# Bondat d'ajust i independència

# Bondat d'ajust

Sovint desitjam saber si una mostra prové o no d'una distribució concreta

## Exemples

- Llençam un dau en l'aire moltes vegades, apuntam els resultats, i d'aquests resultats en volem deduir si el dau està trucat o no
- Hem emprat unes mostres petites en un t-test: perquè el resultat del contrast tengui sentit, aquestes mostres han de provenir de poblacions normals. És el cas?

# Bondat d'ajust

En aquests casos, farem un contrast

$$\begin{cases} H_0 : \text{la mostra prové de la distribució desitjada} \\ H_1 : \text{la mostra no prové de la distribució desitjada} \end{cases}$$

Com sempre:

- Si obtenim evidència que ens permeti rebutjar la hipòtesi nul·la, conclourem que la mostra no prové de la distribució desitjada
- Si no obtenim evidència que ens permeti rebutjar la hipòtesi nul·la, acceptarem que la mostra prové de la distribució desitjada

# Bondat d'ajust

Els tests es basaran bàsicament en

1. Comparar les freqüències observades amb les freqüències teòriques de la distribució que contrastam
2. Determinar si les freq. observades són prou diferents de les freq. teòriques com per poder rebutjar la hipòtesi nul·la

# Exemple 1

Tenim un dau i volem saber si està trucat o no

Si no està trucat, quan llençam el dau i miram el resultat $X$, cada resultat $x = 1, \ldots, 6$ té probabilitat $P(X = x) = 1/6$

Llençam el dau 120 vegades i obtenim els resultats següents:

| Resultat | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Freqüència | 20 | 22 | 17 | 18 | 19 | 24 |

Si el dau no estigués trucat, esperaríem obtenir 20 vegades cada resultat. Hi ha prou evidència que el dau estigui trucat?

# Test $\chi^2$ de Pearson

Suposem que tenim $n$ observacions. Calculam les freqüències observades de $k$ grups de resultats (classes; poden ser els resultats individuals). Aquestes classes han de cobrir tots els resultats possibles.

Volem contrastar si aquestes observacions segueixen una distribució totalment coneguda, per a la qual coneixem la probabilitat que una observació caigui dins cada una de les classes

El contrast és

$$\begin{cases} H_0 : \text{La població té aquesta distribució} \\ H_1 : \text{La població no té aquesta distribució} \end{cases}$$

# Test $\chi^2$ de Pearson

Per a cada classe $i$, diguem

- $o_i$: la freqüència absoluta observada d'aquesta classe
- $p_i$: la probabilitat que una observació pertanyi a aquesta classe si $H_0$ és certa
- $e_i$: la freqüència absoluta esperada, o teòrica, d'aquesta classe si $H_0$ és certa: $e_i = p_i \cdot n$

Rebutjarem $H_0$ si les $o_i$ són prou diferents de les $e_i$

# Test $\chi^2$ de Pearson

## Teorema

*L'estadístic de contrast*

$$X^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

*té aproximadament una distribució* $\chi^2_{k-1}$ *si*

- *n gran* ($n \geqslant 25$ *o 30*)
- *Les classes cobreixen tots els resultats possibles (a la pràctica:* $\sum_{i=1}^{k} e_i = \sum_{i=1}^{k} o_i$*)*
- *Totes les classes tenen prou probabilitat com per tenir-les en compte (a la pràctica:* $e_i \geqslant 5$ *per a tota classe* $i$*; això es pot relaxar una mica, però no ho farem)*

# Test $\chi^2$ de Pearson

Sigui $\chi_0$ el valor que pren l'estadístic de contrast

El p-valor del contrast és

$$P(\chi^2_{k-1} \geqslant \chi_0),$$

amb el significat usual

# Exemple 1

| | Valor obtingut | | | | | |
|---|---|---|---|---|---|---|
| Freqüència | 1 | 2 | 3 | 4 | 5 | 6 |
| Observada, $o_i$ | 20 | 22 | 17 | 18 | 19 | 24 |
| Esperada, $e_i$ | 20 | 20 | 20 | 20 | 20 | 20 |

$$\chi_0 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20}$$

```
> O=c(20,22,17,18,19,24)
> E=rep(20,6)
> sum((O-E)^2/E)
[1] 1.7
```

# Exemple 1

Volem fer el contrast

$$\begin{cases} H_0 : \text{El dau dóna distribució uniforme} \\ H_1 : \text{El dau està trucat} \end{cases}$$

Estam en les condicions del teorema, per tant l'estadístic de contrast segueix una llei $\chi_5^2$:

p-valor: $P(\chi_5^2 \geqslant 1.7) = \texttt{1-pchisq(1.7,5)} = 0.89$. Com que és més gran que 0.05, acceptam la hipòtesi nul·la.

Conclusió: No tenim proves que el dau estigui trucat.

# Codi R

La funció per realitzar un test $\chi^2$ amb R és

```
chisq.test(obs,p=probs)
```

on `obs` és la llista de freqüències observades i `probs` la llista de probabilitats de les observacions; si no s'especifica, s'entén que totes són iguales

La suma de les `probs` ha de ser 1

# Exemple 1

```
> freq.abs.obs.daus=c(20,22,17,18,19,24)
> chisq.test(freq.abs.obs.daus)
        Chi-squared test for given probabilities
data:  freq.abs.obs.daus
X-squared = 1.7, df = 5, p-value = 0.8889
```

Obtenim el valor de l'estadístic, `X-squared`, els graus de llibertat, `df`, i el p-valor, `p-value`

# Exemple 2

Un tècnic de medi ambient vol estudiar l'augment de temperatura de l'aigua a dos quilòmetres dels abocaments d'aigua autoritzats d'una planta industrial.

El responsable de l'empresa afirma que *aquests augments de temperatura segueixen una llei normal amb* $\mu = 3.5$ *dècimes de grau C i* $\sigma = 0.7$ *dècimes de grau C.*

El tècnic ho posa en dubte. Per decidir-ho, pren una mostra aleatòria d'observacions de l'augment de les temperatures (en dècimes de grau).

# Exemple 2

| Rang de temperatures | Freqüències |
|---|---|
| 1.45—1.95 | 2 |
| 1.95—2.45 | 1 |
| 2.45—2.95 | 4 |
| 2.95—3.45 | 15 |
| 3.45—3.95 | 10 |
| 3.95—4.45 | 5 |
| 4.45—4.95 | 3 |
| Total | 40 |

Hi ha evidència que la sospita del tècnic sigui vertadera, a un nivell de significació del 5%?

# Exemple 2

Les classes han de cobrir tots els resultats possibles.
Afegim les cues als resultats extrems.

| Rang de temperatures | Freqüències |
|---|---|
| menys de 1.95 | 2 |
| 1.95—2.45 | 1 |
| 2.45—2.95 | 4 |
| 2.95—3.45 | 15 |
| 3.45—3.95 | 10 |
| 3.95—4.45 | 5 |
| 4.45 o més | 3 |
| Total | 40 |

# Exemple 2

El contrast és:

$$\begin{cases} H_0 : \text{La distribució dels augments de temp.} \\ \qquad\qquad \text{és } N(3.5, 0.7) \\ H_1 : \text{La distribució dels augments de temp.} \\ \qquad\qquad \text{no és } N(3.5, 0.7) \end{cases}$$

Tenim les freqüències observades, cal calcular les freqüències teòriques

# Exemple 2

Sigui $X \sim N(3.5, 0.7)$

$$
\begin{aligned}
p_1 &= P(X \leqslant 1.95) \\
&= P\left(\frac{X-3.5}{0.7} \leqslant \frac{1.95-3.5}{0.7}\right) \\
&= P(Z \leqslant -2.21) = F_Z(-2.21) = 0.0136
\end{aligned}
$$

Per tant

$$e_1 = p_1 \cdot n = 0.0136 \cdot 40 = 0.54$$

# Exemple 2

Sigui $X \sim N(3.5, 0.7)$

$$\begin{aligned} p_2 &= P(1.95 \leqslant X \leqslant 2.45) \\ &= P\left(\frac{1.95-3.5}{0.7} \leqslant \frac{X-3.5}{0.7} \leqslant \frac{2.45-3.5}{0.7}\right) \\ &= P(-2.21 \leqslant Z \leqslant -1.5) \\ &= F_Z(-1.5) - F_Z(-2.21) = 0.0533 \end{aligned}$$

Per tant

$$e_2 = p_2 \cdot n = 0.0533 \cdot 40 = 2.13$$

# Exemple 2

Calculam a mà d'aquesta manera totes les freqüències esperades

| Rang de temperatures | $o_i$ | $e_i$ |
|---|---|---|
| menys de 1.95 | 2 | 0.54 |
| 1.95—2.45 | 1 | 2.13 |
| 2.45—2.95 | 4 | 5.92 |
| 2.95—3.45 | 15 | 10.29 |
| 3.45—3.95 | 10 | 10.67 |
| 3.95—4.45 | 5 | 6.97 |
| més de 4.45 | 3 | 3.48 |

Tenim freqüències esperades $< 5$, el test $\chi^2$ no es pot aplicar amb garanties

# Exemple 2

Agrupam a fi d'obtenir freqüències esperades $\geqslant 5$ amb el màxim de classes.

| Rang de temp. | $o_i$ | $o_i$ acum. | $e_i$ | $e_i$ acum. |
|---|---|---|---|---|
| menys de 1.95 | 2 | | 0.54 | |
| 1.95—2.45 | 1 | | 2.13 | |
| 2.45—2.95 | 4 | 7 | 5.92 | 8.59 |
| 2.95—3.45 | 15 | 15 | 10.29 | 10.29 |
| 3.45—3.95 | 10 | 10 | 10.67 | 10.67 |
| 3.95—4.45 | 5 | | 6.97 | |
| més de 4.45 | 3 | 8 | 3.48 | 10.45 |

# Exemple 2

Calculem l'estadístic de contrast amb les freqüències acumulades ($k = 4$ classes)

$$\chi_0 = \frac{(7-8.59)^2}{8.59} + \frac{(15-10.29)^2}{10.29} + \frac{(10-10.67)^2}{10.67} + \frac{(8-10.45)^2}{10.45} = 3.067$$

El p-valor és

$$P(\chi_3^2 \geqslant 3.067) = \text{entre } 0.35 \text{ i } 0.4$$

No hi ha evidència que els augments de temperatures observats no segueixin la llei normal esmentada.

# Exemple 2 amb R

```
> freq.abs.obs=c(2,1,4,15,10,5,3)
> n=sum(freq.abs.obs)
> lim.esq=c(-Inf,1.95+0.5*(0:5))
> lim.dret=c(1.95+0.5*(0:5),Inf)
> mu=3.5
> sigma=0.7
> prob.esp=pnorm(lim.dret,mu,sigma)
  -pnorm(lim.esq,mu,sigma)
> freq.abs.esp=n*prob.esp
> round(freq.abs.esp,2)
[1]  0.54  2.14  5.97 10.22 10.73  6.91  3.49
```

# Exemple 2 amb R

```
> chisq.test(freq.abs.obs,p=prob.esp)

        Chi-squared test for given probabilities

data:  freq.abs.obs
X-squared = 8.1337, df = 6, p-value = 0.2285

Warning message:
In chisq.test(freq.abs.obs, p = prob.esp) :
  Chi-squared approximation may be incorrect
```

R ens avisa que hi ha freqüències esperades inferiors a 5, i que per tant l'aproximació de l'estadístic del test a la distribució $\chi^2$ pot no ser correcta

# Exemple 2 amb R

Agrupem (ho haurem de fer a mà)

```
> freq.abs.obs.agrup=c(sum(freq.abs.obs[1:3]),
   freq.abs.obs[4:5],sum(freq.abs.obs[6:7]))
> prob.esp.agrup=c(sum(prob.esp[1:3]),
  prob.esp[4:5],sum(prob.esp[6:7]))
> chisq.test(freq.abs.obs.agrup,p=prob.esp.agrup)

     Chi-squared test for given probabilities

data:  freq.abs.obs.agrup
X-squared = 3.1531, df = 3, p-value = 0.3686
```

# Test $\chi^2$ amb paràmetres poblacionals desconeguts

De vegades ens interessarà contrastar si les observacions segueixen algun tipus determinat de distribució (Poisson, normal, ...) amb algun paràmetre indeterminat

En aquest cas, estimam el paràmetre a partir de les observacions

El test és exactament el mateix, excepte que alguns autors recomanen que al nombre de graus de llibertat de la $\chi^2$ li restem el nombre de paràmetres que estimam. Nosaltres seguirem aquesta recomanació.

# Exemple 3

Es vol determinar si el nombre de vegades que apareix la seqüència GATACA en una cadena d'ADN de longitud 1000 segueix una llei Poisson

Es prenen diverses mostres de cadenes d'ADN de longitud 1000 i s'hi compten els nombres de GATACA

| nombre $x_i$ de vegades que hi apareix GATACA | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| freqüència $o_i$ | 229 | 211 | 93 | 35 | 7 | 1 |

Hem fet $n = 229 + 211 + 93 + 35 + 7 + 1 = 576$ observacions

# Exemple 3

Volem realitzar el contrast

$$\begin{cases} H_0 : \text{La mostra prové d'una distribució } Po(\lambda) \\ H_1 : \text{La mostra no prové d'aquesta distribució} \end{cases}$$

Necessitam estimar el paràmetre $\lambda$.

# Exemple 3

$\lambda$ és el valor esperat d'una v.a. $Po(\lambda)$. El podem estimar amb la mitjana mostral:

$$\begin{aligned}\lambda &= \frac{229\cdot 0+211\cdot 1+93\cdot 2+35\cdot 3+7\cdot 4+1\cdot 5}{229+211+93+35+7+1}\\ &= \frac{535}{576}=0.929\end{aligned}$$

# Exemple 3

Ara calculam les probabilitats i les freqüències teòriques. Recordem que, si $X$ és una v.a. de Poisson,

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $o_i$ | 229 | 211 | 93 | 35 | 7 | 1 |
| $p_i$ | 0.395 | 0.367 | 0.170 | 0.053 | 0.012 | 0.002 |
| $e_i = p_i \cdot n$ | 227.49 | 211.34 | 98.17 | 30.40 | 7.06 | 1.31 |

# Exemple 3

Ara calculam les probabilitats i les freqüències teòriques. Recordem que, si $X$ és una v.a. de Poisson,

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| $o_i$ | 229 | 211 | 93 | 35 | 7 | 1 |
| $p_i$ | 0.395 | 0.367 | 0.170 | 0.053 | 0.012 | 0.002 |
| $e_i = p_i \cdot n$ | 227.49 | 211.34 | 98.17 | 30.40 | 7.06 | 1.31 |

No està bé! Recordau que les classes han de cobrir tots els resultats possibles!

Canviarem el 5 per "5 o més"

# Exemple 3

| $x_i$ | 0 | 1 | 2 | 3 | 4 | $\geqslant 5$ |
|---|---|---|---|---|---|---|
| $o_i$ | 229 | 211 | 93 | 35 | 7 | 1 |
| $p_i$ | 0.395 | 0.367 | 0.170 | 0.053 | 0.012 | 0.003 |
| $e_i = p_i \cdot n$ | 227.49 | 211.34 | 98.17 | 30.40 | 7.06 | 1.55 |

on $P(X \geqslant 5) = 1 - (P(X = 0) + \cdots + P(X = 4))$

Hi ha freqüències esperades petites: agruparem les dues últimes columnes.

# Exemple 3

| $x_i$ | 0 | 1 | 2 | 3 | 4 | $\geqslant 5$ |
|---|---|---|---|---|---|---|
| $o_i$ | 229 | 211 | 93 | 35 | 7 | 1 |
| $p_i$ | 0.395 | 0.367 | 0.170 | 0.053 | 0.012 | 0.003 |
| $e_i = p_i \cdot n$ | 227.49 | 211.34 | 98.17 | 30.40 | 7.06 | 1.55 |

on $P(X \geqslant 5) = 1 - (P(X = 0) + \cdots + P(X = 4))$

Hi ha freqüències esperades petites: agruparem les dues últimes columnes.

| $x_i$ | 0 | 1 | 2 | 3 | $\geqslant 4$ |
|---|---|---|---|---|---|
| $o_i$ | 229 | 211 | 93 | 35 | 8 |
| $p_i$ | 0.395 | 0.367 | 0.170 | 0.053 | 0.015 |
| $e_i = p_i \cdot n$ | 227.49 | 211.34 | 98.17 | 30.40 | 8.61 |

# Exemple 3

El nombre $k$ de classes és 5, el nombre $m$ de paràmetres estimats és 1, per tant considerarem que l'estadístic de contrast té distribució $\chi^2_3$.

El valor de l'estadístic és

$$\chi_0 = \sum_{i=1}^{5} \frac{(o_i - e_i)^2}{e_i} = 1.02$$

El p-valor és

$$P(\chi^2_3 \geqslant 1.02) = 0.796$$

Per tant no podem rebutjar que les observacions trobades no segueixin una llei de Poisson. Això significa que no hi ha evidència que les aparicions de GATACA en cadenes d'ADN de longitud 1000 no siguin aleatòries.

# Exemple 3 amb R

Ja prenem les dades agrupades

```
> freq.obs=c(229,211,93,35,8)
> probs=c(dpois(0:3,0.929),1-ppois(3,0.929))
> chisq.test(freq.obs,p=probs)

Chi-squared test for given probabilities

data:  freq.obs
X-squared = 1.0215, df = 4, p-value = 0.9065
```

# Exemple 3 amb R

Ja prenem les dades agrupades

```
> freq.obs=c(229,211,93,35,8)
> probs=c(dpois(0:3,0.929),1-ppois(3,0.929))
> chisq.test(freq.obs,p=probs)

Chi-squared test for given probabilities

data:  freq.obs
X-squared = 1.0215, df = 4, p-value = 0.9065
```

R ha calculat el p-valor prenent $\chi_4^2$ (no sap que hem estimat un paràmetre), nosaltres el calculam amb $\chi_3^2$

```
> 1-pchisq(1.0215,3)
[1] 0.7960498
```

# Test K-S

El test de Kolgomorov-Smirnov (K-S) serveix per contrastar si una mostra segueix o no una distribució contínua, sense restriccions sobre la mida de la mostra

Es pot emprar amb tota distribució contínua completament especificada

Per a distribucions concretes, mides de mostres concretes o quan estimam els paràmetres, existeixen tests específics millors, però no els veurem aquí (sí amb R)

# Test K-S

Partim d'una mostra $x_1, x_2, \ldots, x_n$, amb tots els valors diferents, i volem contrastar si ha estat produïda per una variable $X$ amb distribució $F_X$.

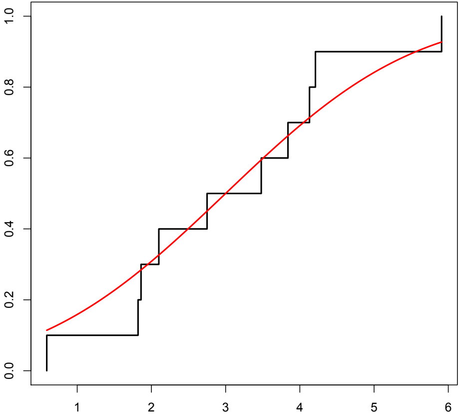
**(1)** Ordenam la mostra: $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$

**(2)** Calculam la funció de distribució mostral $F_n$ d'aquesta mostra, com si cada $x_{(i)}$ tingués probabilitat $1/n$

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \dfrac{k}{n} & \text{si } x_{(k)} \leqslant x < x_{(k+1)} \\ 1 & \text{si } x_{(n)} \leqslant x \end{cases}$$

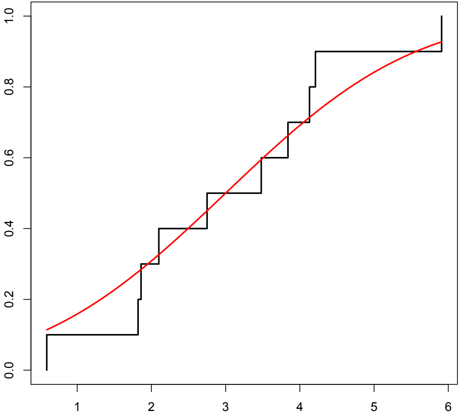
# Test K-S

**(3)** Comparam $F_n(x)$ amb $F_X(x)$. Si són molt diferents, podem rebutjar que la mostra prové de la variable $X$



# Test K-S

**(3)** Calculam $\sup\{|F_n(x) - F_X(x)| \mid x \in \mathbb{R}\}$. Com que $F_X$ és creixent, aquest suprem s'assoleix al voltant de qualque escaló.



# Test K-S

**(3)** Per fer-ho, calculam, per a cada $x_{(i)}$, la discrepància

$$\begin{aligned} D_n(x_{(i)}) &= \text{màx}\left\{ |F_X(x_{(i)}) - F_n(x_{(i)}^-)|, |F_X(x_{(i)}) - F_n(x_{(i)})| \right\} \\ &= \text{màx}\left\{ \left| F_X(x_{(i)}) - \frac{i-1}{n} \right|, \left| F_X(x_{(i)}) - \frac{i}{n} \right| \right\} \end{aligned}$$

(recordau $F(a^-) = \lim\limits_{t \to a^-} F(t)$)

# Test K-S

**(3)** ... i prenem la discrepància màxima

$$D_n = \text{màx}\{D_n(x_{(h)}) \mid h = 1, \ldots, n\}$$

Aquesta discrepància màxima segueix una distribució coneguda (que no depèn de la $X$ mentre sigui contínua) que permet calcular regions de rebuig i p-valors

# Test K-S

**Exemple**: Volem decidir si els valors

$$5.84, 4.57, 1.34, 3.58, 1.54, 2.25$$

provenen d'una distribució normal amb $\mu = 3$ i $\sigma = 1.5$.

Volem fer el contrast

$$\begin{cases} H_0: & \text{aquesta mostra prové d'una } X \sim N(3, 1.5) \\ H_0: & \text{aquesta mostra no prové d'una } X \sim N(3, 1.5) \end{cases}$$

# Test K-S

Ordenam la mostra: $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$

**Exemple**: Ordenam 5.84, 4.57, 1.34, 3.58, 1.54, 2.25

```
> x=c(5.84,4.57,1.34,3.58,1.54,2.25)
> sort(x)
[1] 1.34 1.54 2.25 3.58 4.57 5.84
```

# Test K-S

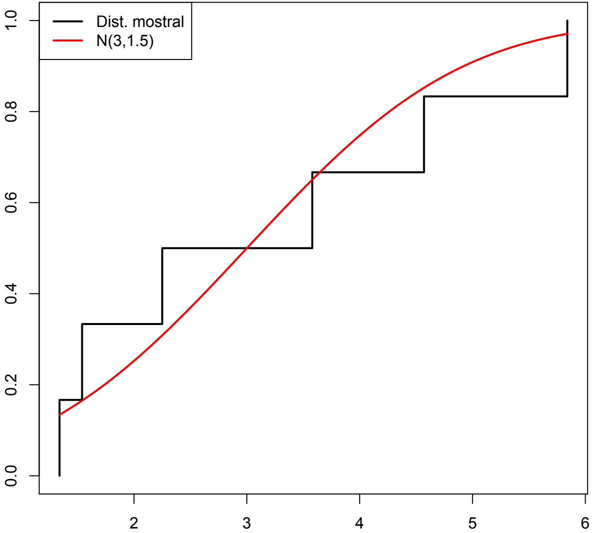
Calculam la funció de distribució mostral $F_n$ d'aquesta mostra com si cada $x_{(i)}$ tingués probabilitat $1/n$

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \dfrac{k}{n} & \text{si } x_{(k)} \leqslant x < x_{(k+1)} \\ 1 & \text{si } x_{(n)} \leqslant x \end{cases}$$

**Exemple**: Ordenats $1.34, 1.54, 2.25, 3.58, 4.57, 5.84$:

$$F_6(x) = \begin{cases} 0 & \text{si } x < 1.34 \\ 1/6 & \text{si } 1.34 \leqslant x < 1.54 \\ 2/6 & \text{si } 1.54 \leqslant x < 2.25 \\ 3/6 & \text{si } 2.25 \leqslant x < 3.58 \\ 4/6 & \text{si } 3.58 \leqslant x < 4.57 \\ 5/6 & \text{si } 4.57 \leqslant x < 5.84 \\ 1 & \text{si } 5.84 \leqslant x \end{cases}$$

# Test K-S



# Test K-S

Calculam la discrepància de cada observació $x_{(i)}$

$$D_n(x_{(i)}) = \text{màx}\left\{\left|F_X(x_{(i)}) - \frac{i-1}{n}\right|, \left|F_X(x_{(i)}) - \frac{i}{n}\right|\right\}$$

**Exemple**: Ordenats $1.34, 1.54, 2.25, 3.58, 4.57, 5.84$;

```
> round(pnorm(sort(x),3,1.5),3)
[1] 0.134 0.165 0.309 0.650 0.852 0.971
```

$$\begin{aligned} D_6(x_{(1)}) &= \text{màx}\{|F_X(1.34) - 0|, |F_X(1.34) - 1/6|\} \\ &= \text{màx}\{|0.134 - 0|, |0.134 - 1/6|\} \\ &= \text{màx}\{0.134, 0.033\} = 0.134 \end{aligned}$$

$$\begin{aligned} D_6(x_{(2)}) &= \text{màx}\{|F_X(1.54) - 1/6|, |F_X(1.54) - 2/6|\} \\ &= \text{màx}\{|0.165 - 1/6|, |0.165 - 2/6|\} \\ &= \text{màx}\{0.002, 0.168\} = 0.168 \end{aligned}$$

# Test K-S

Calculam la discrepància de cada observació $x_{(i)}$

$$D_n(x_{(i)}) = \text{màx}\left\{\left|F_X(x_{(i)}) - \frac{i-1}{n}\right|, \left|F_X(x_{(i)}) - \frac{i}{n}\right|\right\}$$

**Exemple**: Ordenats $1.34, 1.54, 2.25, 3.58, 4.57, 5.84$;

```
> round(pnorm(sort(x),3,1.5),3)
[1] 0.134 0.165 0.309 0.650 0.852 0.971
```

$$\begin{aligned} D_6(x_{(3)}) &= \text{màx}\{|F_X(2.25) - 2/6|, |F_X(2.25) - 3/6|\} \\ &= \text{màx}\{|0.309 - 2/6|, |0.309 - 3/6|\} \\ &= \text{màx}\{0.024, 0.191\} = 0.191 \end{aligned}$$

$$\begin{aligned} D_6(x_{(4)}) &= \text{màx}\{|F_X(3.58) - 3/6|, |F_X(3.58) - 4/6|\} \\ &= \text{màx}\{|0.65 - 3/6|, |0.65 - 4/6|\} \\ &= \text{màx}\{0.15, 0.017\} = 0.15 \end{aligned}$$

# Test K-S

Calculam la discrepància de cada observació $x_{(i)}$

$$D_n(x_{(i)}) = \text{màx}\left\{\left|F_X(x_{(i)}) - \frac{i-1}{n}\right|, \left|F_X(x_{(i)}) - \frac{i}{n}\right|\right\}$$

**Exemple**: Ordenats $1.34, 1.54, 2.25, 3.58, 4.57, 5.84$;

```
> round(pnorm(sort(x),3,1.5),3)
[1] 0.134 0.165 0.309 0.650 0.852 0.971
```

$$\begin{aligned} D_6(x_{(5)}) &= \text{màx}\{|F_X(4.57) - 4/6|, |F_X(4.57) - 5/6|\} \\ &= \text{màx}\{|0.852 - 4/6|, |0.852 - 5/6|\} \\ &= \text{màx}\{0.185, 0.019\} = 0.185 \end{aligned}$$

$$\begin{aligned} D_6(x_{(6)}) &= \text{màx}\{|F_X(5.84) - 5/6|, |F_X(5.84) - 6/6|\} \\ &= \text{màx}\{|0.971 - 5/6|, |0.971 - 1|\} \\ &= \text{màx}\{0.138, 0.029\} = 0.138 \end{aligned}$$

# Test K-S

Definim l'estadístic $D_n$ com la discrepància més gran:

$$D_n = \text{màx}\{D_n(x_{(h)}) \mid h = 1, \ldots, n\}$$

**Exemple**:

$$\begin{aligned} D_6 &= \text{màx}\{0.134, 0.168, 0.191, 0.15, 0.185, 0.138\} \\ &= 0.191 \end{aligned}$$

# Test K-S

La regla de decisió és rebutjar $H_0$ al nivell $\alpha$ si

$$D_n \geqslant D_{n,\alpha}$$

on $D_{n,\alpha}$ és el $\alpha$-quantil de la distribució del test K-S (teniu les taules a Campus Extens)

**Exemple**: Si prenem $\alpha = 0.05$, tenim que $D_{6,0.05} = 0.521$. Com que $0.191 < 0.521$, no podem rebutjar que la mostra provingui d'una variable $N(3, 1.5)$.

# Amb R

Per realitzar un test K-S amb R, tenim la instrucció

```
ks.test(x,"distribució",paràmetres)
```

on `x` és el vector que analitzam, la `distribució` és la distribució que contrastam, i els `paràmetres` són els de la distribució.

**Exemple**:

```
> x=c(5.84,4.57,1.34,3.58,1.54,2.25)
> ks.test(x,"pnorm",3,1.5)
      One-sample Kolmogorov-Smirnov test
data:  x
D = 0.1915, p-value = 0.95
alternative hypothesis: two-sided
```

Dóna el valor de l'estadístic, i un p-valor amb el significat usual

# Test K-S-Lilliefors

Quan es fa el test K-S per contrastar si una mostra provinent d'una distribució normal amb $\mu$ i $\sigma$ desconegudes, es recomana

- Estimar els paràmetres de la normal a partir de la mostra
- Calcular l'estadístic del test K-S amb aquests paràmetres
- Emprar a la decisió els $\alpha$-quantils $D'_{n,\alpha}$ de la distribució del test K-S-Lilliefors (teniu la taules a Campus Extens)

# Test K-S-Lilliefors

Amb R, és la funció `lillie.test` del paquet `nortest`

**Exemple**:

```
> install.packages("nortest",dep=TRUE)
...
> library(nortest)
> x=c(5.84,4.57,1.34,3.58,1.54,2.25)
> lillie.test(x)
     Lilliefors (Kolmogorov-Smirnov) normality
     test
data:  x
D = 0.1991, p-value = 0.6425
```

# Test $\chi^2$ d'independència en taules de contingència

Tenim una taula de contingència que ens dóna les freqüències absolutes conjuntes de dues característiques $X$ i $Y$ d'una població. Volem contrastar si aquestes dues característiques són variables aleatòries independents o no.

# Exemple 5

En un estudi d'una vacuna d'hepatitis hi participen 1083 voluntaris. D'aquests, es trien aleatòriament 549 i són vacunats. Els altres, 534, no són vacunats. Després d'un cert temps, s'observa que 70 dels 534 no vacunats han agafat l'hepatitis, mentre que només 11 dels 549 vacunats l'han agafada.

| | Vacunat? | |
|---|---|---|
| Emmalaltí? | Sí | No |
| Sí | 11 | 70 |
| No | 538 | 464 |

És el fet de contreure hepatitis independent d'haver estat vacunat contra la malaltia?

# Exemple 5

| | Vacunat? | |
|---|---|---|
| Emmalaltí? | Sí | No |
| Sí | 11 | 70 |
| No | 538 | 464 |

És el fet de contreure hepatitis independent d'haver estat vacunat contra la malaltia?

En aquest cas $2 \times 2$ és un contrast de proporcions per a dues mostres independents:

$$\begin{cases} H_0 : p_{\text{Vacunats}} = p_{\text{No vacunats}} \\ H_1 : p_{\text{Vacunats}} \neq p_{\text{No vacunats}} \end{cases}$$

Però en general ...

# Test $\chi^2$ d'independència

Considerem dues característiques $X$ i $Y$ d'una població que poden prendre els valors $X_1, \dots, X_I$ i $Y_1, \dots, Y_J$. Donam en una taula les freqüències absolutes de cada combinació de valors $(X_a, Y_b)$ en una mostra aleatòria de mida $n$

| $X \backslash Y$ | $Y_1$ | $Y_2$ | $\dots$ | $Y_j$ | $\dots$ | $Y_J$ | $n_{i\bullet}$ |
|---|---|---|---|---|---|---|---|
| $X_1$ | $n_{11}$ | $n_{12}$ | $\dots$ | $n_{1j}$ | $\dots$ | $n_{1J}$ | $n_{1\bullet}$ |
| $X_2$ | $n_{21}$ | $n_{22}$ | $\dots$ | $n_{2j}$ | $\dots$ | $n_{2J}$ | $n_{2\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $X_i$ | $n_{i1}$ | $n_{i2}$ | $\dots$ | $n_{ij}$ | $\dots$ | $n_{iJ}$ | $n_{i\bullet}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $X_I$ | $n_{I1}$ | $n_{I2}$ | $\dots$ | $n_{Ij}$ | $\dots$ | $n_{IJ}$ | $n_{I\bullet}$ |
| $n_{\bullet j}$ | $n_{\bullet 1}$ | $n_{\bullet 2}$ | $\dots$ | $n_{\bullet j}$ | $\dots$ | $n_{\bullet J}$ | $n$ |

# Exemple 5

| E\V | $V_1$ | $V_2$ | $n_{i\bullet}$ |
|---|---|---|---|
| $E_1$ | 11 | 70 | 81 |
| $E_2$ | 538 | 464 | 1002 |
| $n_{\bullet j}$ | 549 | 534 | 1083 |

# Test $\chi^2$ d'independència

$$\begin{cases} H_0: & \text{Les variables } X \text{ i } Y \text{ són independents} \\ H_1: & \text{Les variables } X \text{ i } Y \text{ no són independents} \end{cases}$$

Si diem

$$p_{ij} = P(X = X_i, Y = Y_j)$$
$$p_i = P(X = X_i) \qquad p_j = P(Y = Y_j)$$

el test d'independència equival a contrastar

$$\begin{cases} H_0: p_{ij} = p_i \cdot p_j \text{ per a tots } 1 \leqslant i \leqslant I,\ 1 \leqslant j \leqslant J \\ H_1: \text{no totes aquestes igualtats són veritat} \end{cases}$$

# Test $\chi^2$ d'independència

Emprarem l'estadístic

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} \cdot n_{\bullet j}}{n}}$$

que compara cada freqüència observada $n_{ij}$ amb la freqüència esperada si les variables fossin independents

$$\frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} \cdot n = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

Si $n$ és gran i cada freqüència esperada $\frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$ és $\geqslant 5$, aquest estadístic segueix aproximadament una llei $\chi^2$ amb $(I-1) \cdot (J-1)$ graus de llibertat

# Test $\chi^2$ d'independència

Com sempre, si $\chi_0$ és el valor que pren l'estadístic de contrast, el p-valor del contrast és

$$P(\chi^2_{(I-1)\cdot(J-1)} \geqslant \chi_0),$$

amb el significat usual

# Exemple 5

| E\V | $V_1$ | $V_2$ | $n_{i\bullet}$ |
|---|---|---|---|
| $E_1$ | 11 | 70 | 81 |
| $E_2$ | 538 | 464 | 1002 |
| $n_{\bullet j}$ | 549 | 534 | 1083 |

Freqüències esperades

$$\frac{n_{1\bullet} \cdot n_{\bullet 1}}{n} = \frac{81 \cdot 549}{1083} = 41.06$$
$$\frac{n_{1\bullet} \cdot n_{\bullet 2}}{n} = \frac{81 \cdot 534}{1083} = 39.94$$
$$\frac{n_{2\bullet} \cdot n_{\bullet 1}}{n} = \frac{1002 \cdot 549}{1083} = 507.94$$
$$\frac{n_{2\bullet} \cdot n_{\bullet 2}}{n} = \frac{1002 \cdot 534}{1083} = 494.06$$

# Exemple 5

Estadístic:

$$\chi_0 = \frac{(11-41.06)^2}{41.06} + \frac{(70-39.94)^2}{39.94} + \frac{(538-507.94)^2}{507.94} + \frac{(464-494.06)^2}{494.06} = 48.24$$

p-valor:

$$P(\chi_1^2 \geqslant 48.24) < 0.05$$

Per tant podem rebutjar la hipòtesi nul·la: vacunar-se i emmalaltir no són independents

# Exemple 6

Un investigador vol saber si el nombre de cries per lloba és independent de la zona on visqui

Considera 3 zones ($X$): $X_1 =$"Nord", $X_2 =$"Centre" i $X_3 =$"Sud"

Classifica els nombres de cries ($Y$) en $Y_1 =$" Dos o menys", $Y_2 =$" Entre tres i cinc", $Y_3 =$"Entre sis i vuit" i $Y_4 =$"Nou o més"

# Exemple 6

Pren una mostra de 200 llobes i obté la taula següent:

| $X \backslash Y$ | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $n_{i\bullet}$ |
|---|---|---|---|---|---|
| $X_1$ | 5 | 8 | 15 | 22 | 50 |
| $X_2$ | 20 | 26 | 46 | 8 | 100 |
| $X_3$ | 15 | 10 | 15 | 10 | 50 |
| $n_{\bullet j}$ | 40 | 44 | 76 | 40 | 200 |

# Exemple 6

Pren una mostra de 200 llobes i obté la taula següent:

| $X \backslash Y$ | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $n_{i\bullet}$ |
|---|---|---|---|---|---|
| $X_1$ | 5 | 8 | 15 | 22 | 50 |
| $X_2$ | 20 | 26 | 46 | 8 | 100 |
| $X_3$ | 15 | 10 | 15 | 10 | 50 |
| $n_{\bullet j}$ | 40 | 44 | 76 | 40 | 200 |

Les freqüències esperades si les variables són independents són

| $X \backslash Y$ | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $n_{i\bullet}$ |
|---|---|---|---|---|---|
| $X_1$ | 10 | 11 | 19 | 10 | 50 |
| $X_2$ | 20 | 22 | 38 | 20 | 100 |
| $X_3$ | 10 | 11 | 19 | 10 | 50 |
| $n_{\bullet j}$ | 40 | 44 | 76 | 40 | 200 |

# Exemple 6

Volem fer el contrast

$$\begin{cases} H_0: & \text{El nombre de cries per lloba és independent} \\ & \qquad\text{de la zona} \\ H_1: & \text{El nombre de cries per lloba no és independent} \\ & \qquad\text{de la zona} \end{cases}$$

Empram l'estadístic de contrast

$$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{(\text{freq.obs}_{ij} - \text{freq.esp}_{ij})^2}{\text{freq.esp}_{ij}}$$

$X^2$ segueix una llei $\chi^2$ amb $(3-1)\cdot(4-1) = 6$ graus de llibertat.

# Exemple 6

Calculam el valor de l'estadístic

$$\chi_0 = \ldots = 31.6$$

Calculam el p-valor

$$P(\chi_6^2 \geqslant 31.6) < 0.05$$

Per tant rebutjam la hipòtesi nul·la, i concloem que el nombre de cries per lloba és dependent de la zona on visqui.

# Exemple 6

Podem emprar la funció `chisq.test`, però hi hem d'entrar la taula en format `table`:

```
> dades = as.table(matrix(c(5,8,15,22,20,26,46,
 8,15,10,15,10),nrow=4,byrow=TRUE))
> dades
   A  B  C  D
A  5  8 15 22
B 20 26 46  8
C 15 10 15 10
> chisq.test(dades)
     Pearson's Chi-squared test
data: dades
X-squared = 31.6048, df = 6, p-value =1.942e-05
```

El $p$-valor és molt petit, rebutjam la hipòtesi nul·la

# Contrast d'homogeneïtat

Tenim una taula de contingència que ens dóna les freqüències absolutes conjuntes de dues característiques $X$ i $Y$ d'una població. Volem contrastar si, per a cada valor de $X$, les proporcions dels valors de $Y$ són les mateixes o no.

**Exemple:** En el nostre exemple de la vacuna

| | Vacunat? | |
|---|---|---|
| Emmalaltí? | Sí | No |
| Sí | 11 | 70 |
| No | 538 | 464 |

volem determinar si la proporció de malalts és la mateixa entre els vacunats que entre els que no estan vacunats

# Contrast d'homogeneïtat

És exactament el mateix test que el d'independència (si les variables són independents, les proporcions no variaran segons la filera o segons la columna)

Però el disseny de l'experiment sol ser diferent: usualment, l'experimentador tria *a priori* el nombre d'unitats experimentals per a cada valor $X_i$ de $X$ (és el que hem fet a l'exemple de les vacunes, però no el que hem fet a l'exemple dels llops)