13.Proves d'Hipòtesis clàssiques

Estadística
Grau en Matemàtiques

Josep A. Sanchez
Dept. Estadística i I.O.(UPC)

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Distribució assimptòtica de l'estadístic raó de versemblança

Teorema: Sigui $X_1, \ldots, X_n$ una m.a.s. amb $X_i \sim f(x;\theta)$ per algun $\theta \in \Theta$. Volem testar:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$

on $\Theta = \Theta_0 \cup \Theta_1$ amb $\Theta_0 \cap \Theta_1 = \emptyset$

Asumim que:

1. Es donen les condicions del teorema de Cràmer-Rao
2. El paràmetre θ és identificable (diferents valors de θ donen diferents distribucions de probabilitat per X)
3. El suport de la variable $\{x : f(x|\theta) > 0\}$ és el mateix $\forall \theta \in \Theta$
4. $e(\theta_0|\theta) = E\left[\log\left(\frac{f(\underset{\sim}{X};\theta)}{f(\underset{\sim}{X};\theta_0)}\right)\right]$ existeix per cada parell $\theta, \theta_0 \in \Theta$

Aquestes condicions impliquen que les derivades de la funció de versemblança existeixen i són continues i que el suport de la distribució no depèn del paràmetre. Donades aquestes condicions, l'estadístic

$$Q_n = -2\log(\lambda(X_n)) \to \chi^2_d \qquad \text{under } H_0 : \theta = \theta_0$$

amb $d = dim(\Theta) - \dim(\Theta_0)$

Exemple: bondat d'ajust de la distribució multinomial

Sigui $X \sim \mathfrak{F}(x;\theta)$. Volem fer un test de **bondat d'ajust**:

$$\begin{cases} H_0 : \mathfrak{F}(x;\theta) = \mathfrak{F}_0(x;\theta) \\ H_1 : \mathfrak{F}(x;\theta) \neq \mathfrak{F}_0(x;\theta) \end{cases}$$

$\mathfrak{F}_0(x;\theta)$ pot ser qualsevol distribució (Normal, exponencial,...). Tenim valors $X_1, \ldots, X_n$ i els agrupem en classes disjuntes $C_1, \ldots, Cm$, on $C_1 = (-\infty, c_1], C_2 = (c_1, c_2], \ldots, C_m = (c_m - 1, c_m], C_{m+1} = (c_m, +\infty)$

Definim

$$p_j = P(X \in C_j) = p_j(\theta) \qquad j = 1, \ldots, m+1$$
$$p_j^0 = P(X \in C_j | X \sim \mathfrak{F}_0(x;\theta)) = p_j^0(\theta)$$

Podem rescriure el test com

$$\begin{cases} H_0 : p_1 = p_1^0(\theta), \ldots, p_{m+1} = p_{m+1}^0(\theta) \\ H_1 : \exists j : p_j \neq p_j^0(\theta) \end{cases}$$

Sigui $Y_j = \sum_{i=1}^n \mathbb{I}_{\{X_i \in C_j\}}$.

Llavors $(Y_1, \ldots, Y_{m+1}) \sim MN(n, p_1, \ldots, p_{m+1})$

Exemple: bondat d'ajust de la distribució multinomial

La funció de versemblança es:

$$L(\theta|\underset{\sim}{X}) = \frac{n}{y_1! \dots y_{m+1}!} p_1(\theta)^{y_1} \dots p_{m+1}(\theta)^{y_{m+1}}$$

té un màxim en $\hat{p}_j = \frac{y_j}{n}$. Sigui $\hat{(\theta)}$ l'estimador màxim versemblant de θ fent servir la mostra $\underset{\sim}{X}$. Llavors, l'estimador màxim versemblant de p_j sota H_0 és $p_j^0(\hat{\theta})$.

L'estadístic de raó de versemblances és:

$$\Lambda(y_1, \dots, y_{m+1}) = \frac{p_1^0(\hat{\theta})^{y_1} \dots p_{m+1}^0(\hat{\theta})^{y_{m+1}}}{\hat{p}_1^{y_1} \dots \hat{p}_{m+1}^{y_{m+1}}} = \prod_{j=1}^{m+1} \left(\frac{p_j^0(\hat{\theta})}{\hat{p}_j} \right)^{y_j}$$

Exemple: bondat d'ajust de la distribució multinomial

Fent servir la distribució assimptòtica de l'estadístic raó de versemblança:

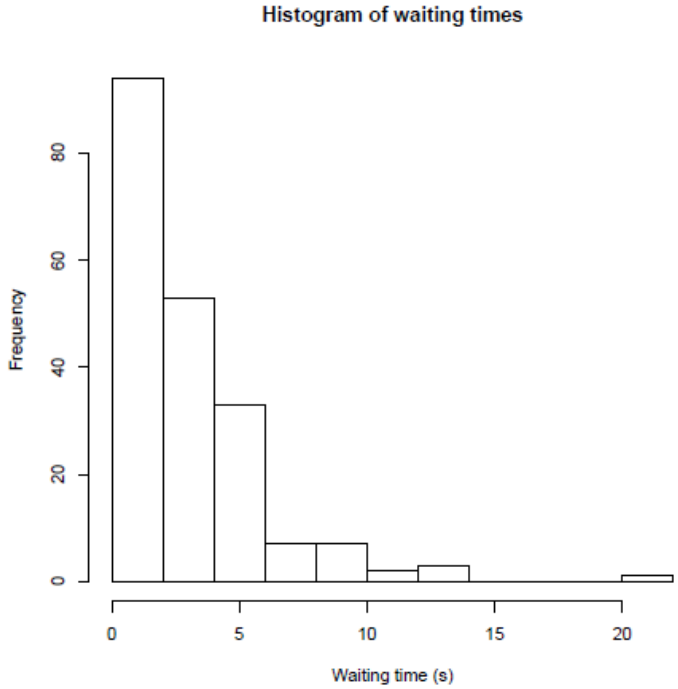
$$-2\log\Lambda = -2\sum_{j=1}^{m+1} y_j \log\left(\frac{p_j^0(\hat{\theta})}{\hat{p}_j}\right) = -2n\sum_{j=1}^{m+1} \hat{p}_j \log\left(\frac{p_j^0(\hat{\theta})}{\hat{p}_j}\right)$$

Per tant, $-2\log\Lambda \sim \chi^2_{m+1-k-1} = \chi^2_{m-k}$ sota H_0 on k és el nombre de paràmetres estimats a partir de les dades. Fent servir l'expansió de Taylor es pt comprovar que assimptòticament,

$$-2\log\Lambda \simeq \sum_{j=1}^{m+1} \frac{(y_j - np_j^0(\hat{\theta}))^2}{np_j^0(\hat{\theta})}$$

Aquest és l'habitual estadístic Chi-quadrat per a la bondat d'ajust.

La distribució dels temps d'espera



	obs	pr	exp	chi2
(0,2]	94.00	0.48	96.80	0.10
(2,4]	53.00	0.25	50.00	0.20
(4,6]	33.00	0.13	25.80	2.00
(6,8]	7.00	0.07	13.20	2.90
(8,10]	7.00	0.03	6.80	0.00
(10,50]	6.00	0.04	7.40	0.30

$$X^2 = 0.1 + 0.2 + \cdots + 0.3 = 5.5$$

$$\chi^2_{4,0.95} = 9.49$$

$$P\left(\chi^2_4 \geq 5.5\right) = 0.24$$

```
   N N*  Mean Stdev   Med    Q1    Q3   Min    Max
X 200  0 3.022   2.9 2.316 0.954 4.274 0.013 21.633
```

Test de la t-Student

Sigui $X \sim N(\mu, \sigma^2)$ i $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < +\infty, 0 < \sigma^2 < +\infty\}$ Volem testar:

$$\begin{cases} H_0 : \mu = 0, \sigma^2 > 0 \\ H_1 : \mu \neq 0, \sigma^2 > 0 \end{cases}$$

$$\Theta_0 = \{(\mu, \sigma^2) : \mu = 0, 0 < \sigma^2 < +\infty\}$$

Tenim una m.a.s. de mida n>1 $X_1, \ldots, X_n$ amb $X \sim N(\mu, \sigma^2)$

Funció de versemblança:

$$L(\mu, \sigma^2 | \underset{\sim}{X}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right]$$

Sota H_0:

$$L(0, \sigma^2 | \underset{\sim}{X}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right]$$

Test de la t-Student

Calculem el màxim de versemblança en Θ_0 i en Θ i construim el tes de raó de versemblança:

- En Θ_0

$$\frac{\partial}{\partial \sigma^2} \log L(0, \sigma^2 | \underset{\sim}{X}) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n x_i^2}{2\sigma^4} = 0 \Leftrightarrow \tilde{\sigma}^2 = \frac{\sum_{i=1}^n x_i^2}{n}$$

$$L(0, \tilde{\sigma}^2 | \underset{\sim}{X}) = \left(\frac{1}{2\pi \frac{\sum_{i=1}^n x_i^2}{n}} \right)^{n/2} \exp \left[-\frac{\sum_{i=1}^n x_i^2}{2\frac{\sum_{i=1}^n x_i^2}{n}} \right] = \left(\frac{ne^{-1}}{2\pi \sum_{i=1}^n x_i^2} \right)^{n/2}$$

- En Θ

$$\hat{\mu}_{ML} = \bar{X} \qquad \hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

$$L(\hat{\mu}, \hat{\sigma}^2 | \underset{\sim}{X}) = \left(\frac{1}{2\pi \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \right)^{n/2} \exp \left[-\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \right] = \left(\frac{ne^{-1}}{2\pi \sum_{i=1}^n (x_i - \bar{x})^2} \right)^{n/2}$$

Test de la t-Student

Raó de versemblances:

$$\Lambda(x_1,\ldots,x_n)=\frac{L(0,\tilde{\sigma}^2|\underset{\sim}{X})}{L(\hat{\mu},\hat{\sigma}^2|\underset{\sim}{X})}=\frac{\left(\frac{ne^{-1}}{2\pi\sum_{i=1}^n x_i^2}\right)^{n/2}}{\left(\frac{ne^{-1}}{2\pi\sum_{i=1}^n (x_i-\bar{x})^2}\right)^{n/2}}=$$

$$\left(\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n x_i^2}\right)^{n/2}=\left(\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2+n\bar{x}^2}\right)^{n/2}=\left(\frac{1}{1+\frac{n\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right)^{n/2}$$

Com sempre, $0\leq\Lambda\leq 1$

Sota $H_0:\mu=0,\sigma^2>0$

- Si $\bar{x}=0$ i $\sum_{i=1}^n x_i^2>0$, les dades confirmen H_0 i $\Lambda=1$
- Si $\bar{x}$ i $\frac{n\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}$ són valors molt més grans que zero, les dades fan rebutjar H_0 i Λ serà petit.

Procediment: Rebutjar H_0 si $0\leq\Lambda\leq\lambda_0$

Test de la t-Student

$$\Lambda \leq \lambda_0 \Leftrightarrow \left(1+\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n\bar{x}^2}\right)^{-n/2} \leq \lambda_0$$

$$\Leftrightarrow \left(1+\frac{n\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)^{n/2} \geq \frac{1}{\lambda_0}$$

$$\Leftrightarrow \frac{n\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2} \geq \frac{1}{\lambda_0^{2/n}} - 1$$

$$\Leftrightarrow \frac{n\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2/(n-1)} \geq (n-1)(\lambda_0^{-2/n}-1)$$

$$\Leftrightarrow \frac{\sqrt{n}|\bar{x}|}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2/(n-1)}} \geq \sqrt{(n-1)(\lambda_0^{-2/n}-1)} = c$$

Sota H_0 tenim:

$$t(\underset{\sim}{X}) = \frac{\sqrt{n}(\bar{x}-0)}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2/(n-1)}} \sim t_{n-1}$$

Basem el test de raó de versemblança de H_0 versus H_1 en aquest estadístic: Rebutgem H_0 si $|t(\underset{\sim}{X})| \geq c$, on c es calcula imposant

$$\alpha = P(|t(\underset{\sim}{X})| \geq c | H_0)$$

Exemple

Un pes estàndard es mesura repetidament en les mateixes balances. Les desviacions del pes estàndard són (en grams):

$$-0.5605, -0.2302, 1.5587, 0.0705, 0.1293, 1.7151, 0.4609, -1.2651, -0.6869, -0.4457$$

Testem la hipòtesi:

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

$$\bar{X} = 0.0746 \qquad S = 0.9538 \qquad n = 10$$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{0.0746 - 0}{0.9538/\sqrt{10}} = 0.2474$$

1. Regió Crítica: $C = \{\underset{\sim}{X} : |T| \geq t_{n-1,1.\alpha/2}\}$ amb $t_{9,0.975} = 2.2622$
2. P-valor$=2P(t_9 \geq 0.2474) = 0.8101$

Interval de Confiança: $IC_{95\%}(\mu) = \bar{X} \pm t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}} = [-0.6077, 0.7569]$

Test de la t-Student per qualsevol μ_0

Si volem testar

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

$$\Theta_0 = \{(\mu, \sigma^2) : \mu \leq \mu_0, 0 < \sigma^2\}$$

Els estimadors màxim-versemblants en Θ_0 són:

$$\hat{\mu}_{ML} = \mu_0 \qquad \hat{\sigma}^2_{ML} = \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{n}$$

i per tant,

$$\Lambda^{-1} = \left(1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^{n/2}$$

i l'estadístic T corresponent és:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Observacions aparellades

Hi ha experiments on les observacions de dues mostres estan aparellades: per a cada observació d'una mostra hi ha una corresponent en l'altra mostra.
Exemples:

- Es mesura la presió arterial de un grup de persones abans i després de fer exercici
- Es mesura el desgast de les soles de les dues sabates que porten els individus de la mostra

Sigui (X_i, Y_i) les parelles amb $i = 1, \ldots, n$ i X_i i Y_i tenen esperances μ_X i μ_Y respectivament i variàncies σ_X^2 i σ_Y^2. Les diferents parelles són independents però les dues components de la parella tenen covariància $Cov(X_i, Y_i) = \sigma_{XY}$ En aquest disseny, es treballa amb les diferències:

$$D_i = X_i - Y_i$$

$$E(D_i) = \mu_X - \mu_Y \qquad V(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}$$

Estimem $\mu_X - \mu_Y$ amb $\bar{D}$

$$E(\bar{D}) = \mu_X - \mu_Y \qquad V(\bar{D}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})$$

Observacions aparellades

Suposem que un experiment s'ha dut a terme amb un disseny NO aparellat i tenim dues mostres de mida n, $\underset{\sim}{X}$ i $\underset{\sim}{Y}$ que són independents entre elles

Estimem $\mu_X - \mu_Y$ mab $\bar{X} - \bar{Y}$:

$$E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y \qquad V((\bar{X} - \bar{Y}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2)$$

En el cas aparellat, la variància de $\bar{D}$ serà més petita si hi ha una covariància positiva. Suposem que les variàncies coincideixen, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, llavors

Cas aparellat: $V(\bar{D}) = \frac{2\sigma^2(1-\rho)}{n}$

Cas independent: $V(\bar{X} - \bar{Y}) = \frac{2\sigma^2}{n}$

El rati serà: $\frac{V(\bar{D})}{V(\bar{X}-\bar{Y})} = 1 - \rho$

Si $\rho = 0.5$ la precisió d'una mostra de n pairs és la mateixa que la d'un disseny independent de mida $2n$ per a cada mostra.

Test de la t-Student per a dues mostres aparellades

Suposem que les diferències $D_i = X_i - Y_i$ són Normals amb

$$E(D_i) = \mu_X - \mu_Y = \mu_D \qquad V(D_i) = \sigma_D^2$$

En general σ_D^2 és desconeguda,

$$T = \frac{\bar{D} - \mu_D}{S_{\bar{D}}} = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}$$

i l'interval de confiança serà:

$$IC_{1-\alpha}(\mu) = \bar{D} \pm t_{n-1,1-\alpha/2} \frac{S_D}{\sqrt{n}}$$

Si $H_0 : \mu_D = 0$, amb nivell de significació α rebutgem H_0 si:

$$|T| \geq t_{n-1,1-\alpha/2} \Leftrightarrow |\bar{D}| \geq t_{n-1,1-\alpha/2} \frac{S_D}{\sqrt{n}}$$

Test de la t-Student per a dues mostres aparellades

Quan cada unitat experimental/subjecte es mesura dues vegades (p.exemple, abans i després d'un tractament) les observacions estan aparellades-

Aquestes dues mesures no són independents, sino correlacionades

El test es planteja:

$$\begin{cases} H_0 : \delta = \mu_1 - \mu_2 = 0 \\ H_1 : \delta \neq 0 \end{cases}$$

Estadístic de test:

$$T = \frac{\bar{D} - \delta}{S_D/\sqrt{n}}$$

Distribució de referència sota H_0:

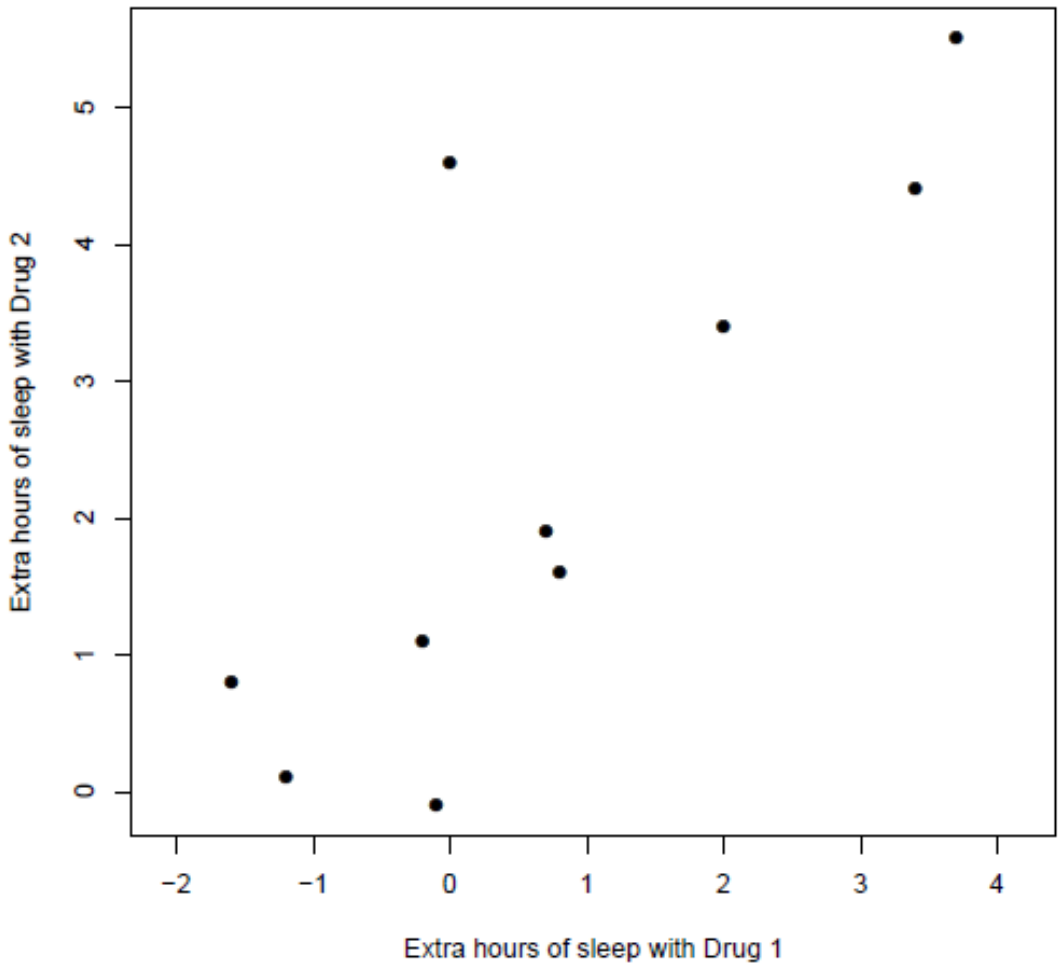
$$T|H_0 \sim t_{n-1}$$

Per tant, un disseny de dues mostres aparellades es resol com un test d'una mostra sobre les diferències.

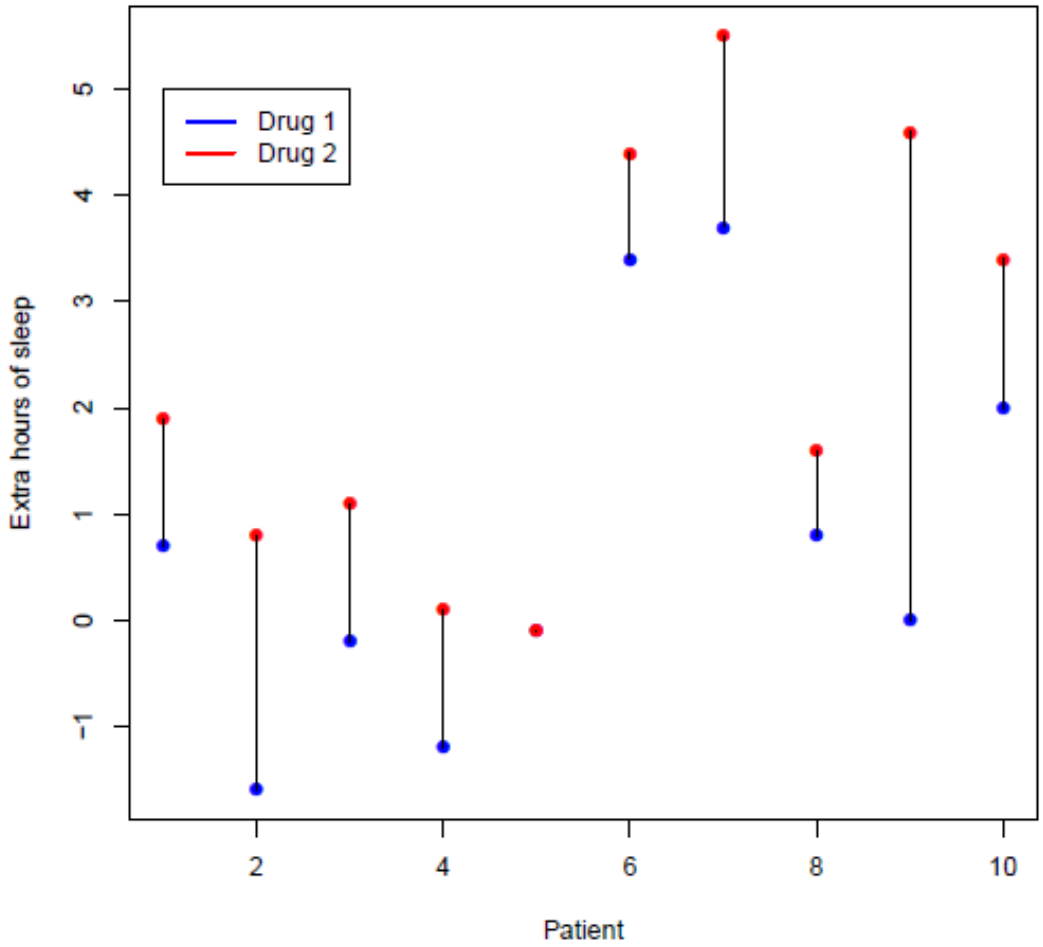
Test de la t-Student per a dues mostres aparellades

Patient	Drug 1	Drug 2	D
1	0.70	1.90	1.20
2	-1.60	0.80	2.40
3	-0.20	1.10	1.30
4	-1.20	0.10	1.30
5	-0.10	-0.10	0.00
6	3.40	4.40	1.00
7	3.70	5.50	1.80
8	0.80	1.60	0.80
9	0.00	4.60	4.60
10	2.00	3.40	1.40
mean	0.75	2.33	1.58
s.d.	1.79	2.00	1.23

Test de la t-Student per a dues mostres aparellades



Test de la t-Student per a dues mostres aparellades



Test d'hipòtesi per dades aparellades

$$\begin{cases} H_0 : \delta = \mu_1 - \mu_2 = 0 \\ H_1 : \delta \neq 0 \end{cases}$$

$$T = \frac{\bar{D} - \delta}{S_D/\sqrt{n}} = \frac{1.58}{1.23/\sqrt{10}} = 4.0621$$

p-valor$= 2P(t_9 \geq 4.0621) = 2 * 0.0014 = 0.0028$

$$IC_{95\%}(\delta) = [0.70, 2.46]$$

Test d'hipòtesi per dades aparellades en R

```
> t.test(drug1,drug2,paired=TRUE)

Paired t-test

data:  drug1 and drug2

t = -4.0621, df = 9, p-value = 0.002833

alternative hypothesis:  true difference in means is not equal to 0

95 percent confidence interval:

-2.4598858 -0.7001142

sample estimates:

mean of the differences

-1.58
```

Test de la t-Student per a dues mostres independents

- Les observacions no estan aparellades
- Comparem dos grups (p.exemple Tractament versus Controls)
- Hi ha dues mostres independents (no necessàriament de la mateixa mida)

Grup tractat: $X_1, \ldots, X_m \quad X_i \sim N(\mu_1, \sigma^2)$

Grup Control: $Y_1, \ldots, Y_n \quad Y_i \sim N(\mu_2, \sigma^2)$

Efecte del tractament: $\mu_1 - \mu_2$

Test d'hipòtesis:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu \Leftrightarrow \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0 \end{cases}$$

- Si σ^2 és comuna i coneguda:

$$\lambda(\underset{\sim}{X}, \underset{\sim}{Y}) = \frac{\sup_{(\mu,\mu,\sigma^2)} L(\mu, \mu, \sigma^2 | \underset{\sim}{X}, \underset{\sim}{Y})}{\sup_{(\mu_1,\mu_2,\sigma^2)} L(\mu_1, \mu_2, \sigma^2 | \underset{\sim}{X}, \underset{\sim}{Y})}$$

  Rebutgem H_0 si $\lambda(\underset{\sim}{X}, \underset{\sim}{Y}) \leq c$. Estimem $\mu_1 - \mu_2$ amb $\bar{X} - \bar{Y}$ i rebutgem si $\bar{X} - \bar{Y}$ és prou "gran".

Test de la t-Student per dues mostres independents (σ^2 comuna)

Tenim,

$$\bar{X} - \bar{Y} \sim N\left[\mu_1 - \mu_2, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right]$$

Si σ^2 és coneguda, l'estadístic del test es basa en la quantitat pivotal:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} | H_0 \sim N(0,1)$$

$$IC_{1-\alpha}(\mu_1 - \mu_2) = \bar{X} - \bar{Y} \pm Z_{1-\alpha/2}\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}$$

En la pràctica, σ^2 no és coneguda i s'estima a partir de les dues mostres:

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2} \qquad T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} | H_0 \sim t_{m+n-2}$$

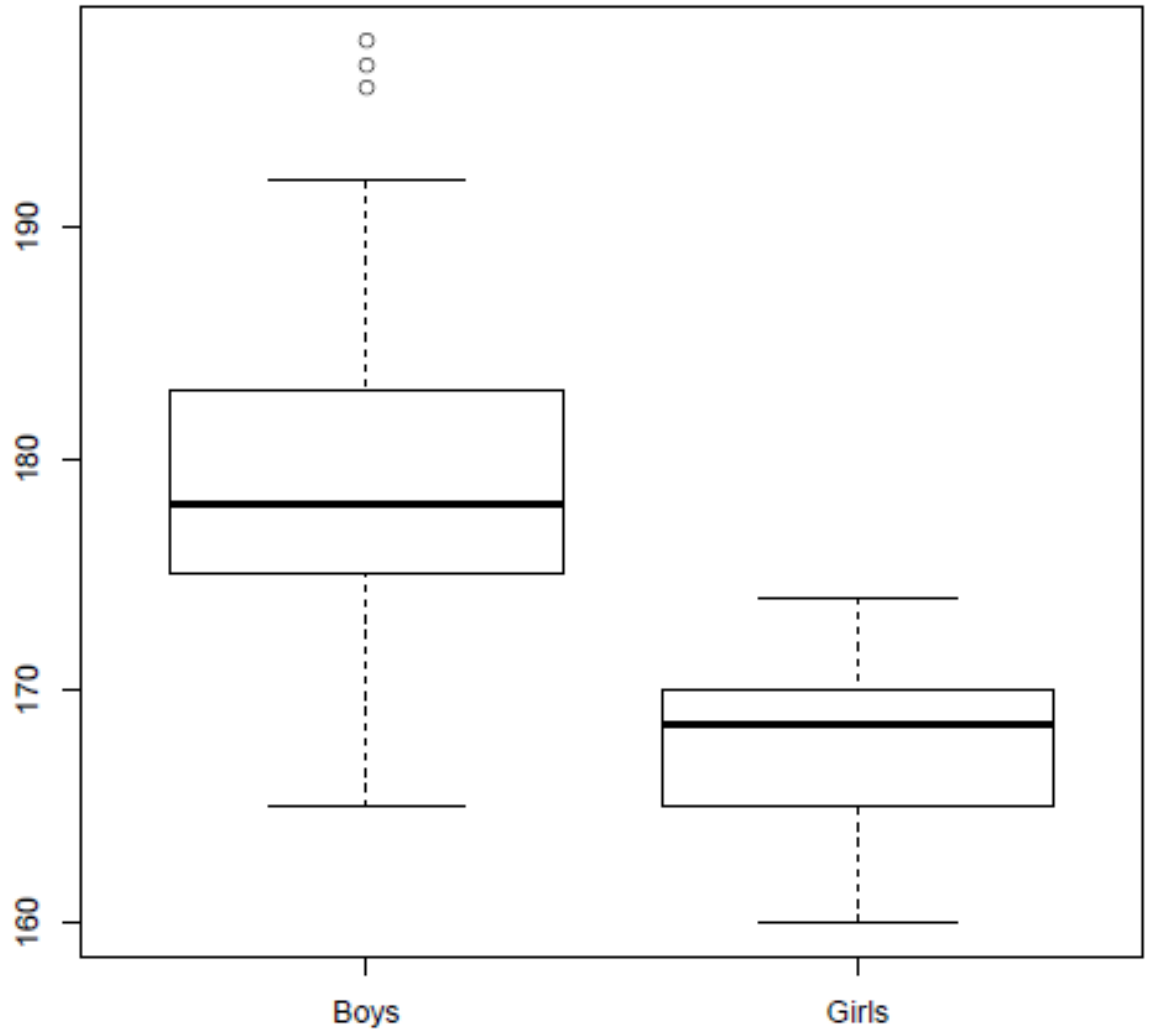
Test de la t-Student per dues mostres independents (σ^2 comuna)

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} =$$

$$= \frac{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}}}{\sqrt{\frac{\frac{(m-1)S_X^2}{\sigma^2} + \frac{(n-1)S_Y^2}{\sigma^2}}{m+n-2}}}$$

Per tant, la distribució sota H_0 correspon a la del quocient entre una N(0,1) i l'arrel quadrada d'una chi-quadrat amb m+n-2 graus de llibertat dividit pel seus graus de llibertat. Això correspon a la definició d'una t_{m+n-2}

Exemple: Alçades de nois i noies



Exemple: Alçades de nois i noies

Nois: $\bar{X} = 179.506 \qquad S = 6.5 \qquad n = 77$

Noies: $\bar{Y} = 167.5 \qquad S = 4.363 \qquad n = 14$

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

$$S_p^2 = \frac{(77-1)6.5^2 + (14-1)4.363^2}{77+14-2} = 38.86232$$

$$T = \frac{179.506 - 167.5}{\sqrt{38.86232}\sqrt{\frac{1}{77}+\frac{1}{14}}} = 6.628885$$

1. Regió Crítica: $C = \{\underset{\sim}{X}, \underset{\sim}{Y} : |T| \geq t_{89,0.975} = 1.986979\}$
2. P-valor$=2P(t_{89} > 6.628885) = 2.52 * 10^{-9}$

$$IC_{0.95}(\mu_1 - \mu_2) = 179.5 - 167.5 \pm t_{89,0.975}\sqrt{38.86}\sqrt{\frac{1}{77}+\frac{1}{14}} = [8.41, 15.61]$$

Exemple: Alçades de nois i noies (output de R)

```
Two Sample t-test

data:  Height by Sexe

t = 6.6289, df = 89, p-value = 2.524e-09

alternative hypothesis:  true difference in means is not equal to 0

95 percent confidence interval:

8.407601 15.605386

sample estimates:

mean in group 0 mean in group 1

179.5065 167.5000
```

Test de la t-Student per dues mostres independents (σ^2 no comuna)

Tenim

$$\bar{X} - \bar{Y} \sim N\left[\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right]$$

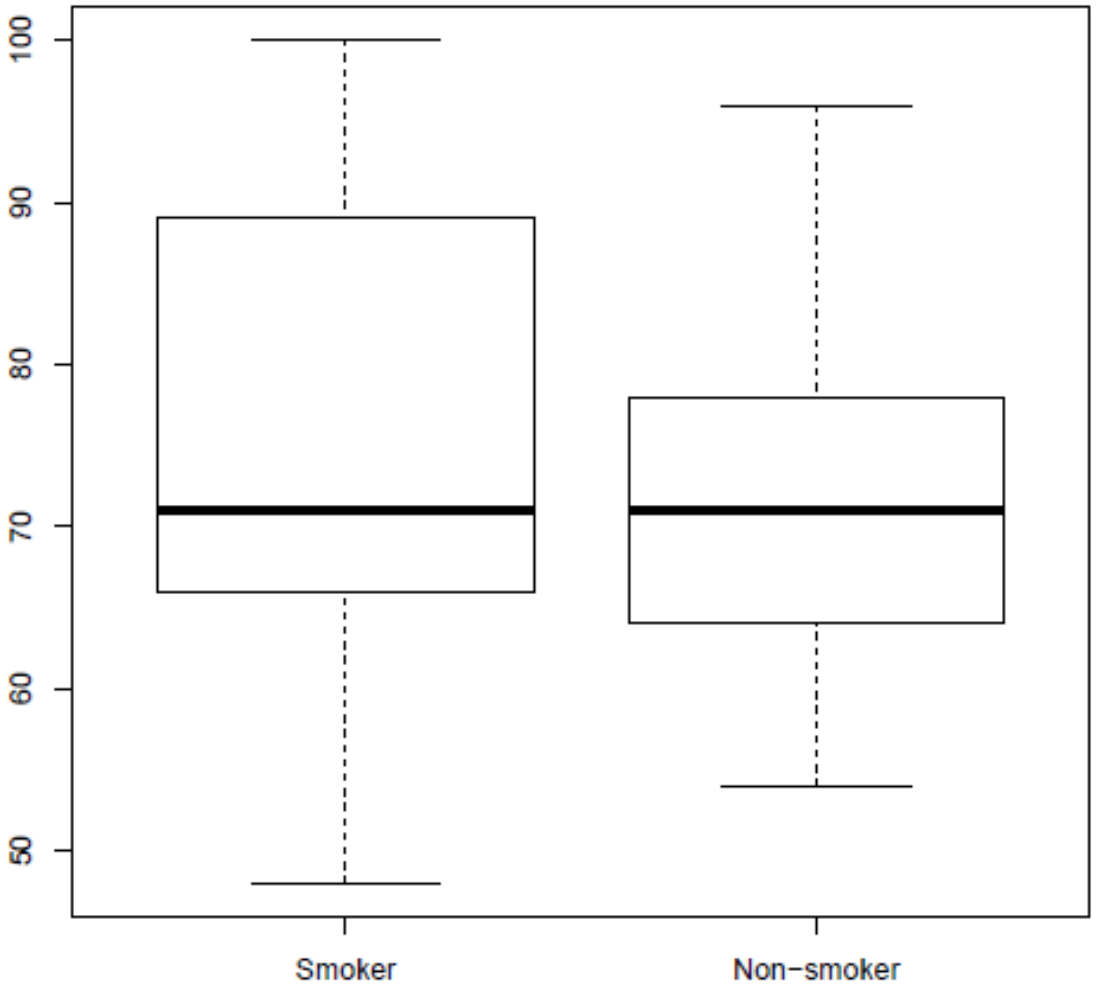
$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} | H_0 \sim N(0,1)$$

$$T' = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} | H_0 \sim t_{\hat{\nu}}$$

La distribució exacta de l'estadístic T' es complicada, però pot ser aproximada per una distribució t-Student amb graus de llibertat modificats $\hat{\nu}$

$$\hat{\nu} = \frac{\left(\frac{S_1^2}{m} + \frac{S_2^2}{n}\right)^2}{\frac{S_1^4}{m^2(m-1)} + \frac{S_2^4}{n^2(n-1)}}$$

Exemple: Pols de fumadors i no fumadors



Exemple: Pols de fumadors i no fumadors

Fumadors: $\bar{X} = 75 \qquad S = 13.493 \qquad n = 28$

No fumadors: $\bar{Y} = 71.938 \qquad S = 9.702 \qquad n = 64$

$$\hat{\nu} = \frac{\left(\frac{13.493^2}{28} + \frac{9.702^2}{64}\right)^2}{\frac{13.493^4}{28^2 27} + \frac{9.702^4}{64^2 63}} = 39.72$$

$$T' = \frac{75 - 71.938}{\sqrt{\frac{13.493^2}{28} + \frac{9.702^2}{64}}} = 1.085$$

```
Welch Two Sample t-test

data:  Pulse by Smoke

t = 1.0846, df = 39.723, p-value = 0.2847

alternative hypothesis:  true difference in means is not equal to 0

95 percent confidence interval:

-2.645663 8.770663

sample estimates:

mean in group Smoker mean in group Non-smoker

75.0000 71.9375
```

Test de comparació de variàncies de dues mostres independents

$$\begin{cases} H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2 \Leftrightarrow \sigma_X^2/\sigma_Y^2 = 1 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 \Leftrightarrow \sigma_X^2/\sigma_Y^2 \neq 1 \end{cases}$$

$$\frac{S_X^2}{S_Y^2} = \frac{\frac{(m-1)S_X^2}{\sigma_X^2}\frac{\sigma_X^2}{m-1}}{\frac{(n-1)S_Y^2}{\sigma_Y^2}\frac{\sigma_Y^2}{n-1}} = \frac{U/(m-1)}{V/(n-1)}\frac{\sigma_X^2}{\sigma_Y^2}$$

on $U \sim \chi^2_{m-1}$ i $V \sim \chi^2_{n-1}$. Sota H_0 es compleix que $\frac{\sigma_X^2}{\sigma_Y^2} = 1$ i per tant

$$F = \frac{S_X^2}{S_Y^2}|H_0 \sim F_{m-1,n-1}$$

Aquest test és un test de raó de versemblança.

Exemple: Pols de fumadors i no fumadors

Fumadors: $\bar{X} = 75 \qquad S = 13.493 \qquad n = 28$

No fumadors: $\bar{Y} = 71.938 \qquad S = 9.702 \qquad n = 64$

$$F = \frac{13.493^2}{9.702^2} = 1.9344$$

1. Regió Crítica: $C = \{\underset{\sim}{X}, \underset{\sim}{Y} : F \leq F_{27,63,0.025} \text{ ó } F \geq F_{27,63,0.975} = 1.8334\}$
2. P-valor$=2P(F_{27,63} > 1.9344) = 0.0328$

```
F test to compare two variances

data:  Pulse by Smoke

F = 1.9344, num df = 27, denom df = 63, p-value = 0.0328

alternative hypothesis:  true ratio of variances is not equal to 1

95 percent confidence interval:

1.055115 3.859449

sample estimates:

ratio of variances

1.934427
```

Test d'una proporció d'una mostra

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

Estadístic de test:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} | H_0 \approx N(0,1)$$

Exemple

En 1983, en Espanya van nèixer 246124 nens i 229619 nenes. Podem considerar que la probabilitat de que un nou nascut sigui nen és 0.5?

X: Nou nascut és nen $\sim Bern(p)$

$$\begin{cases} H_0 : p = 0.5 \\ H_1 : p \neq 0.5 \end{cases}$$

$$\hat{p} = 0.51735$$

$$Z = \frac{0.51735 - 0.5}{\sqrt{0.51735(1 - 0.51735)}} = 23.94$$

P-valor$= 2P(Z \geq 23.94) = 10^{-126}$

$$IC_{0.95}(p) = [0.5159, 0.5187]$$

Exemple

```
> prop.test(boy,tot)

-sample proportions test with continuity correction

data:  boy out of tot, null probability 0.5

X-squared = 572.5402, df = 1, p-value < 2.2e-16

alternative hypothesis:  true p is not equal to 0.5

95 percent confidence interval:

0.5159254 0.5187674

sample estimates:

p

0.5173466
```

Resum

Test	Hypothesis	Statistic	Distribution
One-sample Z	$H_o : \mu = \mu_0$	$Z = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$	$N(0,1)$
	$H_1 : \mu \neq \mu_0$		
One-sample Z	$H_o : p = p_0$	$Z = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}$	$N(0,1)$
(proportion)	$H_1 : p \neq p_0$		
One-sample T	$H_o : \mu = \mu_0$	$T = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$	t_{n-1}
	$H_1 : \mu \neq \mu_0$		
Two-sample T	$H_o : \mu_D = 0$	$T = \frac{\overline{D}-\mu_D}{s_{\overline{D}}}$	t_{n-1}
(paired)	$H_1 : \mu_D \neq 0$		
Two-sample T	$H_o : \mu_x = \mu_y$	$T = \frac{\overline{X}_m-\overline{Y}_n-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$	t_{m+n-2}
(independent)	$H_1 : \mu_x \neq \mu_y$		
Two-sample T	$H_o : \mu_x = \mu_y$	$T = \frac{\overline{X}_m-\overline{Y}_n-(\mu_1-\mu_2)}{\sqrt{\frac{s_m^2}{m}+\frac{s_n^2}{n}}}$	$t_{\hat{\upsilon}}$
(independent)	$H_1 : \mu_x \neq \mu_y$		
Two-sample F	$H_o : \sigma_x^2 = \sigma_y^2$	$F = \frac{s_x^2}{s_y^2}$	$F(n_x-1, n_y-1)$
	$H_1 : \sigma_x^2 \neq \sigma_y^2$		

Test de la Chi-quadrat per bondat d'ajust

Exemples:

- Tenim una mostra del nombre de cotxhe esperant en una benzinera i volem saber si aquest número segueix una distribució de Poisson
- Tenim una mostra de la llargada de l'ala d'un conjunt de mosques i volem saber si aquesta llargada segueix una distribució Normal

El test que volem resoldre és un cas de **bondat d'ajust**:

$$\begin{cases} H_0 : X \sim F_\theta \\ H_1 : X \text{ no segeueix } \sim F_\theta \end{cases}$$

Els paràmetres θ poden ser coneguts o no.

Test de la Chi-quadrat per bondat d'ajust. Procediment

- Fem una partició de $\mathbb{R}$ en el suport de l'espai mostral en intervals disjunts $I_1, \ldots, I_k$
- Calculem $p_j^0 = P(X \in I_j | X \sim F_\theta)$ $\quad 1 \leq j \leq k$ i np_j^0 seran els valors **esperats** sota H_0
- Sigui N_j el nombre d'observacions de la mostra en I_j. Seran els valors **observats**:

$$N_j = \sum_{i=1}^{n} \mathbb{I}_{\{x_i \in I_j\}}$$

- Calculem:

$$X^2 = \sum_{i=1}^{n} \frac{(N_j - np_j^0)^2}{np_j^0}$$

Nota: $E(N_j | H_0) = np_j^0$

Test de la Chi-quadrat per bondat d'ajust

- Hi ha dues possibilitats:
  - La distribució sota H_0 està completament especificada
  - El tipus de distribució sota H_0 està definit, però els paràmetres s'han d'estimar.
- L'estadístic X^2 va ser proposat per Pearson (1900) i té una distribució assimptòtica amb $k-1$ graus de llibertat o $k-1-p$ en el segon cas (p=nombre de paràmetres a estimar)
- Els intervals I_j se solen escollir de forma que np_j^0 són aproximadament del mateix ordre, amb $np_j^0 \geq 5$
- La regió Crítica serà: $C = \{\underset{\sim}{X} : X^2 \geq \chi^2_{k-1,1-\alpha}\}$

Exemple 1

Durant 3 mesos es pren nota del nombre de clients que treuen diners d'un caixer electrònic. Volem saber si la distribució del nombre d'usuaris diaris pot ser una Poisson de paràmetre $\lambda = 2.7$

$$\begin{cases} H_0 : X \sim Poisson(2.7) \\ H_1 : X \nsim Poisson(2.7) \end{cases}$$

La mostra $X_1, \ldots, X_{92}$ amb $X_i \in \mathbb{N}$ és:

k	N_j
0	5
1	10
2	25
3	27
4	14
5	8
≥ 6	3

Exemple 1

Calculem

$$P(X_j = k|H_0) = e^{-2.7}\frac{2.7^k}{k!}$$

k	$P(X_j = k\|H_0)$	$E_j = nP(X_j = k)$
0	0.0672	6.18
1	0.1815	16.69
2	0.2450	22.54
3	0.2205	20.18
4	0.1488	13.69
5	0.0804	7.39
≥ 6	0.0567	5.22

L'estadístic de prova serà:

$$X^2 = \sum_{i=1}^{7}\frac{(N_j - E_j)^2}{E_j} = \frac{(5-6.18)^2}{6.18} + \cdots + \frac{(3-5.22)^2}{5.22} = 6.40$$

Sota H_0, la distribució de referència és una χ^2_{7-1}.

1. Regió Crítica: $C = \{\underset{\sim}{X} : X^2 \geq \chi^2_{6,1-\alpha}\}$. Amb $\alpha = 0.05$, $\chi^2_{6,0.95} = 12.59$

Exemple 1

Donada una mostra de 100 bombetes, el temps de funcionament sense interrupció es mesura en mesos. Els resultats estan recollits en la taula inferior. La mitjana mostral és $\bar{X} = 2.069$. Si X_j és el temps de vida de la bombeta j, volem testar:

$$\begin{cases} H_0 : X \sim exp(\mu) \\ H_1 : X \nsim exp(\mu) \end{cases}$$

X_j	N_j
$[0, 1)$	31
$[1, 2)$	30
$[2, 3)$	13
$[3, 4)$	10
$[4, 5)$	6
$[5, +\infty)$	10

Com μ no està especificat, l'estimem per màxima-versemblança: $\hat{\mu} = \bar{X} = 2.069$

Exemple 2

Calculem els valors esperats sota la hipòtesi nul · la $(X \sim exp(\mu))$

$$P(X \leq x) = 1 - e^{-x/\mu}$$

$$P(a \leq X < b) = (1 - e^{-b/\mu}) - (1 - e^{-a/\mu})$$

$[a, b)$	$P(a \leq X < b)$	E_j
$[0, 1)$	0.3833-0.0000=0.3833	38.3
$[1, 2)$	0.6196-0.3833=0.2364	23.6
$[2, 3)$	0.7654-0.6196=0.1458	14.6
$[3, 4)$	0.8553-0.7654=0.0899	9
$[4, 5)$	0.9108-0.8553=0.0554	5.5
$[5, +\infty)$	1.000-0.9108=0.0892	8.9

Estadístic Chi-quadrat de Pearson: $X^2 = \frac{(31-38.3)^2}{38.3} + \cdots + \frac{(10-8.9)^2}{8.9} = 3.565$

Distribució de referència sota H_0 (un paràmetre estimat): $\chi^2_{6-1-1} = \chi^2_4$

1. Regió Crítica: $C = \{\underset{\sim}{X} : X^2 \geq \chi^2_{4,1-\alpha}\}$. Amb $\alpha = 0.05$, $\chi^2_{4,0.95} = 9.487$
2. P-valor=$P(\chi^2_4 \geq 3.565) = 1 - P(\chi^2_4 \leq 3.565) = 1 - 0.53 = 0.47$

Notes

- El test de la Chi-quadrat de Pearson per a la bondat d'ajust és un test assimptòtic: els resultats seran correctes si la grandària de la mostra es gran.
- Aquest test depèn de la forma en que es particiona el suport de la variable. Amb diferent agrupacions podem obtenir diferent resultats
- Si el valor esperat d'una classe és inferior a 5, l'estadístic pot veure's afectat per la classe on es dona aquesta propietat, ja que la discrepància entre esperats i observats tindrà més pes que la resta.

Test per dades categòriques en taules de contingència

- Si volem establir si hi ha relació entre dues variables categòriques, el test corresponent s'anomena **test d'independència per a taules de contingència**. Per exemple, podem voler veure si la gent amb cabell ros tendeixen a tenir més ulls blaus que els morenos. La depèndencia entre dues variables implica que si coneixem el valor d'una de les variables en un individu, les probabilitats de l'altre variable dependran d'aquest valor.
    - Una població i dues característiques mesurades en cada individu: les variables són independents?
- Si volem establir si la distribució d'una variable és similar en vàries poblacions, el test s'anomena **test d'homogeneïtat per a taules de contingència**. Per exemple, la distribució dels tipus sangüinis en les poblacions d'esquimals, africans i europeus és la mateixa?
    - Vàries poblacions i una característica mesurada en cada individu: les poblacions són homogènies respecte a la variable?

Test per dades categòriques en taules de contingència

Tenim una mostra descrita per una taula de contingència mab m files i n columnes per a les variables A i B:

$$A \in \{A_1, \ldots, A_m\} \qquad B \in \{B_1, \ldots, B_n\}$$

Probabilitats conjuntes: $P_{ij} = P(A = A_i \cap B = B_j)$

Probabilitats marginal: $P_{i+} = P(A = A_i)$ i $P_{+j} = P(B = B_j)$

$$P_{i+} = \sum_{j=1}^{n} P_{ij} \qquad P_{+j} = \sum_{i=1}^{m} P_{ij}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} P_{ij} = 1$$

Test per dades categòriques en taules de contingència

Taula d'efectius observats (O_{ij}= individus amb $A = A_i$ i $B = B_j$)

	B_1	B_2	B_3	...	B_m	
A_1	O_{11}	O_{12}	O_{13}	...	O_{1k}	O_{1+}
A_2	O_{21}	O_{22}	O_{23}	...	O_{2k}	O_{2+}
:	:	:	:	...	:	:
A_m	O_{m1}	O_{m2}	O_{m3}	...	O_{mk}	O_{m+}
	O_{+1}	O_{+2}	O_{+3}	...	O_{+n}	N

Test d'independència:

$$\begin{cases} H_0 : P(A = A_i \cap B = B_j) = P(A = A_i)P(B = B_j) & \forall i, j \\ H_1 : \exists i, j \quad P(A = A_i \cap B = B_j) \neq P(A = A_i)P(B = B_j) \end{cases}$$

Test per dades categòriques en taules de contingència

Efectius esperats: $E_{ij} = NP_{ij}$, sota H_0, $E_{ij} = NP_{i+}P_{+j}$

Estimadors per P_{i+} i P_{+j}:

$$\hat{P}_{i+} = \frac{O_{i+}}{N} \qquad \hat{P}_{+j} = \frac{O_{+j}}{N}$$

Taula d'efectius esperats (E_{ij})

	B_1	B_2	$\ldots$	B_m
A_1	$O_{1+}O_{+1}/N$	$O_{1+}O_{+2}/N$	$\ldots$	$O_{1+}O_{+n}/N$
A_2	$O_{2+}O_{+1}/N$	$O_{2+}O_{+2}/N$	$\ldots$	$O_{2+}O_{+n}/N$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
A_m	$O_{m+}O_{+1}/N$	$O_{m+}O_{+2}/N$	$\ldots$	$O_{m+}O_{+n}/N$

Estadístic Chi-quadrat de Pearson: $X^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

Sota H_0, $X^2 \sim \chi^2$ amb $(m-1)(n-1)$ graus de llibertat

Regió Crítica: $C = \{\underset{\sim}{X} : X^2 \geq \chi^2_{(m-1)(n-1),1-\alpha}\}$

Exemple: Color de cabells i d'ulls en N=5387 individus

	Eye colour			
Hair colour	Light	Blue	Medium	Dark
Fair	688	326	343	98
Red	116	38	84	48
Medium	584	241	909	403
Dark	188	110	412	681
Black	4	3	26	85

	Eye colour				
Hair colour	Light	Blue	Medium	Dark	Total
Fair	688	326	343	98	1455
Red	116	38	84	48	286
Medium	584	241	909	403	2137
Dark	188	110	412	681	1391
Black	4	3	26	85	118
Total	1580	718	1774	1315	5387

Test d'independència:

$$\begin{cases} H_0 : \text{Colors de cabell i ulls són independents} \\ H_1 : \text{Colors de cabell i ulls són dependents} \end{cases}$$

Exemple: Color de cabells i d'ulls en N=5387 individus

Sota $H_0 : E_{ij} = O_{i+}O_{+j}/N$

	Light	Blue	Medium	Dark
Fair	426.75	193.93	479.15	355.17
Red	83.88	38.12	94.18	69.81
Medium	626.78	284.83	703.74	521.65
Dark	407.98	185.40	458.07	339.55
Black	34.61	15.73	38.86	28.80

Estadístic Chi-quadrat de Pearson:
$X^2 = \frac{(688-426.75)^2}{426.75} + \cdots + \frac{(85-28.8)^2}{28.8} = 1240.04$

Distribució de referència sota H_0 (un paràmetre estimat): $\chi^2_{(5-1)(4-1)} = \chi^2_{12}$

1. Regió Crítica: $C = \{\underset{\sim}{X} : X^2 \geq \chi^2_{12,1-\alpha}\}$. Amb $\alpha = 0.05$, $\chi^2_{12,0.95} = 21.03$
2. P-valor$=P(\chi^2_{12} \geq 1240.04) \approx 0$

Exemple: Test d'independència amb R

```
> X

Light Blue Medium Dark

Fair 688 326 343 98

Red 116 38 84 48

Medium 584 241 909 403

Dark 188 110 412 681

Black 4 3 26 85

> chisq.test(X)

Pearson's Chi-squared test

data:  X

X-squared = 1240.039, df = 12, p-value < 2.2e-16
```

Test d'Homogeneïtat en m poblacions

Considerem que la variable A identifica m poblacions. Per a cada població, disposem d'una mostra d'una variable categòrica que descriu una característica dels individus, B amb n nivells.

Volem saber si les probabilitats són homogènies per a cada població.

Test d'homogeneïtat:

$$\begin{cases} H_0 : P(B = B_j | A = A_i) = P(B = B_j) & \forall i, j \\ H_1 : \exists i, j \quad P(B = B_j | A = A_i) \neq P(B = B_j) \end{cases}$$

Sota H_0 les probabilitats comuns a totes les poblacions s'estimen $\hat{P}_{+j} = \frac{O_{+j}}{N}$

Taula d'efectius esperats ($E_{ij} = O_{i+}\hat{P}_{+j} = O_{i+}\frac{O_{+j}}{N}$)

	B_1	B_2	$\ldots$	B_m	
A_1	$O_{1+}O_{+1}/N$	$O_{1+}O_{+2}/N$	$\ldots$	$O_{1+}O_{+n}/N$	O_{1+}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	
A_m	$O_{m+}O_{+1}/N$	$O_{m+}O_{+2}/N$	$\ldots$	$O_{m+}O_{+n}/N$	O_{m+}

El procediment per resoldre el test és el mateix que el del test d'independència

Exemple: Facultats i sistemes operatius

Una empresa de software està interessada en les preferències dels sistemes operatius, i fa una entrevista a 400 estudiants de 4 facultats universitàries:

	Windows	Macintosh	Linux	Total
Informatics	21	9	70	100
Economics	40	39	21	100
Law	65	29	6	100
Chemistry	29	32	39	100
Total	155	109	136	400

Exemple: Facultats i sistemes operatius

Test d'homogeneïtat:

$$\begin{cases} H_0 : \text{Les proporcions dels diferents sistemes són homogènies en les 4 facultats} \\ H_1 : \text{Al menys una facultat té proporcions diferents} \end{cases}$$

Estadístic Chi-quadrat de Pearson:

$$X^2 = \frac{(21 - 100 * 155/400)^2}{100 * 155/400} + \cdots + \frac{(39 - 100 * 136/400)^2}{100 * 136/400} = 113.52$$

Distribució de referència sota H_0 (un paràmetre estimat): $\chi^2_{(4-1)(3-1)} = \chi^2_6$

1. Regió Crítica: $C = \{\underset{\sim}{X} : X^2 \geq \chi^2_{6,1-\alpha}\}$. Amb $\alpha = 0.05$, $\chi^2_{6,0.95} = 12.59$
2. P-valor=$P(\chi^2_6 \geq 113.52) \approx 0$

Naturalesa de l'associació

- Si es rebutja la independència (o l'homogeneïtat), és interessant determinar les categories que estan associades.
- Les contribucions del terme de cada cel · la a l'estadístic X^2 és informatiu
- Un diagrama de barres estratificat o un mosaic plot permet visualitzar la naturalesa de l'associació

Estadística descriptiva per taules de contingència