehat{\theta} \in \mathcal{A}$, acceptem H_0 (o diem que no tenim prou evidència per rebutjar-la). La regió complementària, $\overline{\mathcal{A}}$ té probabilitat petita. Si $\widehat{\theta} \in \overline{\mathcal{A}}$, rebutgem H_0. Llavors, cometem error de tipus I si, sent H_0 certa, $\widehat{\theta} \notin \mathcal{A}$. Per tant:

$$\alpha = P(\widehat{\theta} \in \overline{\mathcal{A}}). \tag{5.3}$$

on la probabilitat es calcula tenint en compte que la distribució de $\widehat{\theta}$ ve donada per H_0.

Habitualment es treballa no directament amb $\widehat{\theta}$ sino amb una variable $W = \varphi(\widehat{\theta}, \theta_0, n)$ de tipus simple: $\mathrm{N}(0,1)$, χ^2, etc. El test sol prendre el nom d'aquesta variable (tipus Z, tipus T, test χ^2,...).

Valors típics per la significació són el 5% i l'1%. La forma de la regió $\mathcal{A}$ depèn de la hipòtesi alternativa. En un test de cua dreta les desviacions significatives han de ser positives de manera que $\mathcal{A}$ és tot el que es troba a l'esquerra de cert punt. En un test de cua esquerra $\mathcal{A}$ és tot el que es troba a la dreta de cert punt. En un test de doble cua triem $\mathcal{A}$ de manera que a cada un dels seus costats correspongui probabilitat $\dfrac{\alpha}{2}$.

5.3 Cas normal. Test Z

És quan construim un estadístic $Z \sim \mathrm{N}(0,1)$. Definim z_α com el valor tal que $P(Z > z_\alpha) = \alpha$. S'obté amb la funció de quantils.

En el test bilateral la regió d'acceptació és $[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}]$. Si $|Z| \leq z_{\frac{\alpha}{2}}$ acceptem H_0. En cas contrari, acceptem la hipòtesi alternativa.

En el test de cua dreta la regió d'acceptació és $(-\infty, z_\alpha]$. Si $Z \leq z_\alpha$ acceptem H_0. En cas contrari, acceptem la hipòtesi alternativa.

En el test de cua esquerra la regió d'acceptació és $[-z_\alpha, \infty)$. Si $Z \geq -z_\alpha$ acceptem H_0. En cas contrari, acceptem la hipòtesi alternativa.

El test Z s'utilitza quan, sota H_0, referida a un paràmetre θ de la població, tenim un estimador $\widehat{\theta}$ que és normal amb variància coneguda. Llavors

$$Z = \frac{\widehat{\theta} - E[\widehat{\theta}]}{\sqrt{V[\widehat{\theta}]}}. \tag{5.4}$$

La hipòtesi nul·la fixa $E[\widehat{\theta}]$.

Es pot aplicar a valors mitjans, proporcions, comparació de mitjanes, comparació de proporcions. En el cas de proporcions calen mostres grans per a ser vàlida l'aproximació normal.

Exemple 5.3 En cert tipus de línies de comunicació el soroll és una variable $\mathrm{N}(\mu, \sigma)$ on $\sigma = 2$ i μ depèn de la qualitat dels components. La línia de cert fabricant ve especificada amb $\mu = 8$. Una vegada instalada fem proves i obtenim una mostra de mida $n = 20$ amb mitjana $\widehat{X} = 9$. Aplicar el test Z amb significació del 5% per comprovar si podem considerar que la línia segueix l'especificació.

En aquest cas, $H_0 = \{\mu = 8\}$ i $H_A = \{\mu > 8\}$. Fem una hipòtesi de cua dreta ja que podem descartar que el nivell de soroll sigui inferior al de l'especificació.

Sabem que $\widehat{X} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

El valor és $Z = \dfrac{\widehat{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \dfrac{9-8}{\frac{2}{\sqrt{20}}} = 2.236$. Amb $\alpha = 0.05$ trobem $z_\alpha = 1.645$. Com $Z > z_\alpha$, la desviació és significativa i rebutgem H_0.

```
> mu<-8
> sigma<-2
> n<-20
> x<-9
> z<-(x-mu)/(sigma/sqrt(n))
> z
[1] 2.236068
> # quantil per alpha=0.05
> a<-0.05
> za<-qnorm(1-a)
> za
[1] 1.644854
```

Que passa si fem el test a l'1% de significació?

```
> # quantil per alpha=0.01
> a<-0.01
> za<-qnorm(1-a)
> za
[1] 2.326348
```

Ara $z_\alpha = 2.326$ i $Z < z_\alpha$. No hi ha evidència suficient per rebutjar H_0. ♦

5.4 Valors P

Com s'ha vist en l'anterior exemple, el resultat del test depèn de la significació triada. En ocasions això resulta útil. Pot ser que canviar la línia o de fabricant fos molt costós. En aquest cas voldrem molta evidència de que no es verifica l'especificació. Llavors triarem un valor petit de α. També

cal tenir en compte que la mida de la mostra, n, juga un paper important. El mateix valor de la mitjana mostral és més significatiu com més gran és n (Z és proporcional a $\sqrt{n}$) així que podem millorar l'evidència prenen mostres més grans, si és possible.

Una manera més apropiada de procedir és no fixar un valor de α i analitzar com depèn el resultat del test d'aquest valor. Clarament, hi ha un valor límit entre les situacions d'acceptar i rebutjar H_0. Aquest valor correspon a $Z = z_\alpha$. La significació límit és $\alpha = P(Z \geq Z^*)$ on Z^* és el valor obtingut per la variable Z. Aquesta α s'anomena valor P en el test. Com més petit és el valor P més significatives són les desviacions.

Exemple 5.4 Continuant l'exemple anterior, calculem el valor P:

```
> P<-1-pnorm(z)
> P
[1] 0.01267366
```

Veient aquest valor queda clar quin seria el resultat del test a $\alpha = 0.05$ i a $\alpha = 0.01$. Però a més veiem que el valor P és més proper a 0.01, fet que desplaça el pes a rebutjar H_0. ♦

En general[1], si $F_Z(z)$ és la funció de distribució de Z:

- Test bilateral: $P = 2F_Z(-|Z^*|)$.
- Test de cua dreta: $P = 1 - F_Z(Z^*)$.
- Test de cua esquerra: $P = F_Z(Z^*)$.

La manera d'interpretar el valor P és:

- $0.1 \leq P < 1$: Acceptem H_0.
- $0 < P \leq 0.01$: Rebutgem H_0.
- $0.01 < P < 0.1$: Zona gris on convé anar amb cautela. És preferible repetir el test, amb mostres més grans si es pot. Si és imprescindible prendre una decisió amb només aquesta informació, el criteri pot ser acceptar H_0 per $P \geq 0.05$ i rebutjar H_0 per $P < 0.05$.

Exemple 5.5 Tenim una moneda i volem comprovar si és justa. Si la probabilitat de cara és p, $H_0 = \left\{p = \frac{1}{2}\right\}$. El test consistirà en tirar la moneda $n = 100$ vegades i contar el nombre X de cares obtingudes. Sota H_0 i atès que la mostra és gran, podem considerar $X \sim \mathrm{N}(np, \sqrt{np(1-p)})$. Llavors $Z = \dfrac{X - np}{\sqrt{np(1-p)}}$. Farem el test bilateral.

Suposem que s'obtenen 41 cares. Aplicar primer el test al 5% de significació. Calcular després el valor P.

```
> n<-100
> x<-41
> p<-0.5
> z<-(x-n*p)/sqrt(n*p*(1-p))
> z
[1] -1.8
> a<-0.05
```

[1] En el cas discret, el valor P ha d'incloure el punt Z^*.

```
> za<-qnorm(1-a/2)
> za
[1] 1.959964
> P<-2*pnorm(-abs(z))
> P
[1] 0.07186064
```

$|Z| = 1.8 < z_{\frac{\alpha}{2}}$ així que acceptem que la moneda és justa. El valor P és troba en la zona gris, així que el resultat no és massa concloent. En qualsevol cas és $P > 0.05$ fet favorable a acceptar H_0.

Què passa si s'obtenen 67 cares en aquestes 100 tirades?

```
> n<-100
> x<-67
> p<-0.5
> z<-(x-n*p)/sqrt(n*p*(1-p))
> z
[1] 3.4
> P<-2*pnorm(-abs(z))
> P
[1] 0.0006738585
```

El valor P molt menor que 0.01 senyala que la moneda no és justa. ♦

Exemple 5.6 Comparació de proporcions

Un problema típic és determinar si la proporció de cert tipus d'elements és la mateixa en dues poblacions diferents. Per exemple, considerem la proporció de productes defectuosos en dues plantes de producció (o en la mateixa planta abans i després de modificar quelcom en el procés de fabricació).

Tenim la població A on una mostra de mida n_A dóna una proporció $\widehat{p}_A$ i la població B amb una mostra de mida n_B i proporció $\widehat{p}_B$.

Tenim $H_0 = \{p_A = p_B\}$. Anomenem p al valor comú de les proporcions en H_0. Si les mostres són grans tenim dues variables normals independents $\widehat{p}_A \sim \mathrm{N}\left(p, \sqrt{\frac{p(1-p)}{n_A}}\right)$ i $\widehat{p}_B \sim \mathrm{N}\left(p, \sqrt{\frac{p(1-p)}{n_B}}\right)$. La variable $\widehat{p}_A - \widehat{p}_B$ és $\mathrm{N}\left(0, \sqrt{\frac{p(1-p)}{n_A} + \frac{p(1-p)}{n_B}}\right)$.

La variable $\mathrm{N}(0,1)$ és:

$$Z = \frac{\widehat{p}_A - \widehat{p}_B}{\sqrt{p(1-p)(\frac{1}{n_A} + \frac{1}{n_B})}}. \tag{5.5}$$

Com en aquest cas p no és coneguda, l'estimarem amb el mètode de *pooling*:

$$\widehat{p} = \frac{n_A\widehat{p}_A + n_B\widehat{p}_B}{n_A + n_B}. \tag{5.6}$$

En (5.5) posem $\widehat{p}$ en lloc de p.

Passem a un exemple concret. En una planta de producció una mostra de mida 200 mostrava un 2% de productes defectuosos. En una altra planta, una mostra de mida 300 mostrava un 4% de productes defectuosos. Podem considerar que en la segona planta la producció funciona de manera pitjor?

Tenim $n_A = 200$, $\widehat{p}_A = 0.02$, $n_B = 300$, $\widehat{p}_B = 0.04$.

$H_0 = \{p_A = p_B\}$ i $H_A = \{p_B > p_A\}$ (cua esquerra ja que és $\widehat{p}_A - \widehat{p}_B < 0$).

```
> nA<-200
> pA<-0.02
> nB<-300
> pB<-0.04
> p<-(nA*pA+nB*pB)/(nA+nB)
> p
[1] 0.032
> Z<-(pA-pB)/sqrt(p*(1-p)*(1/nA+1/nB))
> Z
[1] -1.244824
> P<-pnorm(Z)
> P
[1] 0.1065982
```

El valor gran de P (0.1) indica que hem d'acceptar la hipòtesi nul·la. La diferència entre les dues plantes no és prou significativa. ♦

5.5 Variància desconeguda. Test T

En poblacions normals on no coneixen la variància volem fer tests sobre el valor mitjà. En aquest cas estimem la variància a partir de $\widehat{S}^2$ i treballem amb la t d'Student.

$$T = \frac{\widehat{X} - \mu}{\frac{\widehat{S}}{\sqrt{n}}}. \tag{5.7}$$

T és de tipus t amb $n-1$ graus de llibertat. La distribució de la t és simètrica al voltant de zero així que el test T procedeix igual que el test Z.

Exemple 5.7 A partir de raonaments teórics, es dedueix que el temps que dura cert procés en un ordinador és una variable normal amb valor mitjà 10. En 6 mesures obtenim els temps: 11.7, 13.7, 10.9, 9.5, 10.8, 10.5. Fer un test T sobre la hipòtesi nul·la que el raonament teòric és correcte.

```
> mu<-10
> n<-6
> x<-c(11.7,13.7,10.9,9.5,10.8,10.5)
> m<-mean(x)
> m
[1] 11.18333
> s<-sd(x)
> s
[1] 1.423259
> t<-(m-mu)/(s/sqrt(n))
> P<-2*pt(-abs(t),n-1)
> P
[1] 0.0972934
```

El valor de P proper a 0.1 ens porta a acceptar la hipòtesi nul·la. ♦

Per a decidir si dues mostres independents provenen de la mateixa població normal considerem la variable $\widehat{X} - \widehat{Y}$. Si les dues provenen d'una població $\mathrm{N}(\mu, \sigma)$, l'anterior variable té mitjana zero

i variància $\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}$. La següent variable és de tipus t amb n_1+n_2-2 graus de llibertat:

$$T=\frac{\widehat{X}-\widehat{Y}}{\widehat{S}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}. \tag{5.8}$$

on $\widehat{S}^2$ és la variància comú estimada (pooling):

$$\widehat{S}^2=\frac{(n_1-1)\widehat{S}_x^2+(n_2-1)\widehat{S}_y^2}{n_1+n_2-2}. \tag{5.9}$$

DEM: Les variables $\widehat{X}\sim \mathrm{N}\left(\mu,\frac{\sigma}{\sqrt{n_1}}\right)$, $\widehat{Y}\sim \mathrm{N}\left(\mu,\frac{\sigma}{\sqrt{n_2}}\right)$, $(n_1-1)\frac{\widehat{S}_x^2}{\sigma^2}\sim\chi^2_{n_1-1}$ $(n_2-1)\frac{\widehat{S}_y^2}{\sigma^2}\sim\chi^2_{n_2-1}$ són independents.

Llavors, $\widehat{X}-\widehat{Y}\sim \mathrm{N}\left(0,\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$ d'on $\frac{\widehat{X}-\widehat{Y}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim \mathrm{N}(0,1)$.

També tenim $(n_1-1)\frac{\widehat{S}_x^2}{\sigma^2}+(n_2-1)\frac{\widehat{S}_y^2}{\sigma^2}\sim\chi^2_{n_1+n_2-2}$, és a dir, $(n_1+n_2-2)\frac{\widehat{S}^2}{\sigma^2}\sim\chi^2_{n_1+n_2-2}$.

Així, $T=\frac{\mathrm{N}(0,1)}{\chi_{n_1+n_2-2}/\sqrt{n_1+n_2-2}}=\frac{\widehat{X}-\widehat{Y}}{\widehat{S}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ és t amb $df=n_1+n_2-2$. ♣

Exemple 5.8 El temps de durada de la bateria d'un mòvil és una variable normal. Pel model original de bateria es mesuren 10 temps de durada amb valor mitjà $\widehat{X}=7.5$ i desviació $\widehat{S}_x=2.2$. Per un model clònic es fan 15 mesures amb $\widehat{Y}=6.0$ i $\widehat{S}_y=2.5$. Hi ha diferència apreciable entre la durada dels dos models?

Com el model clònic sembla durar meny temps farem un test de cua dreta ($H_A=\{\mu_x-\mu_y>0\}$).

```
> n1<-10
> mx<-7.5
> sx<-2.2
> n2<-15
> my<-6
> sy<-2.5
> s<-sqrt(((n1-1)*sx^2+(n2-1)*sy^2)/(n1+n2-2))
> t<-(mx-my)/(s*sqrt(1/n1+1/n2))
> P<-1-pt(t,n1+n2-2)
> P
[1] 0.06870034
```

El valor de P no és prou petit per rebutjar la hipòtesi nul·la, si bé som a la zona gris i convindria analitzar-ho amb mostres més grans. ♦

5.6 Tests per a la variància

Per poblacions normals, les variables (4.27) i (4.29) permeten contrastar hipòtesis sobre la variància d'una població i la comparació de variàncies de dues poblacions, respectivament. La hipòtesi nul·la dirà que una població té un determinat valor de σ o que dues poblacions tenen el mateix valor de σ. Com es tracta de variables amb valors positius, pel valor P hem de veure si el valor obtingut es troba a la dreta o esquerra del valor "central" ($F=1$ o $\chi^2=n-1$) i, en el cas del test bilateral, multiplicar per 2 la probabilitat de la cua.

Exemple 5.9 En l'exemple anterior hem suposat que la desviació era la mateixa en les dues poblacions. Així $H_0 = \{\sigma_1 = \sigma_2\}$. Llavors

$$F = \frac{\widehat{S}_x^2}{\widehat{S}_y^2}$$

és de tipus F amb $n_1 - 1$ i $n_2 - 1$ graus de llibertat. Fem un test bilateral.

```
> n1<-10
> sx<-2.2
> n2<-15
> sy<-2.5
> f<-sx^2/sy^2
> f
[1] 0.7744
> P<-2*pf(f,n1-1,n2-1)
> P
[1] 0.7148895
```

Com $F = 0.77 < 1$ hem calculat la probabilitat a l'esquerra i l'hem multiplicat per 2. El valor de P és molt gran i permet acceptar clarament la hipòtesi que la desviació és la mateixa en els dos tipus de bateria. ♦

Exemple 5.10 Segons cert model, la desviació en les notes d'un examen ha de valer $\sigma = 1.5$. En un examen amb 90 estudiants resulta $\widehat{S} = 2$. És consistent aquest resultat amb el model teòric?

Sota $H_0 = \{\sigma = 1.5\}$, la variable $\chi^2 = (n-1)\dfrac{\widehat{S}^2}{\sigma^2} = 89\dfrac{\widehat{S}^2}{1.5^2}$ és khi quadrat amb 89 graus de llibertat. Fem un test bilateral.

```
> sigma<-1.5
> n<-90
> s<-2
> chi2<-(n-1)*s^2/sigma^2
> chi2
[1] 158.2222
> P<-2*(1-pchisq(chi2,n-1))
> P
[1] 1.772824e-05
```

P és molt petit així que rebutgem la validesa del model. ♦

5.7 Test χ^2

Considerem un població on cada element pertany a una de les classes $A_1, A_2, \ldots, A_r$. Sigui p_i la probabilitat que un element pertanyi a la classe A_i, $i = 1, \ldots, r$. En una mostra de mida n el nombre d'elements en cadascuna de les classes forma una variable multinomial. El següent test determina si els valors obtinguts per aquestes freqüències són consistents amb els valors p_i. Si les freqüències observades són X_i, $i = 1, \ldots, r$ definim la variable

$$\chi^2 = \sum_{i=1}^{r} \frac{(X_i - np_i)^2}{np_i}. \tag{5.10}$$

que s'aproxima bé per una khi quadrat amb $r-1$ graus de llibertat si n és prou gran ($np_i > 5$, $\forall i$). Sobre aquest variable és fa el test de cua dreta.

DEM: Les variables X_i són binomials i s'aproximen a normals quan $n \to \infty$, però no són independents ja que $X_1 + X_2 + \cdots + X_r = n$. Utilitzarem el següent resultat:

La distribució conjunta de r variables de Poisson independents, $Y_i \sim \text{Poisson}(\alpha_i), i = 1, \ldots, r$ sotmeses a la condició $Y_1 + Y_2 + \cdots + Y_r = n$ és una variable r-nomial amb probabilitats $p_i = \dfrac{\alpha_i}{\alpha}$ on $\alpha = \sum_j \alpha_j$. En efecte:

$$P\left(Y_1 = n_1, \cdots, Y_r = n_r \mid \sum_j Y_j = n\right) = \frac{P(Y_1 = n_1, \cdots, Y_r = n_r, \sum_j Y_j = n)}{P\left(\sum_j Y_j = n\right)}.$$

El numerador de l'expressió anterior val zero si $\sum_j Y_j \neq n$ i $\prod_j P(Y_j = n_j) = \prod_j e^{-\alpha_j} \dfrac{\alpha_j^{n_j}}{n_j!} = e^{-\alpha} \prod_j \dfrac{\alpha_j^{n_j}}{n_j!}$ si $\sum_j Y_j = n$.

Ara, tenint en compte que la suma de variables de Poisson independents també és de Poisson amb paràmetre donat per la suma dels paràmetres, tenim que $\sum_j Y_j$ és una variable de poisson de paràmetre α. El denominador val doncs $e^{-\alpha} \dfrac{\alpha^n}{n!}$. El quocient de les dues probabilitats dóna:

$$P\left(Y_1 = n_1, \cdots, Y_r = n_r | \sum_j Y_j = n\right) = \frac{n!}{\prod_j n_j!} \cdot \frac{\prod_k \alpha_k^{n_k}}{\alpha^n} = \frac{n!}{\prod_j n_j!} \left(\frac{\alpha_1}{\alpha}\right)^{n_1} \cdots \left(\frac{\alpha_r}{\alpha}\right)^{n_r}.$$

que és la fórmula de les probabilitats multinomials.

Així, considerem les variables $X_i \sim \text{Poisson}(\alpha_i)$ amb $\alpha_i = np_i$ verificant la condició $\sum_i X_i = n$. Una variable de Poisson de paràmetre np_i es pot considerar com a suma de n variables Poisson(p_i) independents. Pel Teorema del Límit Central aquesta variable s'aproxima a una $\text{N}(np_i, \sqrt{np_i})$. Llavors, per n gran la variable 5.10 és

$$\chi^2 = \sum_{i=1}^{r} Z_i^2.$$

on $Z_i = \dfrac{X_i - np_i}{\sqrt{np_i}} \sim \text{N}(0,1)$ i es verifica la condició $\sum_i \sqrt{p_i} Z_i = 0$.

La variable $(Z_1, Z_2, \ldots, Z_r)$ té densitat conjunta dependent només de $z_1^2 + z_2^2 + \cdots + z_r^2$ i, per tant, té simetria esfèrica. La condició ens limita a un hiperplà donat, que passa per l'origen de coordenades. Degut a la simetria esfèrica qualsevol altre hiperplà que passi per l'origen donarà lloc a la mateixa variable. Podem triar l'hiperplà $Z_r = 0$ amb el que

$$\chi^2 = \sum_{i=1}^{r-1} Z_i^2.$$

sense cap restricció. Així χ^2 és khi quadrat amb $r-1$ graus de llibertat. ♣

Exemple 5.11 Volem comprovar si un dau és just tirant-lo 180 vegades i aplicant el test χ^2 a les freqüències obtingudes pels 6 resultats possibles. En fer-ho obtenim:

i	1	2	3	4	5	6
X_i	18	25	40	37	15	45

La hipòtesi nul·la és que el dau és just i $p_i = \dfrac{1}{6}$ per $i = 1, \ldots, 6$.

```
> n<-180
> x<-c(18,25,40,37,15,45)
> r<-length(x)
```

```
> p<-1/r
> chi2<-sum((x-n*p)^2/(n*p))
> Pvalue<-1-pchisq(chi2,r-1)
> Pvalue
[1] 0.0001066716
```

P és molt petit així que el dau no està equilibrat.

Repetim el problema simulant en R la tirada d'un dau:

```
> n<-180
> d<-sample(1:6,n,replace=TRUE)
> d
  [1] 1 2 3 1 2 4 2 4 5 2 4 1 5 1 3 6 1 1 2 3 4 4 6 5 6 2 6 2 1 4 2 2 5 5 5 3 1
 [38] 4 4 2 1 3 5 4 4 1 3 1 4 3 4 1 5 1 5 1 2 6 5 2 3 5 4 3 1 1 4 6 5 1 1 4 3 3
 [75] 6 2 2 1 4 5 2 6 2 2 6 6 5 1 5 1 5 1 3 1 3 2 6 1 6 3 3 2 2 4 6 4 5 1 6 4 2
[112] 4 2 6 2 2 1 3 6 3 6 1 2 5 1 5 6 5 6 6 4 5 2 2 3 3 6 4 1 6 4 5 3 3 2 3 3 4
[149] 1 4 6 1 3 2 6 3 1 6 2 4 5 6 3 6 2 5 2 2 6 6 2 5 1 6 6 3 2 1 1 4
> d1<-sum(d==1)
> d2<-sum(d==2)
> d3<-sum(d==3)
> d4<-sum(d==4)
> d5<-sum(d==5)
> d6<-sum(d==6)
> x<-c(d1,d2,d3,d4,d5,d6)
> x
[1] 35 35 27 27 25 31
> r<-length(x)
> p<-1/r
> chi2<-sum((x-n*p)^2/(n*p))
> Pvalue<-1-pchisq(chi2,r-1)
> Pvalue
[1] 0.6794384
```

Ara el valor P és gran i acceptem que el dau és just. ♦

Una aplicació particular del test χ^2 és **comprovar si un conjunt de dades s'ajusta a certa distribució de probabilitat**. En aquest cas dividim els possibles valors de la variable de manera que en cada interval en trobem més de 5. Les probabilitats de cada interval s'obtenen de la distribució suposada. Normalment no coneixerem els valors dels paràmetres que conté aquesta distribució i els haurem d'estimar a partir de les dades. Un teorema que no demostrem diu que:

Si dels r valors p_i n'estimem k a partir de les dades pel mètode de la màxima versemblança, la variable χ^2 definida en (5.10) és de tipus khi quadrat amb $r-k-1$ graus de llibertat.

Exemple 5.12 Estudiem la durada en minuts de les sessions dels usuaris d'un servei per internet. Amb una mostra de 100 sesions obtenim:

minuts	23	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
sessions	1	3	6	2	5	4	3	6	7	14	3	4	5	10	4	4	5

minuts	42	43	44	45	47	50
sessions	4	3	3	1	2	1

Estudiem si s'ajusten bé a una variable normal. En aquest cas hem de determinar dos paràmetres a partir de les dades. El mètode de màxima versemblança estima μ a través de la mitjana mostral $\widehat{X}$ i σ a través de la desviació mostral S. Com les freqüencies que es veuen a la taula són petites triarem intervals més grans:

Interval	$(-\infty, 29.5)$	$(29.5, 33.5)$	$(33.5, 37.5)$	$(37.5, 41.5)$	$(41.5, \infty)$
freqüència	17	20	26	23	14

```
> x<-c(23,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,47,50)
> fx<-c(1,3,6,2,5,4,3,6,7,14,3,4,5,10,4,4,5,4,3,3,1,2,1)
> m<-weighted.mean(x,fx)
> s<-sqrt(weighted.mean(x^2,fx)-m^2)
> n<-100
> r<-5
> k<-2
> f<-c(17,20,26,23,14)
> p1<-pnorm(29.5,m,s)
> p2<-pnorm(33.5,m,s)-pnorm(29.5,m,s)
> p3<-pnorm(37.5,m,s)-pnorm(33.5,m,s)
> p4<-pnorm(41.5,m,s)-pnorm(37.5,m,s)
> p5<-1-pnorm(41.5,m,s)
> p<-c(p1,p2,p3,p4,p5)
> chi2<-sum((f-n*p)^2/(n*p))
> Pvalue<-1-pchisq(chi2,r-k-1)
> Pvalue
[1] 0.5469445
```

El valor P gran indica que l'ajust és bo. ♦

Capítol 6

Regressió

6.1 Formulació del problema

Volem establir connexions entre diverses variables aleatòries associades a certa població. Cada element de la població té associades vàries variables. La variable principal l'anomenem Y. Les variables secundàries són $X^{(\gamma)}$, $\gamma = 1, 2, \ldots, r$. L'objectiu és obtenir funcions que aplicades a les variables secundàries ens donin una bona estimació de la variable principal. És a dir, obtenir un estimador de Y, $\widehat{Y}$:

$$\widehat{Y} = \varphi(X^{(1)}, X^{(2)}, \ldots, X^{(r)}). \tag{6.1}$$

En el capítol d'estimació s'ha vist que, a nivell de variables aleatòries, la millor estimació la dóna l'esperança condicionada: $\widehat{Y} = E[Y|X^{(1)}, X^{(2)}, \ldots, X^{(r)}]$. Ara considerem el problema a nivell estadístic on el que tenim és una mostra de la variable $r+1$ dimensional $(X^{(1)}, X^{(2)}, \ldots, X^{(r)}, Y)$. Notem que el que farem és obtenir a partir de les dades estadístiques una estimació de la funció $\varphi(X^{(1)}, X^{(2)}, \ldots, X^{(r)}) = E[Y|X^{(1)}, X^{(2)}, \ldots, X^{(r)}]$. Recordem també que pel cas de variables normals la funció $\varphi(X^{(1)}, X^{(2)}, \ldots, X^{(r)})$ és lineal, és a dir, té la forma $\alpha^{(1)}X^{(1)} + \alpha^{(2)}X^{(2)} + \cdots + \alpha^{(r)}X^{(r)} + \beta$. Ens concentrarem per tant en aquest tipus d'estimació.

6.2 Significat de les variables secundàries

Encara que inicialment considerem que les variables $X^{(\alpha)}$ són variables aleatòries que tenen, junt amb la variable Y, una distribució de probabilitat conjunta, en molts casos es tracten com paràmetres tals que la distribució de Y queda fixada pel seus valors. En aquest cas el mostreig consisteix en donar valors a les variables secundàries i fer una (o vàries) mesures de Y per aquestes $X^{(\alpha)}$ donades.

Per exemple, Y pot ser una variable que depèn del temps. En aquest cas la variable secundària és el temps t. Típicament, triem una seqüencia de valors de t i anem mesurant Y per a cada un d'ells.

Un altre cas és el de dues magnituds físiques lligades de manera que una depèn determinísticament de l'altra: $y = g(x)$. Si no coneixem la funció $g(x)$, aniríem donant valors a x i mesurant y per determinar g de manera empírica. Ara suposem que les mesures venen acompanyades d'un error i, fixada x, el resultat de la mesura és una variable aleatòria Y.

En aquestes situacions tenim també que $E[Y|X^{(1)}, X^{(2)}, \ldots, X^{(r)}]$ és una funció de les variables $X^{(\alpha)}$. El tractament que es fa és el mateix que quan $X^{(\alpha)}$ són variables aleatòries i, per tant, seguim parlant de valors mitjans, desviacions mostrals, etc. per aquests paràmetres.

6.3 Notació

Amb una mostra de mida n expressarem amb minúscules els valors obtinguts per les variables: $x_i^{(\alpha)}$, $\alpha = 1, \ldots, r$, i y_i, $i = 1, \ldots, n$. Per a cada variable tenim un vector de $\mathbb{R}^n$ amb components donats pels valors obtinguts en la mostra d'aquesta variable: $x^{(\alpha)} = (x_1^{(\alpha)}, \ldots, x_n^{(\alpha)})$, $y = (y_1, \ldots, y_n)$.

Definim els valors mitjans, covariàncies i variàncies mostrals:

$$\overline{x}^{(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\alpha)}, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{6.2}$$

$$S_{x^{(\alpha)},x^{(\beta)}} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\alpha)}x_i^{(\beta)} - \overline{x}^{(\alpha)}\overline{x}^{(\beta)}, \qquad S_{x^{(\alpha)},y} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(\alpha)}y_i - \overline{x}^{(\alpha)}\overline{y}. \tag{6.3}$$

$$S^2_{x^{(\alpha)}} = S_{x^{(\alpha)},x^{(\alpha)}} = \frac{1}{n}\sum_{i=1}^{n}(x_i^{(\alpha)})^2 - (\overline{x}^{(\alpha)})^2, \qquad S_y^2 = S_{y,y} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2. \tag{6.4}$$

Amb el producte escalar ordinari a $\mathbb{R}^n$, $< v, w >= \sum_{i=1}^{n} v_i w_i$ i definint el vector "constant" $u = (1, 1, \ldots, 1) \in \mathbb{R}^n$:

$$< u, u >= n, \qquad < u, x^{(\alpha)} >= n\overline{x}^{(\alpha)}, \qquad < u, y >= n\overline{y}. \tag{6.5}$$

$$< x^{(\alpha)}, x^{(\beta)} >= n(S_{x^{(\alpha)},x^{(\beta)}} + \overline{x}^{(\alpha)}\overline{x}^{(\beta)}), \qquad < x^{(\alpha)}, y >= n(S_{x^{(\alpha)},y} + \overline{x}^{(\alpha)}\overline{y}). \tag{6.6}$$

$$< y, y >= n(S_y^2 + \overline{y}^2). \tag{6.7}$$

6.4 Models lineals. Regressió multivariable

A partir de les dades mostrals hem de determinar un estimador lineal:

$$\widehat{y}_i = a^{(1)}x_i^{(1)} + a^{(2)}x_i^{(2)} + \cdots + a^{(r)}x_i^{(r)} + b, \quad i = 1 \ldots, n \tag{6.8}$$

on les constants $a^{(\alpha)}$ i b estimen les corresponents constants $\alpha^{(\alpha)}$ i β de l'estimador en termes de variables aleatòries

$$\widehat{Y} = \alpha^{(1)}X^{(1)} + \alpha^{(2)}X^{(2)} + \cdots + \alpha^{(r)}X^{(r)} + \beta \tag{6.9}$$

Recordem que $\alpha^{(\alpha)}$ i β es determinaven amb el principi d'ortogonalitat que minimitzava l'error definit a partir d'una distància euclidiana corresponent al producte escalar entre variables aleatòries ($< U, V >= E[UV]$). Ara no coneixem aquests valors i treballem amb els valors mostrals. El resultat, però, serà anàleg, treballant a $\mathbb{R}^n$. Escrivint (6.8) en forma vectorial:

$$\widehat{y} = a^{(1)}x^{(1)} + a^{(2)}x^{(2)} + \cdots + a^{(r)}x^{(r)} + bu, \tag{6.10}$$

imposarem que $\widehat{y} = \sum_{\beta=1}^{r} a^{(\beta)}x^{(\beta)} + bu$ sigui el més proper possible a y. Per això hem de minimitzar la distància entre els vectors $\widehat{y}$ i y, fet que es dóna quan $\widehat{y}$ és la projecció ortogonal de y sobre

l'espai generat pels vectors $x^{(1)}, x^{(2)}, \ldots, x^{(r)}, u$. Llavors ha de ser $(\widehat{y} - y) \perp x^{(\alpha)}$, $\alpha = 1, \ldots, r$ i $(\widehat{y} - y) \perp u$. Les equacions resultants són:

$$\begin{cases} < \sum_{\beta=1}^{r} a^{(\beta)} x^{(\beta)} + bu, x^{(\alpha)} > = < y, x^{(\alpha)} >, \quad \alpha = 1, \ldots, r, \\ < \sum_{\beta=1}^{r} a^{(\beta)} x^{(\beta)} + bu, u > = < y, u >, \end{cases} \tag{6.11}$$

Utilitzant les relacions (6.5) i (6.6) i dividint per n els dos costats:

$$\begin{cases} \sum_{\beta=1}^{r} a^{(\beta)} (S_{x^{(\alpha)}, x^{(\beta)}} + \overline{x}^{(\alpha)} \overline{x}^{(\beta)}) + b\overline{x}^{(\alpha)} = S_{x^{(\alpha)}, y} + \overline{x}^{(\alpha)} \overline{y}, \quad \alpha = 1, \ldots, r, \\ \sum_{\beta=1}^{r} a^{(\beta)} \overline{x}^{(\beta)} + b = \overline{y}, \end{cases} \tag{6.12}$$

La segona equació ens dóna

$$b = \overline{y} - \sum_{\beta=1}^{r} a^{(\beta)} \overline{x}^{(\beta)}, \tag{6.13}$$

Substituint b en les restants equacions arribem al sistema

$$\sum_{\beta=1}^{r} S_{x^{(\alpha)}, x^{(\beta)}} a^{(\beta)} = S_{x^{(\alpha)}, y}, \quad \alpha = 1, \ldots, r. \tag{6.14}$$

Aquest sistema sempre té solució ja que la projecció ortogonal sempre existeix. La solució és única (si els vectors $x^{(1)}, x^{(2)}, \ldots, x^{(r)}, u$ són linealment independents). Així obtenim les constants $a^{(\alpha)}$ i substituint-les en (6.13) obtenim b.

6.5 Residuals. Expressió de l'error

Una vegada determinat el model lineal (6.8), aquest dóna una estimació $\widehat{y}_i$ del valor de y_i. L'error és defineix $\epsilon_i = y_i - \widehat{y}_i$. El conjunt de variables ϵ_i s'anomenen **residuals**. L'**error quadràtic mitjà** és

$$\overline{\epsilon}^2 = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2, \tag{6.15}$$

A partir dels coeficients de l'estimació:

$$\overline{\epsilon}^2 = S_y^2 - \sum_{\alpha=1}^{r} a^{(\alpha)} S_{x^{(\alpha)}, y}. \tag{6.16}$$

DEM:

$$\overline{\epsilon}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \frac{1}{n} < y - \widehat{y}, y - \widehat{y} > = \frac{1}{n} < y - \widehat{y}, y >$$

(ja que $\widehat{y}$ és ortogonal a $y - \widehat{y}$)

$$= \frac{1}{n} \left(< y, y > - < \sum_{\beta=1}^{r} a^{(\beta)} x^{(\beta)} + bu, y > \right) = S_y^2 + \overline{y}^2 - \sum_{\beta=1}^{r} a^{(\beta)} (S_{x^{(\beta)}, y} + \overline{x}^{(\beta)} \overline{y}) - b\overline{y}$$

$$= S_y^2 + \overline{y}^2 - \sum_{\beta=1}^{r} a^{(\beta)} (S_{x^{(\beta)}, y} + \overline{x}^{(\beta)} \overline{y}) - \left(\overline{y} - \sum_{\beta=1}^{r} a^{(\beta)} \overline{x}^{(\beta)} \right) \overline{y} = S_y^2 - \sum_{\alpha=1}^{r} a^{(\alpha)} S_{x^{(\alpha)}, y}. \clubsuit$$

6.6 Coeficient de correlació generalitzat

Considerem els vectors de $\mathbb{R}^n$: y, $\widehat{y}$ i ϵ. Tenim que $y = \widehat{y} + \epsilon$. Hem descompost y en una part que descriu el model lineal més un "error". Diem que $\widehat{y}$ és la **part explicada** de y i que ϵ és la **part no explicada** de y. La qualitat del model es mesura a partir de com és de petita la part no explicada. Això ho determinem amb les variàncies mostrals.

La **variació total** és $V_T = S_y^2$. La **variació no explicada** és $V_{NE} = \overline{\epsilon}^2$. La **variació explicada** és V_E igual a la variància mostral de $\widehat{y}$. Resulta

$$V_T = V_E + V_{NE}. \tag{6.17}$$

DEM: La última equació del sistema (6.12) mostra que $\widehat{y}$ i y tenen la mateixa mitjana mostral i, per tant, ϵ té mitjana mostral zero. De la descomposició ortogonal obtenim $< y, y >=< \widehat{y}, \widehat{y} > + < \epsilon, \epsilon >$. Dividint per n els dos costats i restant $\overline{y}^2$ tenim

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i^2 - \overline{y}^2 + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$$

on $V_T = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2$, $V_E = \frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i^2 - \overline{y}^2$ i $V_{NE} = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$.♣

Definim el coeficient de correlació generalitzat, R^2

$$R^2 = \frac{V_E}{V_T} = \frac{S_y^2 - \overline{\epsilon}^2}{S_y^2}. \tag{6.18}$$

Aïllant l'error mitjà de (6.18):

$$\overline{\epsilon}^2 = S_y^2(1 - R^2). \tag{6.19}$$

En el cas de tenir un ajust perfecte, $\overline{\epsilon}^2 = 0$ i $R^2 = 1$. Pel contrari, si les dades $x^{(\alpha)}$ no contribueixen a la predicció de y, trobarem $a^{(\alpha)} = 0$ i la predicció és reduirà a la constant $b = \overline{y}$ amb el que $V_E = 0$ i $R^2 = 0$.

6.7 Regressió univariable

En el cas $r = 1$ tenim una sola variable secundària X. L'estimador és la recta de regressió $\widehat{y} = ax+b$. El sistema (6.14) es redueix a una sola equació: $S_{x,x}a = S_{x,y}$, d'on

$$a = \frac{S_{x,y}}{S_x^2}. \tag{6.20}$$

$$b = \overline{y} - \frac{S_{x,y}}{S_x^2}\overline{x}. \tag{6.21}$$

Amb (6.16):

$$\overline{\epsilon}^2 = S_y^2 - \frac{S_{x,y}^2}{S_x^2}. \tag{6.22}$$

El coeficient de correlació generalitzat és

$$R^2 = \frac{S_{x,y}^2}{S_x^2 S_y^2}. \tag{6.23}$$

Notem que $R^2 = r^2$ on r és el **coeficient de correlació mostral** entre x i y:

$$r = \frac{S_{x,y}}{S_x S_y}. \tag{6.24}$$

r conté més informació que R ja que R sempre és positiu mentre que r té un signe que indica si la correlació és positiva (Y tendeix a créixer quan X creix) o negativa (Y tendeix a decréixer quan X creix). r es pot utilitzar per estudiar la relació de Y amb cadascuna de les variables secundàries $X^{(\alpha)}$ per separat mentre que R^2 dóna informació de la relació entre Y i tot el conjunt de variables secundàries.

Notem, finalment, que la recta de regressió es pot escriure:

$$\frac{y-\overline{y}}{S_y} = r\frac{x-\overline{x}}{S_x}. \tag{6.25}$$

DEM: La recta és $y = ax + b = ax + \overline{y} - a\overline{x}$ d'on $y - \overline{y} = a(x - \overline{x})$. De (6.20) i (6.24) resulta $a = r\frac{S_y}{S_x}$. ♣

Exemple 6.1 Aviat veurem que amb el programa R obté la regressió i les seves propietats sense que haguem de fer cap càlcul. En aquest exemple les dades són prou simples per poder fer els càlculs a mà. És important entendre d'on provenen tots els elements de la regressió per poder analitzar els resultats que ens dóna R, en cas de dubte, per poder programar la regressió en plataformes on no tinguem software addicional, etc.

Un sistema de comunicació transmet paquets de dades. El nombre d'errors al transmetre un paquet és aleatori i varia també amb la mida del paquet. Fem $n = 5$ proves amb paquets de diferents mides x_i i anomenem y_i al nombre d'errors obtinguts per a cada paquet, respectivament. Les dades resultants són:

i	1	2	3	4	5
x_i	1	2	3	5	6
y_i	1	3	2	4	5

A partir de la taula calculem:

$\sum_i x_i = 17$, $\sum_i x_i^2 = 75$, $\sum_i y_i = 15$, $\sum_i y_i^2 = 55$, $\sum_i x_i y_i = 63$.

Amb aquests sumatoris obtenim:

$\overline{x} = \frac{17}{5} = 3.4$, $\overline{y} = \frac{15}{5} = 3$.

$S_x^2 = \frac{75}{5} - \left(\frac{17}{5}\right)^2 = \frac{86}{25} = 3.44$, $S_y^2 = \frac{55}{5} - 3^2 = 2$, $S_{x,y} = \frac{55}{5} - \frac{17}{5}\cdot 3 = \frac{12}{5} = 2.4$.

La recta de regressió és $\widehat{y} = ax + b$ on

$$a = \frac{S_{x,y}}{S_x^2} = \frac{12/5}{86/25} = \frac{30}{43} = 0.6977,$$

$$b = \overline{y} - a\overline{x} = 3 - \frac{30}{43}\cdot\frac{17}{5} = \frac{27}{43} = 0.6279.$$

Pel coeficient de correlació:

$$R^2 = \frac{S_{x,y}^2}{S_x^2 S_y^2} = \frac{144/25}{(86/25)\cdot 2} = \frac{36}{43} = 0.8372$$

d'on $r = \frac{6}{\sqrt{43}} = 0.915$. L'ajust és bo si bé tenim poques dades i el valor de r tampoc és excessivament proper a 1.

Respecte al balanç de variacions:

$V_T = S_y^2 = 2$, $V_E = R^2 V_T = \dfrac{72}{43} = 1.6744$. $V_{NE} = V_T - V_E = \dfrac{14}{43} = 0.3256$. Així, l'error mitjà és $\overline{\epsilon}^2 = V_{NE} = 0.3256$.

Els valors estimats ($\widehat{y}_i = ax_i + b = \dfrac{30}{43}x_i + \dfrac{27}{43}$) i els residuals ($\epsilon_i = y_i - \widehat{y}_i$) es mostren en la següent taula:

i	1	2	3	4	5
x_i	1	2	3	5	6
y_i	1	3	2	4	5
$\widehat{y}_i$	1.3256	2.0233	2.7209	4.1163	4.8140
ϵ_i	−0.3256	0.9767	−0.7209	−0.1163	0.1860

♦

6.8 Regressió en R

La comanda principal per a fer regressions en el programa R és `lm()` (*linear model*). El seu argument utilitza vectors que contenen les mostres de les variables principal i secundàries, a través d'una sintaxi que descriu la forma de l'estimació. El resultat de `lm()` és una llista que conté, no només els coeficients de la regressió sino també propietats estadístiques d'aquests coeficients i de les variables residuals així com alguns tests de significació. Per a obtenir aquesta informació podem invocar directament `lm()` o aplicar-li comandes com `summary()` i `anova()`.

Mostrarem aquests aspectes sobre diversos exemples.

Exemple 6.2 Considerem l'evolució de la taxa d'atur i del PIB a Espanya els anys 2013, 2014 i 2015. El temps és la variable secundària (en unitats de trimestre). La taxa d'atur i el PIB (posant el nivell a zero a l'inici) les tractarem com variables principals ja que ens interessa la seva evolució en el temps. De *www.ine.es* obtenim:

t	13-1	13-2	13-3	13-4	14-1	14-2	14-3	14-4	15-1	15-2	15-3	15-4
Atur	26.9	26.1	25.7	25.7	25.9	24.5	23.7	23.7	23.8	22.4	21.2	20.9
PIB	0.0	-0.3	-0.2	0.0	0.4	0.9	1.5	2.2	3.1	4.1	4.9	5.7

El primer pas és entrar les dades en R i fer els gràfics corresponents. La visió dels gràfics ja indica la conveniència d'una forma de regressió determinada (ajust a una recta o potser un altre tipus de corba).

```
> n<-12
> t<-1:n
> atur<-c(26.9,26.1,25.7,25.7,25.9,24.5,23.7,23.7,23.8,22.4,21.2,20.9)
> pib<-c(0.0,-0.3,-0.2,0.0,0.4,0.9,1.5,2.2,3.1,4.1,4.9,5.7)
> plot(t,atur)
> plot(t,pib)
```

Les gràfiques resultants són:

En la gràfica de l'atur s'observa un efecte estacional. Fora d'aquest efecte un ajust lineal sembla correcte. El la gràfica del PIB la variació no sembla lineal. Començarem fent regressions lineals de la forma $at + b$ per ambdues variables. Després cercarem formes de millorar l'ajust.

```
> reg_atur<-lm(atur~t)
```

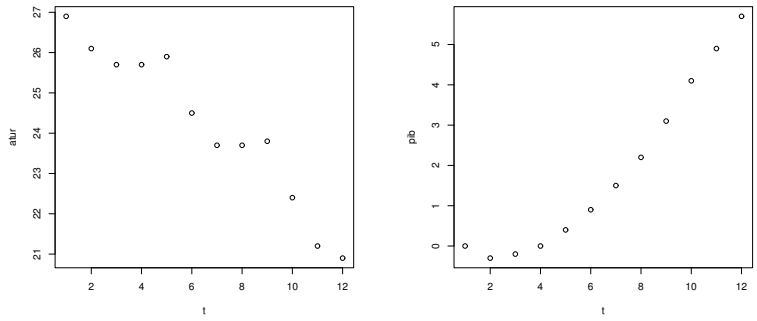

Figura 6.1: Atur i PIB en funció del temps.

```
> reg_pib<-lm(pib~t)
> plot(t,atur)
> abline(reg_atur,col="red")
> plot(t,pib)
> abline(reg_pib,col="red",lwd=2)
```

Ara les gràfiques inclouen la recta de regressió:

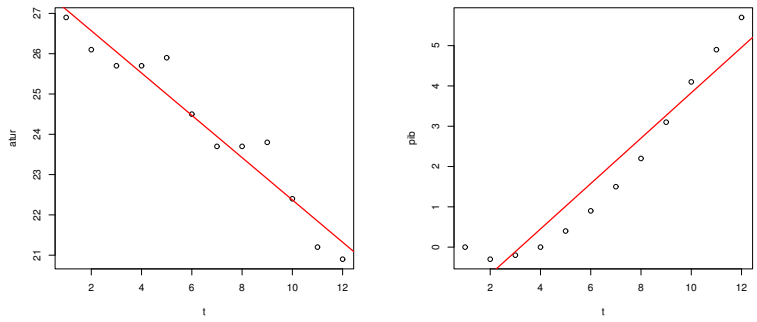

Figura 6.2: Rectes de regressió.

Notem que ara disposem del resultat de la regressió en forma de variables (`reg_atur` i `reg_pib`). La comanda `plot()` sempre obre una nova finestra. Altres comandes gràfiques permeten afegir contingut a la finestra ja oberta. Per exemple `abline()` afegeix una línia recta. El seu argument són les constants a i b (pendent i intercepte) que en aquest cas s'extreuen de la variable que conté la regressió. Utilitzem també altre arguments per fixar el color i gruix de la recta (`col` i `lwd`).

Quan invoquem les variables que contenen la regressió obtenim la forma de regressió que hem demanat i els coeficients de la regressió. La constant b és `Intercept` i els pendents apareixen junt al nom de la variable que multipliquen:

```
> reg_atur

Call:
lm(formula = atur ~ t)

Coefficients:
(Intercept)            t
    27.6197      -0.5248
```

```
> reg_pib

Call:
lm(formula = pib ~ t)

Coefficients:
(Intercept)            t
    -1.8030       0.5633
```

Informació més àmplia s'obté amb la comanda `summary()`:

```
> summary(reg_atur)

Call:
lm(formula = atur ~ t)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64662 -0.36436 -0.08316  0.20443  0.90443

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.6197     0.3246    85.1 1.23e-15 ***
t            -0.5248     0.0441   -11.9 3.16e-07 ***
---
Signif. codes:  0 ‘***´ 0.001 ‘**´ 0.01 ‘*´ 0.05 ‘.´ 0.1 ‘ ´ 1

Residual standard error: 0.5274 on 10 degrees of freedom
Multiple R-squared:  0.9341,    Adjusted R-squared:  0.9275
F-statistic: 141.6 on 1 and 10 DF,  p-value: 3.158e-07

> summary(reg_pib)

Call:
lm(formula = pib ~ t)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6767 -0.5308 -0.1267  0.4091  1.2397

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.80303    0.40341  -4.469   0.0012 **
t            0.56329    0.05481  10.276 1.24e-06 ***
---
Signif. codes:  0 ‘***´ 0.001 ‘**´ 0.01 ‘*´ 0.05 ‘.´ 0.1 ‘ ´ 1

Residual standard error: 0.6555 on 10 degrees of freedom
Multiple R-squared:  0.9135,    Adjusted R-squared:  0.9048
F-statistic: 105.6 on 1 and 10 DF,  p-value: 1.237e-06
```

♦

En aquest cas obtenim:

- la forma demanada de regressió,
- algunes propietats estadístiques del conjunt de residuals ϵ_i,
- els coeficients de la regressió $a^{(\alpha)}$ i b amb alguns valors referits al test de la hipòtesi que valguin zero,
- l'error mitjà $\overline{\epsilon}$,
- el coeficient de correlació generalitzat R^2 (més una versió ajustada d'aquest coeficient)
- un test de la significació del model en conjunt (test F).

En els nostres exemples les rectes de regressió per l'atur i el PIB donen un bon ajust ($R^2 = 0.93$, i $R^2 = 0.91$ respectivament) si bé observant les gràfiques notem una estructura diferent, fet que podem utilitzar per millorar la regressió.

6.9 Altres formes de la regressió

En l'anterior exemple, el PIB mostra un tipus de variació que no es correspon amb una recta. Quan fem regressió lineal, la linealitat radica en la dependència en els paràmetres $a^{(\alpha)}$ i b. Podem utilitzar la mateixa metodologia si l'estimador és una combinació lineal de funcions de les nostres dades estadístiques.

Per exemple, podem utilitzar un ajust parabòlic:

$$\widehat{y} = ax^2 + bx + c. \tag{6.26}$$

El tractament és el mateix que hem fet abans considerant dues variables secundàries $x^{(1)} = x^2$ i $x^{(2)} = x$.

De manera similar podem ajustar a polinomis de grau més alt.

Un altre tipus d'ajust comú és l'exponencial:

$$\widehat{y} = be^{ax}. \tag{6.27}$$

Ara la dependència en a i b no és lineal però podem prendre logaritmes als dos costats:

$$\ln \widehat{y} = ax + \ln b. \tag{6.28}$$

Llavors definim $z = \ln y$, fem la regressió lineal $\widehat{z} = ax + b'$ amb el que tenim l'ajust (6.24) fent $b = e^{b'}$.

Amb la mateixa transformació podem tractar l'ajust potencial:

$$\widehat{y} = bx^a. \tag{6.29}$$

$$\ln \widehat{y} = a \ln x + \ln b. \tag{6.30}$$

Llavors definim $z = \ln y$ i $t = \ln x$, fem la regressió lineal $\widehat{z} = at + b'$ amb el que tenim l'ajust (6.24) fent $b = e^{b'}$.

Exemple 6.3 Seguint l'exemple anterior, millorarem l'ajust del PIB amb una regressió parabòlica:

```
> reg2_pib<-lm(pib~poly(t,2,raw=TRUE))
> summary(reg2_pib)

Call:
lm(formula = pib ~ poly(t, 2, raw = TRUE))

Residuals:
     Min       1Q   Median       3Q      Max
-0.26511 -0.08783 -0.01643  0.07259  0.25182

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.134091   0.174351  -0.769   0.4615
poly(t, 2, raw = TRUE)1 -0.151973   0.061664  -2.465   0.0359 *
poly(t, 2, raw = TRUE)2  0.055020   0.004618  11.915 8.18e-07 ***
---
Signif. codes:  0 ‘***´ 0.001 ‘**´ 0.01 ‘*´ 0.05 ‘.´ 0.1 ‘ ´ 1

Residual standard error: 0.1687 on 9 degrees of freedom
Multiple R-squared:  0.9948,    Adjusted R-squared:  0.9937
F-statistic: 868.2 on 2 and 9 DF,  p-value: 5.077e-11

> co<-coef(reg2_pib)
> par<-co[1]+co[2]*t+co[3]*t^2
> plot(t,pib)
> lines(par,col="red",lwd=2)
```

Notem que no cal definir una nova variable que contingui t^2; la sintaxi de l'argument de `lm()` permet molts tipus d'ajust. En aquest cas demanen un ajust polinòmic de grau 2. El paràmetre `raw=TRUE` força l'ús de polinomis ordinaris (si no, s'utilitzarien polinomis ortogonals).

Com es veu, ara $R^2 = 0.9948$ millorant el valor anterior 0.9135. Per fer la gràfica de la paràbola hem copiat els coeficients de l'ajust en una variable, `co`, i els hem utilitzat per avaluar la paràbola de regressió en els diferents valors de t. El resultat forma el vector `par`. La comanda gràfica `lines` dibuixa la poligonal que uneix aquests punts sobre la gràfica que ja està oberta. El resultat és la figura 6.3.

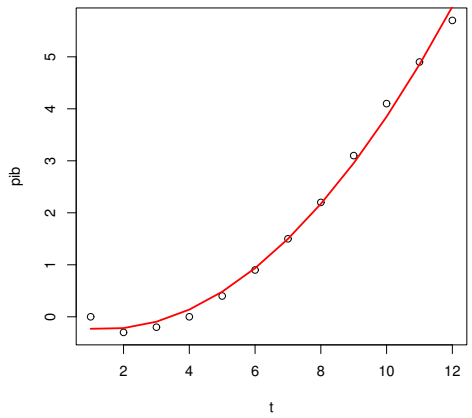

Figura 6.3: Paràbola de regressió.

Vist el resultat, podem tenir la temptació de ajustar amb polinomis de grau més alt. Per exemple, amb grau 6:

```
> reg6_pib<-lm(pib~poly(t,6,raw=TRUE))
> co<-coef(reg6_pib)
> par<-co[1]+co[2]*t+co[3]*t^2+co[4]*t^3+co[5]*t^4+co[6]*t^5+co[7]*t^6
> plot(t,pib)
> lines(t,par,col="red",lwd=2)
```

s'obté $R^2 = 0.9998$ i la gràfica 6.4.

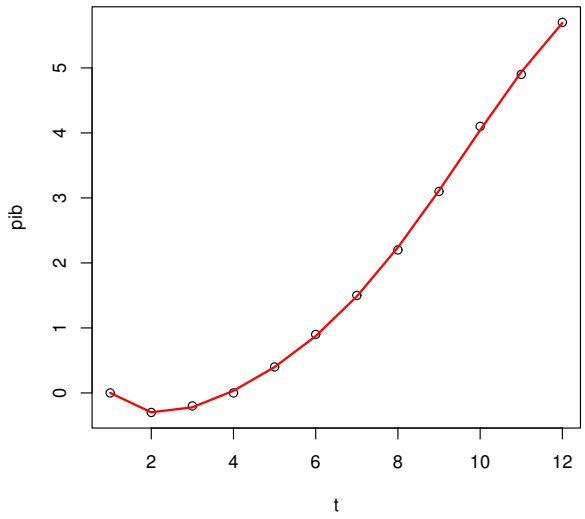

Figura 6.4: Sobreajust, polinomi de grau 6.

La millora no té massa sentit ja que la obtenim al cost d'introduir moltes constants. Sempre podríem tenir un ajust perfecte amb un polinomi de grau prou alt que passés per tots els punts. Aquestes situacions s'anomenen **sobreajust** (*overfitting*) i s'han d'evitar. No hi ha un criteri absolut de com s'ha de fer l'ajust. En general ha de primar la simplicitat del model. Hem d'entendre que el model lineal separa la variable principal en una part explicada i un residu que, si el model és correcte, ja podem considerar com "soroll". La part de soroll és pròpiament aleatòria i no s'ha d'intentar ajustar. ♦

Capítol 7

Anàlisi de la regressió

En el tema anterior hem formulat la regressió a partir de dades estadístiques. La idea era obtenir un equivalent de l'estimació sobre variables aleatòries, $\widehat{Y} = E[Y|X^{(1)}, X^{(2)}, \ldots, X^{(r)}]$, a partir de la informació que ens donen les mostres. El cas principal és el de l'estimació lineal on suposem que l'esperança anterior és una combinació lineal de les variables secundàries. Aquesta és la situació quan el conjunt de totes les variables es de tipus normal.

En qualsevol cas, suposem que l'estimador anterior té forma lineal. Recordem que això pot incloure dependències de tipus polinòmic, exponencial, etc. Recordem també que les variables secundàries poden ser paràmetres de manera que al prendre les mostres, decidim nosaltres els seus valors, o aquests no són aleatoris.

En aquest tema analitzem la relació entre els coeficients de la regressió estadística i els coeficients de l'estimació a nivell de variables aleatòries. Això inclou: la seva significació, intervals de confiança per als coeficients i per als valors predits, etc.

7.1 Hipòtesis i notació

En el que segueix suposarem:

- La resposta Y és normal (les variables y_i són normals).
- Les variables $x^{(\alpha)}$ són paràmetres no aleatoris.
- $$E[y_i] = \sum_{\gamma=1}^{r} \alpha^{(\gamma)} x_i^{(\gamma)} + \beta. \tag{7.1}$$
- La resposta té variància constant σ^2. És a dir, aquesta no depèn dels valors de les variables $x^{(\gamma)}$:
$$V[y_i] = \sigma^2. \tag{7.2}$$

Calculant la regressió obtenim els coeficients $a^{(\gamma)}$, $\gamma = 1, \cdots, r$ i b. Observem que són combinació lineal de les y_i i, per tant, són estadístics de tipus normal. Llavors, $\widehat{y}_i = \sum_{\gamma} a^{(\gamma)} x_i^{(\gamma)} + b$ també són variables normals.

Notem que
$$\widehat{y}_i = \sum_{\gamma} a^{(\gamma)} (x_i^{(\gamma)} - \overline{x}^{(\gamma)}) + \overline{y}. \tag{7.3}$$

Introduim la notació:

$$\Delta x_i^{(\gamma)} = x_i^{(\gamma)} - \overline{x}^{(\gamma)}. \tag{7.4}$$

Es verifica

$$\sum_{i=1}^{n} \Delta x_i^{(\gamma)} = 0. \tag{7.5}$$

Ara podem escriure:

$$S_{x^{(\gamma)},x^{(\delta)}} = \frac{1}{n}\sum_{i=1}^{n} \Delta x_i^{(\gamma)} \Delta x_i^{(\delta)}, \qquad S_{x^{(\gamma)},y} = \frac{1}{n}\sum_{i=1}^{n} \Delta x_i^{(\gamma)} y_i. \tag{7.6}$$

$$\widehat{y}_i = \sum_{\gamma} a^{(\gamma)} \Delta x_i^{(\gamma)} + \overline{y} \tag{7.7}$$

En la segona relació de (7.6) $\overline{y}$ no apareix degut a (7.5). Similarment, $S_{x^{(\gamma)},x^{(\delta)}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \Delta x_i^{(\gamma)} x_i^{(\delta)}$.

7.2 Propietats estadístiques dels coeficients de regressió

Amb les hipòtesis de l'apartat anterior, tractem els nombres $x_i^{(\gamma)}$ com a constants i resulta que els coeficients de la regressió $a^{(\gamma)}$ i b són funcions de les variables aleatòries y_i. En aquesta secció calcularem els seus paràmetres. Amb ell posteriorment podrem determinar intervals de confiança per alguns valors associats a la regressió.

Les equacions (6.13) i (6.14) permeten calcular $a^{(\gamma)}$ i b. El caràcter aleatori prové de que els coeficients $\overline{y}$ i $S_{x^{(\gamma)},y}$ de les equacions depenen de les variables y_i. Aquests coeficients són combinació lineal de les y_i així que es tracta de variables normals. En quant als seus paràmetres:

$$E[\overline{y}] = \sum_{\gamma=1}^{r} \alpha^{(\gamma)} \overline{x}^{(\gamma)} + \beta, \qquad E[S_{x^{(\gamma)},y}] = \sum_{\delta=1}^{r} S_{x^{(\gamma)},x^{(\delta)}} \alpha^{(\delta)}. \tag{7.8}$$

DEM: Amb (6.2), (7.6) i (7.1): $E[\overline{y}] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} E[y_i] = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{\gamma} \alpha^{(\gamma)} x_i^{(\gamma)} + \beta\right) = \sum_{\gamma} \alpha^{(\gamma)} \overline{x}^{(\gamma)} + \beta.$

$$E[S_{x^{(\gamma)},y}] = \frac{1}{n}\sum_{i=1}^{n} \Delta x_i^{(\gamma)} E[y_i] = \frac{1}{n}\sum_{i=1}^{n} \Delta x_i^{(\gamma)} \left(\sum_{\delta} \alpha^{(\delta)} x_i^{(\delta)} + \beta\right) = \sum_{\delta} S_{x^{(\gamma)},x^{(\delta)}} \alpha^{(\delta)}. \ \clubsuit$$

Els coeficients de la regressió també són variables normals ja que

$$a^{(\gamma)} = \sum_{\delta=1}^{r} S^{-1}_{x^{(\gamma)},x^{(\delta)}} S_{x^{(\delta)},y}, \qquad b = \overline{y} - \sum_{\gamma=1}^{r} a^{(\gamma)} \overline{x}^{(\gamma)}. \tag{7.9}$$

Ara podem demostrar que $a^{(\gamma)}$ i b són estimadors centrats de $\alpha^{(\gamma)}$ i β respectivament.

$$E[a^{(\gamma)}] = \alpha^{(\gamma)}, \qquad E[b] = \beta. \tag{7.10}$$

DEM: $E[a^{(\gamma)}] = \displaystyle\sum_{\beta} S^{-1}_{x^{(\gamma)},x^{(\delta)}} E[S_{x^{(\delta)},y}] = \sum_{\beta}\sum_{\delta} S^{-1}_{x^{(\gamma)},x^{(\delta)}} S_{x^{(\delta)},x^{(\delta)}} \alpha^{(\delta)} = \alpha^{(\gamma)}$. Amb (6.13): $E[b] = E[\overline{y}] - \displaystyle\sum_{\gamma} E[a^{(\gamma)}]\overline{x}^{(\gamma)} = \sum_{\gamma} \alpha^{(\gamma)}\overline{x}^{(\gamma)} + \beta - \sum_{\gamma} \alpha^{(\gamma)}\overline{x}^{(\gamma)} = \beta.$ $\clubsuit$

Considerem ara els paràmetres de segon ordre. Utilitzant (7.2) i la independència de les y_i:

$$C[y_i, y_j] = \sigma^2 \delta_{i,j} \tag{7.11}$$

on $\delta_{i,j} = 0$ per $i \neq j$ i $\delta_{i,i} = 1$.

$$C[\overline{y}, y_i] = \frac{\sigma^2}{n}, \qquad C[S_{x^{(\gamma)},y}, y_i] = \frac{\sigma^2}{n}\Delta x_i^{(\gamma)}. \tag{7.12}$$

DEM: $C[\overline{y}, y_i] = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} C[y_j, y_i] = \frac{\sigma^2}{n}$. $C[S_{x^{(\gamma)},y}, y_i] = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} \Delta x_j^{(\gamma)} C[y_j, y_i] = \frac{\sigma^2}{n}\Delta x_i^{(\gamma)}$. ♣

$$C[\overline{y}, \overline{y}] = \frac{\sigma^2}{n}, \qquad C[\overline{y}, S_{x^{(\gamma)},y}] = 0, \qquad C[S_{x^{(\gamma)},y}, S_{x^{(\delta)},y}] = \frac{\sigma^2}{n} S_{x^{(\gamma)},x^{(\delta)}}. \tag{7.13}$$

DEM: $C[\overline{y}, \overline{y}] = V[\overline{y}] = \dfrac{\sigma^2}{n}$ (utilitzant (3.3)). $C[\overline{y}, S_{x^{(\gamma)},y}] = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} C[y_i, S_{x^{(\gamma)},y}] = \frac{\sigma^2}{n^2}\sum_{i=1}^{n} \Delta x_i^{(\gamma)} = 0.$

$$C[S_{x^{(\gamma)},y}, S_{x^{(\delta)},y}] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta x_i^{(\gamma)} \Delta x_j^{(\delta)} C[y_i, y_j] = \frac{1}{n^2}\sum_{i=1}^{n} \Delta x_i^{(\gamma)} \Delta x_i^{(\delta)} \sigma^2 = \frac{\sigma^2}{n} S_{x^{(\gamma)},x^{(\delta)}}. \text{ ♣}$$

Notem que $\overline{y}$ és independent de $S_{x^{(\gamma)},y}$ (normals incorrelades). Atès (7.11) també tenim

$$C[\overline{y}, a^{(\gamma)}] = 0, \qquad C[\overline{y}, b] = \frac{\sigma^2}{n}. \tag{7.14}$$

Finalment, per als coeficients de regressió:

$$C[a^{(\gamma)}, a^{(\delta)}] = \frac{\sigma^2}{n} S^{-1}_{x^{(\gamma)},x^{(\delta)}}, \tag{7.15}$$

$$C[a^{(\gamma)}, b] = -\frac{\sigma^2}{n}\sum_{\delta=1}^{r} S^{-1}_{x^{(\gamma)},x^{(\delta)}} \overline{x}^{(\delta)}, \tag{7.16}$$

$$C[b, b] = \frac{\sigma^2}{n}\left(1 + \sum_{\gamma,\delta=1}^{r} \overline{x}^{(\gamma)} S^{-1}_{x^{(\gamma)},x^{(\delta)}} \overline{x}^{(\delta)}\right). \tag{7.17}$$

DEM: A partir de (7.9) i (7.11):

$$C[a^{(\gamma)}, a^{(\delta)}] = \sum_{\mu=1}^{r}\sum_{\nu=1}^{r} S^{-1}_{x^{(\gamma)},x^{(\mu)}} S^{-1}_{x^{(\delta)},x^{(\nu)}} C[S_{x^{(\mu)},y}, S_{x^{(\nu)},y}] = \sum_{\mu=1}^{r}\sum_{\nu=1}^{r} S^{-1}_{x^{(\gamma)},x^{(\mu)}} S^{-1}_{x^{(\delta)},x^{(\nu)}} \frac{\sigma^2}{n} S_{x^{(\mu)},x^{(\nu)}} = \frac{\sigma^2}{n} S^{-1}_{x^{(\gamma)},x^{(\delta)}}.$$

$$C[a^{(\gamma)}, b] = C[a^{(\gamma)}, \overline{y} - \sum_{\delta=1}^{r} a^{(\delta)} \overline{x}^{(\delta)}] = -\sum_{\delta=1}^{r} C[a^{(\gamma)}, a^{(\delta)}] \overline{x}^{(\delta)} = -\frac{\sigma^2}{n}\sum_{\delta=1}^{r} S^{-1}_{x^{(\gamma)},x^{(\delta)}} \overline{x}^{(\delta)}.$$

$$C[b, b] = C[\overline{y} - \sum_{\delta=1}^{r} a^{(\delta)} \overline{x}^{(\delta)}, b] = \frac{\sigma^2}{n} - \sum_{\delta=1}^{r} \overline{x}^{(\delta)} C[a^{(\delta)}, b] = \frac{\sigma^2}{n} + \frac{\sigma^2}{n}\sum_{\gamma,\delta=1}^{r} \overline{x}^{(\delta)} S^{-1}_{x^{(\delta)},x^{(\gamma)}} \overline{x}^{(\gamma)}. \text{ ♣}$$

A (7.15) i (7.17) veiem que $a^{(\gamma)}$ i b són estimadors consistents de $\alpha^{(\gamma)}$ i β, respectivament, ja que les seves variàncies tendeixen a zero quan $n \to \infty$.

7.3 Propietats estadístiques dels residuals (errors)

Els errors o residuals són les diferencies entre els valors obtinguts y_i i els valors estimats segons la regressió $\widehat{y}_i$:

$$\epsilon_i = y_i - \widehat{y}_i = y_i - \overline{y} - \sum_{\gamma=1}^{r} a^{(\gamma)} \Delta x_i^{(\gamma)}. \tag{7.18}$$

De (7.10) i (7.1) deduïm:

$$E[\widehat{y}_i] = E\left[\sum_{\gamma=1}^{r} a^{(\gamma)} x_i^{(\gamma)} + b\right] = \sum_{\gamma=1}^{r} \alpha^{(\gamma)} x_i^{(\gamma)} + \beta = E[y_i], \tag{7.19}$$

d'on

$$E[\epsilon_i] = 0. \tag{7.20}$$

Considerem ara propietats de segon ordre. Alguns resultats auxiliars:

$$C[a^{(\gamma)}, y_i] = C[a^{(\gamma)}, \widehat{y}_i] = \frac{\sigma^2}{n} \sum_{\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \Delta x_i^{(\delta)}. \tag{7.21}$$

$$C[y_i, \widehat{y}_j] = C[\widehat{y}_i, \widehat{y}_j] = \frac{\sigma^2}{n} \left(1 + \sum_{\gamma,\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \Delta x_i^{(\gamma)} \Delta x_i^{(\delta)}\right). \tag{7.22}$$

DEM:

$$C[a^{(\gamma)}, y_i] = \sum_{\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \frac{1}{n} \sum_{j=1}^{n} \Delta x_j^{(\delta)} C[y_j, y_i] = \frac{\sigma^2}{n} \sum_{\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \Delta x_i^{(\delta)}.$$

$$C[a^{(\gamma)}, \widehat{y}_i] = C[a^{(\gamma)}, \overline{y} + \sum_{\delta=1}^{r} a^{(\delta)} \Delta x_i^{(\delta)}] = \sum_{\delta=1}^{r} C[a^{(\gamma)}, a^{(\delta)}] \Delta x_i^{(\delta)} = \frac{\sigma^2}{n} \sum_{\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \Delta x_i^{(\delta)}.$$

$$C[\widehat{y}_i, \widehat{y}_j] = C[\overline{y} + \sum_{\delta=1}^{r} a^{(\delta)} \Delta x_i^{(\delta)}, \overline{y} + \sum_{\gamma=1}^{r} a^{(\gamma)} \Delta x_j^{(\gamma)}] = C[\overline{y}, \overline{y}] + \sum_{\delta,\gamma=1}^{r} C[a^{(\delta)}, a^{(\gamma)}] \Delta x_i^{(\delta)} \Delta x_j^{(\gamma)}$$

$$= \frac{\sigma^2}{n} + \sum_{\gamma,\delta=1}^{r} \frac{\sigma^2}{n} S^{-1}_{x^{(\delta)}, x^{(\gamma)}} \Delta x_i^{(\delta)} \Delta x_j^{(\gamma)}.$$

$$C[y_i, \widehat{y}_j] = C[y_i, \overline{y} + \sum_{\delta=1}^{r} a^{(\delta)} \Delta x_j^{(\delta)}] = \frac{\sigma^2}{n} + \sum_{\gamma=1}^{r} \frac{\sigma^2}{n} \sum_{\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \Delta x_i^{(\gamma)} \Delta x_j^{(\delta)}. \clubsuit$$

Ara podem obtenir:

$$V[\epsilon_i] = \frac{\sigma^2}{n} \left(n - 1 - \sum_{\gamma,\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \Delta x_i^{(\gamma)} \Delta x_i^{(\delta)}\right). \tag{7.23}$$

I per a $i \neq j$:

$$C[\epsilon_i, \epsilon_j] = -\frac{\sigma^2}{n} \left(1 + \sum_{\gamma,\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \Delta x_i^{(\gamma)} \Delta x_j^{(\delta)}\right). \tag{7.24}$$

DEM: $C[\epsilon_i, \epsilon_j] = C[y_i - \widehat{y}_i, y_j - \widehat{y}_j] = C[y_i - \widehat{y}_i, y_j] = \sigma^2 \delta_{i,j} - \frac{\sigma^2}{n} \left(1 + \sum_{\gamma,\delta=1}^{r} S^{-1}_{x^{(\gamma)}, x^{(\delta)}} \Delta x_i^{(\gamma)} \Delta x_i^{(\delta)}\right). \clubsuit$

Els residuals ϵ_i són independents dels coeficients $a^{(\gamma)}$ i dels valors estimats $\widehat{y}_j$:

$$C[\epsilon_i, a^{(\gamma)}] = 0, \qquad C[\epsilon_i, \widehat{y}_j] = 0. \tag{7.25}$$

DEM: A partir de (7.21) i (7.22): $C[\epsilon_i, a^{(\gamma)}] = C[y_i - \widehat{y}_i, a^{(\gamma)}] = C[y_i, a^{(\gamma)}] - C[\widehat{y}_i, a^{(\gamma)}] = 0$. $C[\epsilon_i, \widehat{y}_j] = C[y_i - \widehat{y}_i, \widehat{y}_j] = C[y_i, \widehat{y}_j] - C[\widehat{y}_i, \widehat{y}_j] = 0$. ♣

Notem ara que

$$E\left[\sum_{i=1}^{n} \epsilon_i^2\right] = \sum_{i=1}^{n} E[\epsilon_i^2] = \sum_{i=1}^{n} V[\epsilon_i] = \frac{\sigma^2}{n}\left((n-1)n - \sum_{\gamma,\delta=1}^{r} S^{-1}_{x^{(\gamma)},x^{(\delta)}} \sum_{i=1}^{n} \Delta x_i^{(\gamma)} \Delta x_i^{(\delta)}\right)$$

$$= \sigma^2\left(n - 1 - \sum_{\gamma,\delta=1}^{r} S^{-1}_{x^{(\gamma)},x^{(\delta)}} S_{x^{(\gamma)},x^{(\delta)}}\right) = \sigma^2(n-1-r).$$

Llavors, obtenim un estimador S^2 de σ^2 definint:

$$S^2 = \frac{1}{n-1-r}\sum_{i=1}^{n} \epsilon_i^2. \tag{7.26}$$

De (7.25) es dedueix que S^2 és independent dels coeficients $a^{(\gamma)}$ i dels valors $\widehat{y}_i$.

7.4 Tests de significació del model

Hi ha dos tipus de test per a determinar la significació del model.

Un és sobre la rellevància de cada variable secundària. Es tracta de la hipòtesi nul·la $H_0 = \{\alpha^{(\gamma)} = 0\}$. Com s'ha vist a (7.10) i (7.15):

$$a^{(\gamma)} \sim \mathrm{N}\left(\alpha^{(\gamma)}, \sigma\sqrt{\frac{S^{-1}_{x^{(\gamma)},x^{(\gamma)}}}{n}}\right)$$

Normalment no coneixem σ i l'estimem amb S. Llavors, fem el test amb la variable

$$T = \frac{a^{(\gamma)} - \alpha^{(\gamma)}}{S\sqrt{\frac{S^{-1}_{x^{(\gamma)},x^{(\gamma)}}}{n}}},$$

que és d'Student amb $n-r-1$ graus de llibertat. Sota la hipòtesi nul·la calculem $T = \dfrac{a^{(\gamma)}}{S\sqrt{\frac{S^{-1}_{x^{(\gamma)},x^{(\gamma)}}}{n}}}$ i obtenim un valor P que si és prou petit indica la rellevància de la variable corresponent.

L'altre tipus de test es basa en els mètodes d'anàlisi de variància (ANOVA).

Com hem vist a (6.17):

$$\sum_i (y_i - \overline{y})^2 = \sum_i (y_i - \widehat{y}_i)^2 + \sum_i (\widehat{y}_i - \overline{y})^2 \tag{7.27}$$

Definim $SS_T = \sum_i (y_i - \overline{y})^2$, $SS_{NE} = \sum_i (y_i - \widehat{y}_i)^2$, $SS_E = \sum_i (\widehat{y}_i - \overline{y})^2$.

Si dividim per σ^2 es tracta de variables khi quadrat. SS_T correspon a la variació total amb $n-1$ graus de llibertat. SS_{NE} i SS_E donen varibles khi quadrat amb $n-r-1$ i r graus de llibertat, respectivament.

Definim $MS_{NE} = \dfrac{SS_{NE}}{n-r-1}$ (és a dir, S^2, definit a (7.26)), $MS_E = \dfrac{SS_E}{r}$. Per a la significació del model considerem la variable

$$F = \frac{MS_E}{MS_{NE}}$$

que és de Fisher amb $r, n-r-1$ graus de llibertat. La probabilitat a la dreta del valor obtingut per a F ens dóna el valor P que hauria de ser petit si el model és significant.

Exemple 7.1 Considerem (X, Y) on $X \sim \mathrm{N}(3, 2)$ i $Y = 2X + W$ on $W \sim \mathrm{N}(0, 3)$ independent de X. Llavors $\mu_X = 3$, $\mu_Y = 6$, $\sigma_X = 2$, $\sigma_Y = 5$ (ja que $V[Y] = 4V[X] + V[W] = 25$), $C[X, Y] = 8$, $\rho = \dfrac{8}{2 \cdot 5} = 0.8$.

$E[Y|X] = 2X$ d'on $\alpha = 2$, $\beta = 0$.

Ara generem una mostra de mida 50 i observem els coeficients de regressió, la seva significació i la significació del model.

```
> n<-50
> x<-rnorm(n,3,2)
> y<-2*x+rnorm(n,0,3)
> reg<-lm(y~x)
> plot(x,y)
> # Dibuixem la regressió obtinguda a partir de les dades
> abline(reg,col="red",lwd=2)
> # Dibuixem la línia exacta de regressió:
> abline(0,2,col="orange",lwd=2)
> summary(reg)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9426 -1.9421  0.1022  2.5702  6.8404

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7381     0.8590   0.859    0.394
x             1.6173     0.2156   7.503 1.25e-09 ***
---
Signif. codes:  0 ‘***´ 0.001 ‘**´ 0.01 ‘*´ 0.05 ‘.´ 0.1 ‘ ´ 1

Residual standard error: 3.785 on 48 degrees of freedom
Multiple R-squared:  0.5397,    Adjusted R-squared:  0.5301
F-statistic: 56.29 on 1 and 48 DF,  p-value: 1.253e-09
```

Hem obtingut $a = 1.6173$, $b = 0.7381$. Notem que b té un valor P alt (0.39) que indica que podem accepar la hipòtesi $\beta = 0$. El valor P associat a a és molt petit (10^{-9}) fet que permet descartar la possibilitat que α valgués 0 i la variable X fos irrellevant per a l'estimació de Y.

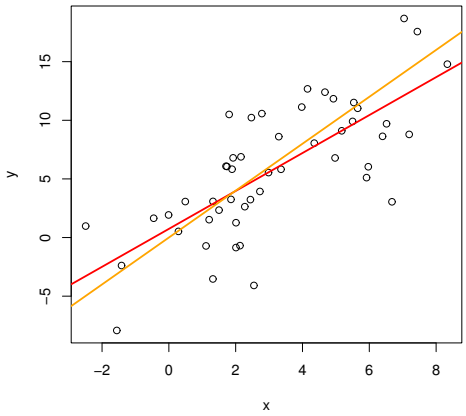

Figura 7.1: En vermell, la recta de regressió estadística $y = ax + b$. en taronja, la recta de regressió exacta $y = \alpha x + \beta$.

Notem que $R^2 = 0.5397$. En aquest cas la variància de Y, fixada X és $\sigma = 3$. Aquesta dispersió es mostra en el valor de R^2 lluny de 1. Notem que $R = 0.7346$ és una estimació del valor exacte $\rho = 0.8$.

El model és significant en conjunt ja que el valor P del test F és petit (10^{-9}). Podem detallar-ho més amb la taula ANOVA:

```
> anv<-anova(reg)
> anv
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 806.55  806.55  56.288 1.253e-09 ***
Residuals 48 687.79   14.33
---
Signif. codes:  0 ‘***´ 0.001 ‘**´ 0.01 ‘*´ 0.05 ‘.´ 0.1 ‘ ´ 1
```

Les variacions V_T, V_E, V_{NE} s'obtenen d'aquesta taula ja que la segona columna conté les sumes de quadrats SS_E, SS_{NE}:

```
> Vt<-var(y)*(n-1)/n
> Vt
[1] 29.88694
> Ve<-anv[1,2]/n
> Ve
[1] 16.13105
> Vne<-anv[2,2]/n
> Vne
[1] 13.75589
> Ve+Vne
[1] 29.88694
```

♦

7.5 Intervals per a la regressió i per a la predicció

7.5.1 Intervals per a la regressió

Amb la regressió obtenim un estimador (6.8) que considerem ara sobre valors arbitraris de les $x^{(\gamma)}$. Fixem-nos en un punt qualsevol, possiblement diferent dels utilitzats en la mostra, $x_p^{(\gamma)}$, $\gamma = 1, \cdots, r$. El nostre estimador en aquest punt ens dóna un valor

$$\widehat{y}_p = \sum_\gamma a^{(\gamma)} x_p^{(\gamma)} + b = \sum_\gamma a^{(\gamma)} (x_p^{(\gamma)} - \overline{x}^{(\gamma)}) + \overline{y}. \tag{7.28}$$

$\widehat{y}_p$ és una estimació de la posició exacta de la funció de regressió (6.9):

$$\mu_p = \sum_\gamma \alpha^{(\gamma)} x_p^{(\gamma)} + \beta. \tag{7.29}$$

Si obtenim un interval de confiança per a μ_p tindrem una mesura de l'error en la posició de la funció de regressió. Notem que $\widehat{y}_p$ és una variable normal i:

$$E[\widehat{y}_p] = \mu_p, \qquad V[\widehat{y}_p] = \frac{\sigma^2}{n} \left(1 + \sum_{\gamma,\delta=1}^{r} (x_p^{(\gamma)} - \overline{x}^{(\gamma)}) S^{-1}_{x^{(\gamma)},x^{(\delta)}} (x_p^{(\delta)} - \overline{x}^{(\delta)}) \right). \tag{7.30}$$

DEM: La primera és immediata, prenent esperances en (7.28). Per a la segona, en ser $\overline{y}$ i $a^{(\gamma)}$ incorrelades:

$$V[\widehat{y}_p] = V[\overline{y}] + \sum_{\gamma,\delta=1}^{r} (x_p^{(\gamma)} - \overline{x}^{(\gamma)}) C[a^{(\gamma)}, a^{(\delta)}] (x_p^{(\delta)} - \overline{x}^{(\delta)}),$$

d'on s'obté el resultat, aplicant (7.13) i (7.15). ♣

Si coneixem σ, obtenim l'interval de la manera habitual (4.21). Si no la coneixem la podem estimar a través de S i tenim l'interval en la forma (4.24):

$$\widehat{y}_p \pm t_c \frac{S}{\sqrt{n}} \sqrt{1 + \sum_{\gamma,\delta=1}^{r} (x_p^{(\gamma)} - \overline{x}^{(\gamma)}) S^{-1}_{x^{(\gamma)},x^{(\delta)}} (x_p^{(\delta)} - \overline{x}^{(\delta)})} \tag{7.31}$$

on t_c s'obté de la t d'Student amb $n - r - 1$ graus de llibertat.

7.5.2 Intervals per a la predicció

Com abans, fixem valors $x_p^{(\gamma)}$ i cerquem intervals de confiança pel valor de la variable Y corresponent. Aquest problema és diferent de l'estimació de paràmetres que hem considerat fins ara en estadística i també és diferent de l'estimació de variables aleatòries feta en la part de variables aleatòries ja que ara tampoc coneixem els paràmetres.

El procediment es basa en la variable $Y - \widehat{y}_p$. Y és independent de les variables y_i de la mostra. Tenim:

$$E[Y - \widehat{y}_p] = 0, \qquad V[Y - \widehat{y}_p] = \sigma^2 + \frac{\sigma^2}{n} \left(1 + \sum_{\gamma,\delta=1}^{r} (x_p^{(\gamma)} - \overline{x}^{(\gamma)}) S^{-1}_{x^{(\gamma)},x^{(\delta)}} (x_p^{(\delta)} - \overline{x}^{(\delta)}) \right) \tag{7.32}$$

DEM: $E[Y - \widehat{y}_p] = E[Y] - E[\widehat{y}_p] = \mu_p - \mu_p = 0$. $V[Y - \widehat{y}_p] = V[Y] + V[\widehat{y}_p] = \sigma^2 + V[\widehat{y}_p]$. Aquesta última variància és a (7.30). Notem que $V[Y]$ val σ^2 ja que Y està condicionada als valors donats de $x_p^{(\delta)}$. ♣

L'interval és

$$\widehat{y}_p \pm t_c S \sqrt{\frac{n+1}{n} + \frac{1}{n} \sum_{\gamma,\delta=1}^{r} (x_p^{(\gamma)} - \overline{x}^{(\gamma)}) S^{-1}_{x^{(\gamma)},x^{(\delta)}} (x_p^{(\delta)} - \overline{x}^{(\delta)})} \tag{7.33}$$

on t_c s'obté de la t d'Student amb $n - r - 1$ graus de llibertat.

7.5.3 Intervals quan hi ha només una variable secundària

$r = 1$. Tenim una variable secundària x. La matriu $S_{x^{(\gamma)},x^{(\delta)}}$ es redueix a un element, $S_{x,x} = S_x^2$. trobem intervals sobre el punt x_p.

Intervals per a la regressió:

$$\widehat{y}_p \pm t_c \frac{S}{\sqrt{n}} \sqrt{1 + \frac{(x_p - \overline{x})^2}{S_x^2}} \tag{7.34}$$

on t_c s'obté de la t d'Student amb $n - 2$ graus de llibertat.

Intervals per a la predicció:

$$\widehat{y}_p \pm t_c S \sqrt{\frac{n+1}{n} + \frac{1}{n} \frac{(x_p - \overline{x})^2}{S_x^2}} \tag{7.35}$$

on t_c s'obté de la t d'Student amb $n - 2$ graus de llibertat.

Notem que l'amplada del intervals és mínima quan $x_p = \overline{x}$ i augmenta a mesura que ens allunyem de $\overline{x}$. És important tenir-ho en compte quan volem fer prediccions fora de l'interval de valors de x utilitzats en la mostra.

Per a n molt gran els intervals de regressió es redueixen a un punt mentre que els de predicció no desapareixen ja que Y és una variable aleatòria i té una dispersió pròpia.

Si dibuixem els límits de confiança (7.34) i (7.35) en funció de x_p s'obtenen les bandes de regressió i de predicció. Això dóna una visió gràfica d'aquests intervals.

Exemple 7.2 Continuem l'exemple anterior, afegint a la gràfica les bandes de regressió i de predicció. La comanda `predict()` troba els intervals de manera directa.

```
> new=data.frame(x)
> prec=predict(reg,new,interval="confidence")
> prep=predict(reg,new,interval="prediction")
> s<-sort(x,index.return=TRUE)
> xs<-s$x
> ix<-s$ix
> plot(x,y)
> abline(reg,col="red",lwd=2)
> abline(0,2,col="orange",lwd=2)
> lines(xs,prec[,2][ix],col="blue",lwd=2)
> lines(xs,prec[,3][ix],col="blue",lwd=2)
> lines(xs,prep[,2][ix],col="green",lwd=2)
> lines(xs,prep[,3][ix],col="green",lwd=2)
```

El codi anterior inclou la ordenació de les dades en x creixents a efectes de dibuixar la línia que uneix els punts.

Tal com s'observa, les bandes de predicció han de contenir la major part dels punts mentre que les de regressió han de contenir la recta de regressió exacta (la recta taronja en la gràfica).

♦

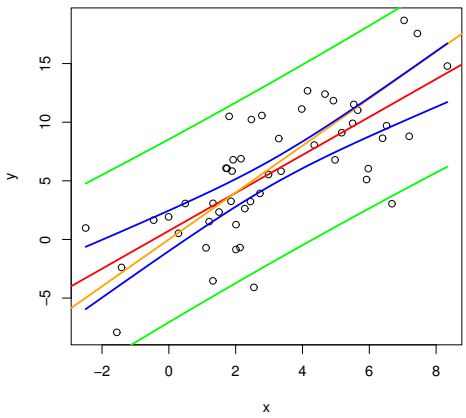

Figura 7.2: En blau les bandes de regressió. En verd les bandes de predicció.

Problemes

1. Una variable aleatòria contínua X té funció de densitat $f(x) = K(4-x)$ per $0 \leq x \leq 4$:
   (a) Calculeu el valor de la constant K. Trobeu la funció de distribució de X, $F(x)$.
   (b) Calculeu els moments de la variable X ($E[X^k]$).
   (c) Calculeu l'esperança, la variància, la desviació, i els coeficients d'asimetria i de curtosi de X.
   (d) Calculeu la mediana, els quartils Q_1 i Q_3 i l'interval interquartílic (IQR).
   (e) A partir d'un generador de nombres aleatòris uniformes en $[0, 1]$, com faríeu per generar valors de la variable X? Feu-ho amb el programa R: genereu 5000 valors uniformes i, a partir d'ells, 5000 valors de X. Dibuixeu els histogrames dels dos conjunts de nombres, comparant-los a les respectives densitats.

2. Resoleu el problema anterior per a una variable X amb funció de densitat $f(x) = \dfrac{K}{x^5}$ per $1 \leq x < \infty$.

3. Calculeu el IQR d'una variable $X \sim \mathrm{N}(\mu, \sigma)$. Calculeu la probabilitat que un valor de X sigui un valor atípic (*outlier*). Genereu mostres normals i dibuixeu el seu *boxplot*. Comproveu si el nombre observat de valors atípics està d'acord amb la probabilitat calculada.
   (Indicació: feu el càlcul per a una $\mathrm{N}(0, 1)$ i tingueu en compte que $\dfrac{X-\mu}{\sigma} \sim \mathrm{N}(0, 1)$.)

4. Una fàbrica produeix dos tipus de tub. El tipus A té diàmetre 5 mm. Es produeix amb toleràncies tals que aquest diàmetre es pot considerar una variable normal de mitjana 5 mm i desviació 0.3 mm. Pel tipus B tenim mitjana 6 mm i desviació 0.4 mm.
   (a) Si volem retenir només els tubs de tipus A amb diàmetre entre 4.4 mm i 5.6 mm, quin percentatge de la producció hem de retirar.
   (b) Tenim dos tubs, un de cada tipus, i hem de decidir quin és el de tipus A. El criteri lògic és decidir que és el de menor diàmetre. Quina és la probabilitat d'equivocar-nos?.
   (c) Prenem una mostra a l'atzar de 10 tubs de tipus B i calculem la seva mitjana aritmètica $\widehat{X}$. Què valen la mitjana i la desviació de $\widehat{X}$? Quina és la probabilitat que $\widehat{X}$ difereixi de la mitjana en menys que 0.15 mm? Compara-ho amb la probabilitat per una mostra d'un sol tub.

5. A un node d'una xarxa arribem paquets de dades de forma que el temps entre un paquet i el següent és una variable exponencial de valor mitjà 5 segons i aquests temps són independents.
   (a) Comprobeu que podem prendre els paràmetres α i β d'una variable de tipus gamma per tal d'obtenir una variable exponencial de paràmetre λ.
   (b) Sigui T el temps transcorregut fins que arriben tres nous paquets. Quin tipus de variable és? Calculeu les següents probabilitats (amb el programa R): $P(T > 25)$, $P(T < 10)$, $P(12 < T < 18)$.

(c) Quina esperança i desviació té T? Obteniu-ho a partir dels moments de la variable gamma i, alternativament, dels paràmeters de l'exponencial i la forma de T.

Genereu en R una mostra de 100 valors de T i compareu la seva mitjana i desviacions mostrals amb els paràmetres de T.

6. Considereu una variable X de Cauchy amb paràmetre α.

   (a) Trobeu la seva funció característica $C_X(\omega)$ (feu servir taules de transformades).

   (b) Què passa a l'intentar aplicat la propietat (4) de les funcions característiques per a calcular els moments de X?

   (c) Quina variable resulta al sumar vàries variables de Cauchy independents? Què implica això respecte al TLC?

7. Demostreu la propietat (3) de la variable Gamma calculant directament la funció característica de Y.

8. Repetiu l'exemple 3.11 amb $n = 1000$ i amb $n = 10000$ i compareu amb els resultats de l'exemple 1.5 i del problema 3.

9. A partir de l'interval de confiança per a proporcions:

   (a) Demostreu que $|\widehat{p} - p| < \dfrac{z_c}{\sqrt{4n}}$.

   (b) Amb la fita anterior podem determinar la mida n per a que l'error sigui menor que un valor ϵ donat. Feu una taula amb els valors de n per $c = 95\%$, $c = 99\%$ i $\epsilon = 10^{-1}$, $\epsilon = 10^{-2}$, $\epsilon = 10^{-3}$.

10. En l'estimació de proporcions hem utilitzat l'aproximació normal considerant mostres grans. Convé recordar que l'aproximació normal a la binomial requereix $np(1-p)$ gran. Si p és molt petit cal assegurar no només que n és gran ($n > 30$) sinó que np és gran.

    Suposem que en una població molt gran certa propietat A es troba amb $p = 10^{-4}$.

    (a) Considereu una mostra de mida $n = 10^4$ i sigui N_A el nombre d'elements que tenen la propietat en aquesta mostra. Llavors, $\widehat{p} = \dfrac{N_A}{n}$. Calculeu les probabilitats que $N_A = 0$, que $N_A = 1$ i que $N_A = 2$. Doneu els corresponents valors de $\widehat{p}$. La conclusió és que ja no val l'aproximació normal. Si p és molt petit i desconegut l'estimació de p requereix especial cura en l'elecció de n.

    (b) Preneu ara $n = 10^5$. Dibuixeu les probabilitats $P(N_A = k)$ per $0 \leq k \leq 20$. Noteu que comença a ser vàlida l'aproximació normal.

11. Apliqueu el mètode de màxima versemblança per a estimar els paràmetres μ i σ d'una variable normal a partir d'una mostra $X_1, X_2, \cdots, X_n$.

12. Apliqueu el mètode de màxima versemblança per a estimar el paràmetre α d'una variable de Poisson a partir d'una mostra $X_1, X_2, \cdots, X_n$.

13. Es mesura una resistència al laboratori obtenint, a partir de 10 mesures, una mitjana de 935 Ω i una desviació de 70 Ω.

    (a) Obteniu els límits de confiança del 95% i del 99%.

    (b) Repetiu l'apartat anterior pel cas que el nombre de mesures hagués estat 100.

    (c) Compareu els resultats anteriors amb els que s'obtindrien si la desviació fos exactament 70 Ω (estadístic normal). Seria el resultat vàlid en algun dels dos casos ($n = 10$ o $n = 100$)?

(d) Quantes mesures calen si volem que l'error sigui menor que 10 Ω (al 95% i al 99%)?

14. Una variable de Poisson X té paràmetre $\alpha = \lambda T$ (X compta el nombre d'esdeveniments durant el temps T).

    (a) Justifiqueu amb el TLC que si T és tal que el valor obtingut per a X és gran, l'interval de confiança per a α és $X \pm z_c\sqrt{X}$.

    (b) Un servidor rep 1428 connexions durant una hora. Trobeu l'interval de confiança del 95% per al nombre mitjà de connexions per hora.

15. En un test de cultura general sobre una mostra de 200 persones d'una gran ciutat s'obte una desviació mostral $\widehat{S_1}$ de 2.3 punts en les puntuacions obtingudes. Una mostra de la mateixa mida sobre la població d'un poble dóna una desviació mostral $\widehat{S_2}$ de 2.8.

    (a) Obteniu intervals del 95% de confiança per a les desviacions σ_1 i σ_2 de les respectives poblacions.

    (b) Feu un test per determinar si al poble la desviació és realment superior a la de la ciutat.

    (c) Són consistents les respostes dels dos apartats anteriors?

16. En 1000 tirades d'una moneda s'obtenen 453 cares. Estudieu mitjançant valors P la hipòtesi que la moneda sigui justa.

    (a) Fent servir la distribució binomial (resultat exacte).

    (b) Fent servir la distribució normal (mostra gran).

    (c) Fent servir la distribució normal amb la correcció del mig punt.

17. En una planta de fabricació de condensadors s'ha establert que la desviació típica en la capacitat dels condensadors produïts és del 3% d'aquesta capacitat. Amb una mostra de 10 condensadors s'obté una desviació mostral $\widehat{S}$ igual al 4.5% de la capacitat. Apliqueu un test a la hipòtesi H_0 = "La producció segueix sent correcta" obtenint un valor P i dient quines són les conclusions al 5% de significació i a l'1% de significació.

18. En una escola, el grup de matí, de 60 alumnes, té nota mitjana 6.3 amb desviació 1.5 mentre que el grup de tarda, de 40 alumnes, té nota mitjana 5.4 amb desviació 1.7. Estudieu si la diferència és significativa, fent servir el valor P.

19. Demostreu que pel cas $r = 2$ el test χ^2 coincideix amb el test Z.

20. L'equació (6.24) ens dóna la recta de regressió de y sobre x. Escriviu també la recta de regressió de x sobre y. Determineu el punt on es talles les dues rectes. Calculeu l'angle que formen les dues rectes. Què val aquest angle en els casos extrems $r = 0$ i $r = \pm 1$?

21. Verifiqueu en R els resultats de l'exemple 6.1. Feu la gràfica amb les dades i la recta de regressió.

22. L'arxiu `elusage.txt` conté dades sobre el consum d'energia i la temperatura en cada més, per diferents anys. Les variables són `Month`, `Year`, `KWH` and `Temp`. Carregueu les dades en R (`elusage<-read.table("elusage.txt")`, `attach(elusage)`).

    (a) Mireu el gràfic de dispersió per `Month` i `KWH`. Volem utilitzar `Month` com a predictor per `KWH`. Quin és el grau adequat d'un ajust polinòmic? Calculeu-lo, així com el corresponent valor de R^2. Dibuixeu les dades junt amb l'ajust polinòmic. Mireu la taula ANOVA d'aquest ajust i expliqueu com el valor de R^2 s'obté d'ella.

(b) Per què no hagués estat lògic ajustar amb una línia recta en l'apartat anterior? mireu ara de predir `Temp` a partir de `Month`. Compareu els valors de R^2 d'un ajust parabòlic i d'un estimador $a\cos(\omega M) + b\sin(\omega M) + c$, on M és el mes, ω és una constant fixada i a, b, c els paràmetres de l'ajust. Dibuixeu les dades junt amb les dues regressions.

(c) Calculeu la recta de regressió per la resposta `KWH` i predictor `Temp`. Dibuixeu les dades junt amb la recta de regressió, les bandes de confiança i les bandes de predicció.

```
Month Year KWH Temp
8 1989 24.828 73
9 1989 24.688 67
10 1989 19.310 57
11 1989 59.706 43
12 1989 99.667 26
1 1990 49.333 41
2 1990 59.375 38
3 1990 55.172 46
4 1990 55.517 54
5 1990 25.938 60
6 1990 20.690 71
7 1990 24.333 75
8 1990 22.759 74
9 1990 24.688 66
10 1990 22.759 61
11 1990 50.588 49
12 1990 79.000 41
1 1991 87.188 35
2 1991 47.333 41
3 1991 38.621 42
4 1991 27.931 56
5 1991 25.000 69
6 1991 17.241 73
7 1991 17.000 77
8 1991 22.188 74
9 1991 18.276 66
10 1991 27.241 56
11 1991 39.706 47
12 1991 81.667 38
1 1992 79.688 34
2 1992 67.667 37
3 1992 65.862 39
4 1992 54.839 51
5 1992 34.667 60
6 1992 23.448 70
7 1992 29.586 73
8 1992 32.448 71
9 1992 30.033 66
10 1992 43.586 54
11 1992 48.029 45
12 1992 52.424 36
1 1993 62.333 34
2 1993 71.156 32
3 1993 44.621 39
```

```
4 1993 19.357 55
5 1993 23.733 64
6 1993 10.414 72
7 1993 17.875 79
8 1993 24.207 75
9 1993 24.333 65
10 1993 40.935 54
11 1993 42.938 48
12 1993 65.250 35
1 1994 101.300 24
2 1994 101.660 32
```

SOLUCIONS DELS PROBLEMES

1. (a) $K = \dfrac{1}{8}$. $F(x) = \dfrac{x}{2} - \dfrac{x^2}{16}$, $0 < x < 4$.

   (b) $E[X^k] = \dfrac{2 \cdot 4^k}{(k+1)(k+2)}$.

   (c) $\mu_X = \dfrac{4}{3} = 1.33$, $\sigma_X^2 = \dfrac{8}{9} = 0.88$, $\sigma_X = \dfrac{2\sqrt{2}}{3} = 0.94$, asimetria $= \dfrac{2\sqrt{2}}{5} = 0.56$, curtosi $= \dfrac{12}{5} = 2.4$.

   (d) Amb la funció de quantils $F^{-1}(p) = 4(1 - \sqrt{1-p})$: $Q_1 = 4 - 2\sqrt{3} = 0.54$, $m = Q_2 = 4 - 2\sqrt{2} = 1.17$. $Q_3 = 2$. IQR $= 2\sqrt{3} - 2 = 1.46$.

   (e) $X = 4(1 - \sqrt{U})$.

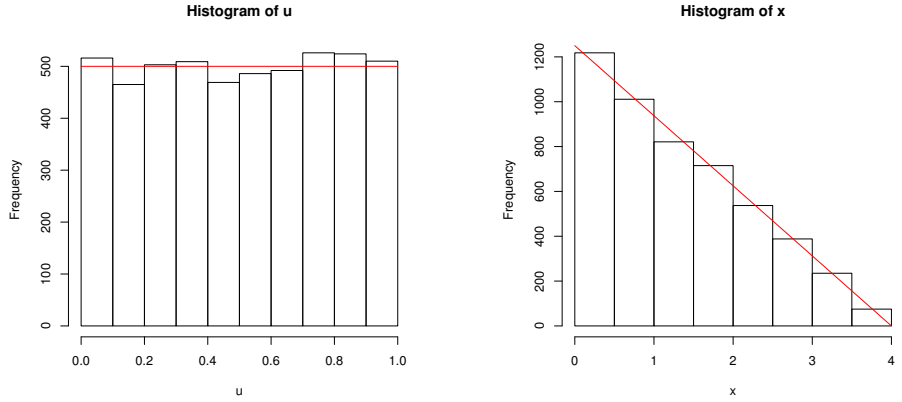

Figura 7.3: Problema 1(e).

2. (a) $K = 4$. $F(x) = 1 - \dfrac{1}{x^4}$, $x \geq 1$.

   (b) $E[X^k] = \dfrac{4}{4-k}$, $0 \leq k \leq 3$. Els moments d'ordre igual o superior a 4 són infinits.

   (c) $\mu_X = \dfrac{4}{3} = 1.33$, $\sigma_X^2 = \dfrac{2}{9} = 0.22$, $\sigma_X = \dfrac{\sqrt{2}}{3} = 0.47$, asimetria $= 5\sqrt{2} = 7.07$, la curtosi no està definida.

   (d) Amb la funció de quantils $F^{-1}(p) = \dfrac{1}{\sqrt[4]{1-p}}$: $Q_1 = \sqrt[4]{\frac{4}{3}} = 1.07$, $m = Q_2 = \sqrt[4]{2} = 1.19$. $Q_3 = \sqrt{2} = 1.41$. IQR $= 0.34$.

   (e) $X = \dfrac{1}{\sqrt[4]{U}}$.

3. Per a una variable $\mathrm{N}(0,1)$, $Q_1 = -0.6744898$ (`>qnorm(0.25)`), $m = Q_2 = 0$, $Q_3 = 0.6744898$ (`>qnorm(0.75)`). Segons l'apartat (a), per a una variable $\mathrm{N}(\mu, \sigma)$, $Q_1 = \mu - 0.6744898\sigma$, $m = Q_2 = \mu$, $Q_3 = \mu + 0.6744898\sigma$. IQR $= 1.34898\sigma$.

   $P(\text{outlier}) = 2P(X > Q_3 + 1.5 \cdot \text{IQR}) = 2P\left(\dfrac{X-\mu}{\sigma} > 2.697959\right) = 0.0069766$. Esperem un 0.7% d'outliers. En la figura es mostra un boxplot d'una mostra de mida 500 amb 5 outliers (en mitjana se n'observen 3.5).

4. (a) 4.5%. (b) 0.023. (c) $\mu_{\widehat{X}} = 6$, $\sigma_{\widehat{X}} = 0.126$. $P = 0.764$. Amb un sol tub, $P = 0.292$.

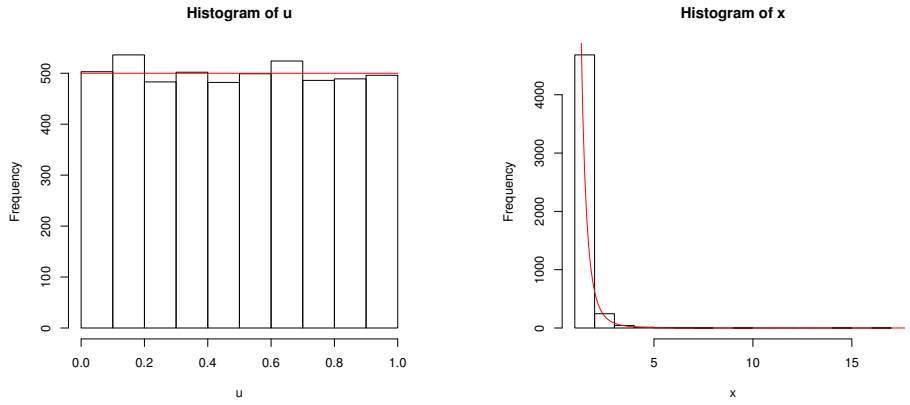

Figura 7.4: Problema 2(e).

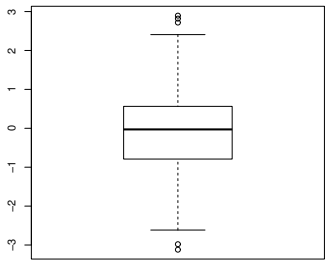

Figura 7.5: Problema 3(b).

5. (a) $\alpha = 1$ i $\beta = \dfrac{1}{\lambda}$.

   (b) $T = X_1 + X_2 + X_3$ on les X_i són Gamma$(1, 5)$ independents, d'on $T \sim$ Gamma$(3, 5)$. $P(T > 25) = 0.1246$, $P(T < 10) = 0.3233$, $P(12 < T < 18) = 0.2669$.

   (c) $\mu_T = 15$, $\sigma_T = 5\sqrt{3} = 8.6602$. Amb una mostra de 100 valors s'obté, per exemple, $\widehat{X} = 15.75$ i $\widehat{S} = 8.49$.

6. (a) $C_X(\omega) = e^{-\alpha|\omega|}$.

   (b) Degut al valor absolut, $C_X(\omega)$ no admet derivades en 0, d'acord amb la no existència dels moments de X.

   (c) Resulta una variable de Cauchy, de manera que no es verifica el TLC (el TLC requereix l'existència dels dos primers moments).

7.

8.

9.

	$\epsilon = 10^{-1}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-3}$
$c = 95\%$	n=96	n=9600	n=960000
$c = 99\%$	n=166	n=16600	n=1660000

10. (a) $P(N_A = 0) = 0{,}368$ amb $\widehat{p} = 0$, $P(N_A = 1) = 0{,}368$ amb $\widehat{p} = 10^{-4}$. $P(N_A = 2) = 0{,}184$ amb $\widehat{p} = 2 \cdot 10^{-4}$.

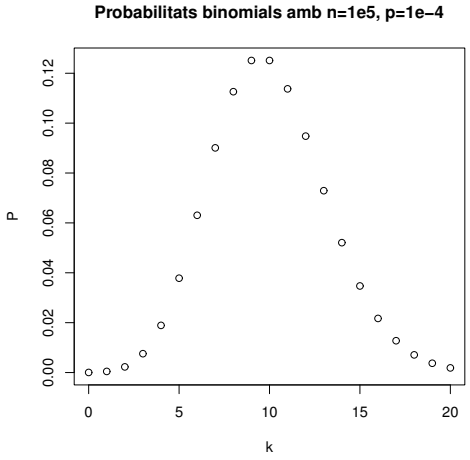

Figura 7.6: problema 6(b).

11. L'estimador de μ és $\widehat{X}$ i l'estimador de σ és S.

12. L'estimador de α és $\widehat{X}$.

13. (a) Al 95%: $(885, 985)$. Al 99%: $(863, 1007)$. (b) Al 95%: $(921, 949)$. Al 99%: $(917, 953)$. (c) $n = 10$, al 95%: $(892, 978)$, al 99%: $(878, 992)$. $n = 100$, al 95%: $(921, 949)$, al 99%: $(917, 953)$. Per $n = 100$ l'aproximació és bona. (d) Al 95%: $n \geq 189$, al 99%: $n \geq 326$ (191 i 329 sense l'aproximació normal).

14. (a) Si Y_i, $i = 1, \ldots, n$, són Poisson de paràmetre α_1 llavors, per a n gran, $X = \sum_i Y_i$ s'aproxima a una normal amb $\mu = n\alpha_1$ i $\sigma = \sqrt{n\alpha_1}$. Tenim que X és Poisson de paràmetre $\alpha = n\alpha_1$ i s'aproxima a una $\mathrm{N}(\alpha, \sqrt{\alpha})$ on α és gran. De $\frac{X-\alpha}{\sqrt{\alpha}} \sim N(0,1)$ treiem un interval que, de manera similar al que passa amb les proporcions, s'aproxima com $\alpha \in (X - z_c\sqrt{X}, X + z_c\sqrt{X})$. (b) $(1354, 1502)$.

15. (a) $\sigma_1 \in (2.09, 2.55), \sigma_2 \in (2.55, 3.10)$. (b) $P = 0.0029$ d'on es rebutja que $\sigma_1 = \sigma_2$ en favor de $\sigma_2 > \sigma_1$. (c) Els intervals de confiança no es solapen, fet consistent amb que les desviacions siguin diferents.

16. (a) 0.0032526. (b) 0.0029534. (c) 0.0032724.

17. $P = 0.0164$. Rebutgem H_0 al 5% però l'acceptem a l'1%.

18. Fent servir pooling per a la variància i amb un test T: $P = 0.0032$. La diferència és significativa i rebutgem la hipòtesi que els grups siguin iguals.

19. En aquest cas $X_1 = X$, $X_2 = n - X$, $p_1 = p$, $p_2 = q$.
$$\chi^2 = \frac{(X-np)^2}{np} + \frac{(n-X-nq)^2}{nq} = \frac{(X-np)^2}{npq} = Z^2.$$

20. L'altra recta és $\dfrac{y-\overline{y}}{S_y} = r^{-1}\dfrac{x-\overline{x}}{S_x}$. Es tallen en el punt $(\overline{x}, \overline{y})$. L'angle que formen és $\operatorname{arctg}\dfrac{1-r^2}{2r}$. Per $r = 0$ les rectes són ortogonals (angle $\pi/2$). Per $r = \pm 1$ les rectes coincideixen (angle 0).

21.

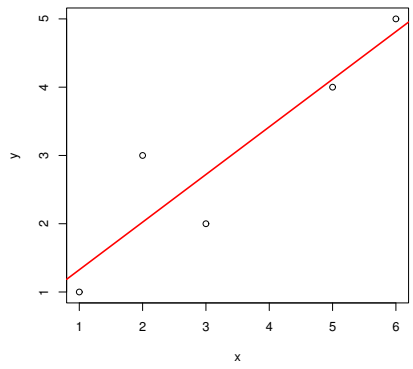

Figura 7.7: Problema 21.

22. (a) El grau adequat és 2. L'estimador és $aM^2 + bM + c$ amb $a = 1.8359$, $b = -25.2869$, $c = 107.1111$. $R^2 = 0.727$. La taula és:

```
Response: KWH
            Df  Sum Sq Mean Sq  F value    Pr(>F)
Month        1  1573.8  1573.8   9.6303  0.003093 **
I(Month^2)   1 21055.7 21055.7 128.8402 1.093e-15 ***
Residuals   52  8498.1   163.4
```

R^2 s'obté dividint la variació explicada ($1573.8 + 21055.7$) per la variació total ($1573.8 + 21055.7 + 8498.1$).

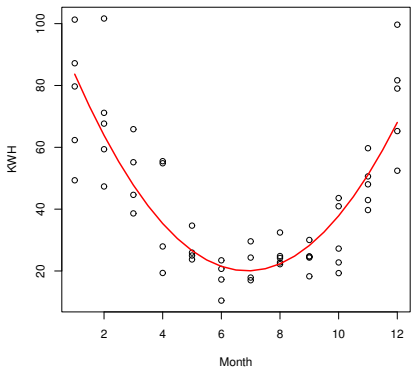

Figura 7.8: Problema 22 (a).

(b) L'ajust amb una recta (de fet, amb un polinomi) no pot ser òptim ja que les dades segueixen la variació anual que és de tipus periòdic. És $\omega = \frac{2\pi}{12} = \frac{\pi}{6}$. Resulta $a = -17.5139$, $b = -11.9434$, $c = 54.4422$. $R^2 = 0.9564$.

(c) S'obté pendent $a = -1.365$ i intercept $b = 116.716$.

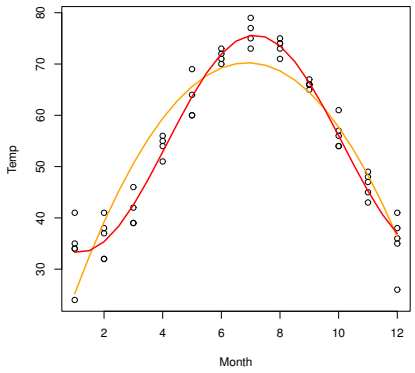

Figura 7.9: Problema 22 (b).

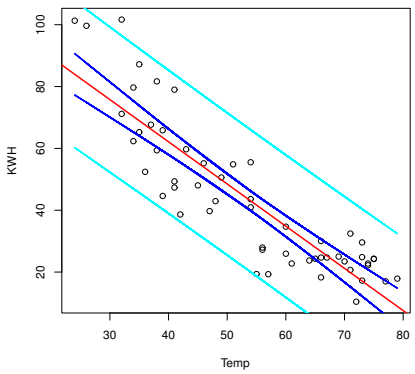

Figura 7.10: Problema 22 (c).