

BLOC 4= Tema 6 + Tema 7: Proves no paramètriques basades en la llei χ^2 + Proves no paramètriques basades en rangs

LLIÇONS TEMA 6. Proves no paramètriques basades en la llei χ^2

- 6.1** La prova χ^2 de Pearson per l'ajust de la mostra a una distribució (pàgines Peña 461-463 + exemple 12.1)
- 6.2** Proves de normalitat (pàgines Peña 473-475 + exemple 12.6)
- 6.3** Proves χ^2 d'independència per a dades categòriques
- 6.4** Proves χ^2 d'homogeneïtat per a dades categòriques (pàgines Peña 508-512)

6.1 La prova χ^2 de Pearson per l'ajust de la mostra a una distribució

Fins ara hem suposat que coneixiem la llei d'on venien les dades, i a partir d'aquesta suposició proposavem estimadors, construïem intervals de confiança o feïem proves d'hipòtesis.

És a dir, treballem usualment **suposant** que les dades són una mostra aleatòria d'una distribució coneguda llevat del valor concret dels seus paràmetres.

Es tracta ara de desenvolupar proves que ens permetin estudiar si aquestes suposicions bàsiques no estan en contradicció amb la mostra observada, és a dir es tracta de fer un **diagnòstic del model**.

En aquest curs només fem una petita introducció i en cursos propers anireu ampliant i aprofundint.

Si voleu més informació i exhaustivitat podeu llegir tot el **capítol 12 de Peña**.

El diagnòstic del model

El diagnòstic del model pot incloure diversos procediments:

- 1 **Proves d'ajust o validesa (goodness-of-fit):** Per comprovar si la distribució suposada (per exemple la normal) és consistent amb les dades observades,
- 2 Proves per comprovar si les observacions són independents,
- 3 **Proves d'homogeneïtat:** Per comprovar si la mostra és homogènia, és a dir, si totes les observacions corresponen a la mateixa població

En aquest curs ens centrem en els punts 1 i 3 (per a dades categòriques).

La prova χ^2 de Pearson

La prova χ^2 deguda a Karl Pearson (1857-1936) és la prova de validesa més antiga.

Consisteix en comparar

- les **freqüències observades** de les dades mostrals i
- les **freqüències esperades** si el model teòric que s'està suposant és cert.

Aplicació a

- distribucions discretes i contínues,
- quan la grandària mostral es prou gran (al menys $n \geq 25$)
- la freqüència observada a cada categoria no és massa petita (al menys 3)

Exemples

- L : llargària de les ales d'una espècie de papallones.
A partir de les mesures de la longitud de les ales de 30 papallones, volem contrastar si la distribució de L és normal.
 $H_0 : X$ es distribueix com $N(\mu, \sigma^2)$ versus
 $H_1 : X$ no es distribueix com $N(\mu, \sigma^2)$
- X : nombre de cotxes fent cua en una benzinera a les 8 del vespre d'un dia feiner
A partir de les dades recollides durant 20 setmanes (100 dies de dilluns a divendres), volem contrastar si la distribució s'ajusta a un model de Poisson.
Mostra X_1, \dots, X_{100} de X =nombre de cotxes fent cua a les 8pm que pot prendre valors $0, 1, 2, \dots$
 $H_0 : X$ es distribueix com $\text{Poisson}(\lambda)$
 $H_1 : X$ NO es distribueix com $\text{Poisson}(\lambda)$
a on el paràmetre λ és en general desconegut.

Exemple clients caixer (exemple amb paràmetre conegut)

Una sucursal bancària està considerant afegir un caixer en el seu vestíbul per tal que no es produeixen cues a primeres hores del matí.

A partir de dades recollides anteriorment, saben que la mitjana del nombre de clients fent servir el caixer entre les 7:50 i les 8:00 és 2.7 ($\lambda = 2.7$).

Durant 3 mesos (92 dies) anoten el nombre de clients X que utilitzen el caixer entre les 7:50 i les 8:00. El rang de valors possibles per X és $\{0, 1, 2, 3, \dots\}$. Suposarem que el nombre de clients d'un dia és independent de l'altre.

Volem saber si X es comporta com una Poisson amb mitjana 2.7, és a dir, contrastem:

$$H_0 : X \sim \text{Poisson}(2.7) \quad \text{versus} \quad H_1 : X \not\sim \text{Poisson}(2.7)$$

Dades per l'exemple caixers

Disposen d'una mostra: X_1, X_2, \dots, X_{92} , on $X_i \in \{0, 1, 2, \dots\}$.
Sigui O_j el nombre de dies en que exactament j clients van fer servir el caixer.

classe l_j	j	O_j
l_0	0	5
l_1	1	10
l_2	2	25
l_3	3	27
l_4	4	14
l_5	5	8
l_6	més de 6	3

Podem escriure $O_j = \sum_{i=1}^{92} \mathbf{1}\{X_i = j\}$ per $0 \leq j \leq 5$ i
 $O_6 = \sum_{i=1}^{92} \mathbf{1}\{X_i \geq 6\}$

El procediment segueix la metodologia de les proves d'hipòtesis:

- 1 $H_0 : X$ es distribueix com F_θ
 $H_1 : X$ no es distribueix com F_θ .

A l'exemple dels clients al caixer:

$$H_0 : X \sim \text{Poisson}(2.7) \quad \text{versus} \quad H_1 : X \not\sim \text{Poisson}(2.7)$$

- 2 **Mesura de discrepància entre H_0 i les dades mostrals.**

Calculem primer la probabilitat, sota H_0 , que el model suposat assigna a cada classe, és a dir,

$$p_j = \mathbf{P}\{X \in I_j | X \sim F_\theta\} \quad 0 \leq j \leq 6.$$

Observem que $\sum_{j=0}^6 p_j = 1$.

Anomenem $E_j = np_j$ a la **freqüència esperada en la mostra de la classe j** d'acord amb el model F_θ .

Procediment a seguir (II)

- 2 Definició de la discrepància entre les freqüències observades i les esperades pel model:

$$\chi^2 = \sum_{j=0}^6 \frac{(\text{Observades}_j - \text{Esperades}_j)^2}{\text{Esperades}_j} = \sum_{j=0}^6 \frac{(O_j - E_j)^2}{E_j}$$

La distribució de l'estadístic χ^2 és aproximadament una χ^2 amb $\nu = 6$ graus de llibertat (són el nombre de classes que tenim menys 1).

- 3 Distribució de l'estadístic χ^2 sota la hipòtesi nul·la ens permet construir la regió de rebuig per decidir el test:

$$R = \{\mathbf{X} : \chi^2 \geq \chi^2_{\nu}(\alpha)\} \quad \text{on } \alpha = \mathbf{P}(\mathbf{X} \in R | H_0)$$

Càlcul probabilitats per l'exemple dels caixers

Sota H_0 la taula d'efectius esperats es calcula en base a la probabilitat teòrica de la Poisson:

$$\begin{aligned}\mathbf{P}(X \in I_j | X \sim \text{Poisson}(2.7)) &= \mathbf{P}(X = j | X \sim \text{Poisson}(2.7)) \\ &= e^{-2.7} \frac{2.7^j}{j!} \quad j = 0, 1, 2, 3, 4, 5\end{aligned}$$

$$\begin{aligned}\mathbf{P}(X \in I_6 | X \sim \text{Poisson}(2.7)) &= \mathbf{P}(X \geq 6 | X \sim \text{Poisson}(2.7)) \\ &= 1 - \sum_{i=0}^5 e^{-2.7} \frac{2.7^i}{i!}\end{aligned}$$

Taula efectius observats i esperats per l'exemple dels caixers

Les diferents probabilitats p_j i valors esperats E_j són les que es mostren a la taula

j	O_j	$p_j = \mathbf{P}(X = j)$	$E_j = n\mathbf{P}(X = j)$
0	5	0.0672	6.18
1	10	0.1815	16.69
2	25	0.2450	22.54
3	27	0.2205	20.28
4	14	0.1488	13.69
5	8	0.0804	7.39
més de 6	3	0.0567	5.22

Discrepància per l'exemple dels caixers

$$\begin{aligned} \chi^2 &= \sum_{j=0}^6 \frac{(\text{Observades}_j - \text{Esperades}_j)^2}{\text{Esperades}_j} = \sum_{j=0}^6 \frac{(O_j - E_j)^2}{E_j} \\ &= \frac{(5 - 6.18)^2}{6.18} + \frac{(10 - 16.69)^2}{16.69} + \frac{(25 - 22.54)^2}{22.54} \\ &+ \frac{(27 - 20.28)^2}{20.28} + \frac{(14 - 13.69)^2}{13.69} + \frac{(8 - 7.39)^2}{7.39} + \frac{(3 - 5.22)^2}{5.22} \\ &= 6.40 \end{aligned}$$

La distribució de l'estadístic χ^2 és aproximadament una χ^2 amb $\nu = 7 - 1 = 6$ graus de llibertat (són el nombre de classes que tenim menys 1).

Nota: Fixem-nos que hi han 7 classes porque les numerem de $j = 0$ a $j = 6$. En les fòrmules en general es posa de $j = 1$ a $j = k$

Conclusió a l'exemple dels caixers

Treballant amb $\alpha = 0.05$, la regió de rebuig per decidir el test és

$$R = \{\mathbf{X} : X^2 \geq \chi^2_\nu(0.05)\} \quad \text{on } 0.05 = \mathbf{P}(\mathbf{X} \in R | H_0)$$

Tenim que $\chi^2_6(0.05) = 12.59$, és a dir

$$\mathbf{P}(\chi^2_6 \geq 12.59) = 0.05$$

i per tant

$$R = \{\mathbf{X} : X^2 \geq 12.59\}$$

i com que hem observat $X^2 = 6.40 < 12.59 = \chi^2_6(0.05)$, acceptem H_0 , la distribució Poisson de mitjana 2.7 pot ser acceptada com a model per a les dades dels clients al caixer.

- 1 Aquesta prova no contrasta un model concret sino tots aquells models que donarien les mateixes probabilitats als intervals construïts. Per aquest motiu és recomanable que el nombre de classes k sigui gran i sempre ≥ 5 ($k \geq 5$).
- 2 Si les dades són contínues (normals, exponencials, etc), comencem fent una partició del rang de valors on X està definida, en k subintervals disjunts I_1, I_2, \dots, I_k (agrupem les n dades en k classes disjunctes)
Els intervals I_j s'han d'escollir de forma que $E_j = np_j$ siguin aproximadament iguals i tals que $E_j = np_j > 3$
- 3 És convenient que el nombre de dades a cada classe O_j sigui aproximadament el mateix i que sigui ≥ 3 ($O_j \geq 3$).
- 4 Cal notar que la prova és insensible a pautes de desviació sistemàtiques ja que es prenen les diferències al quadrat

Exemple bombardeig a Londres (exemple amb paràmetre desconegut)

Durant la Segona Guerra Mundial els alemanys varen bombardejar Londres. Es calcula que **varen fer caure 535 bombes**. Per esbrinar si el bombardeig anava dirigit a objectius militars o si pel contrari eren aleatoris es va dividir l'àrea de la ciutat de Londres en **576 quadrícules de $1/4 \text{ Km}^2$** i es va comptar quantes bombes varen caure per quadrícula. **Si el bombardeig era aleatori el nombre de bombes per quadrícula es distribuïria com una Poisson**. Ens plantegem, doncs, la prova següent:

$$H_0 : X \sim \text{Poisson}(\lambda) \quad H_1 : X \not\sim \text{Poisson}(\lambda)$$

j	O_j
0	229
1	211
2	93
3	35
4	7
5	1

Bombardeig a Londres (II)

Què fer quan el paràmetre λ és desconegut?

S'estima primer λ , es calcula $\mathbf{P}(X = j)$ amb el valor estimat de λ , es fa la taula amb els valors observats i els esperats.

Estimem $\hat{\lambda} = \bar{X} = \frac{\sum_{j=0}^5 j O_j}{576} = \frac{535}{576} = 0.929$ i per tant

$$\hat{p}_j = \widehat{\mathbf{P}(X = j)} = e^{-\hat{\lambda}} \frac{\hat{\lambda}^j}{j!} = e^{-0.929} \frac{0.929^j}{j!}$$

$$E_j = n \hat{p}_j = 576 \cdot e^{-0.929} \frac{0.929^j}{j!}$$

j	O_j	\hat{p}_j	E_j
0	229	0.395	227.5
1	211	0.367	211.3
2	93	0.170	98.1
3	35	0.053	30.4
4	7	0.012	7.1
≥ 5	1	0.003	1.5

Bombardeig a Londres (III)

Calculem la discrepància

$$\begin{aligned}\chi^2 &= \sum_{j=0}^5 \frac{(O_j - E_j)^2}{E_j} \\&= \frac{(229 - E_0)^2}{E_0} + \frac{(211 - E_1)^2}{E_1} + \frac{(93 - E_2)^2}{E_2} \\&\quad + \frac{(35 - E_3)^2}{E_3} + \frac{(7 - E_4)^2}{E_4} + \frac{(1 - E_5)^2}{E_5}\end{aligned}$$

$$\begin{aligned}\chi^2 &= \frac{(229 - 227.5)^2}{227.5} + \frac{(211 - 211)^2}{211} + \frac{(93 - 98)^2}{98} \\&\quad + \frac{(35 - 30)^2}{30} + \frac{(7 - 7)^2}{7} + \frac{(1 - 1.5)^2}{1.5} \\&= 0.01 + 0 + 0.26 + 0.83 + 0 + 0.17 = 1.27\end{aligned}$$

Bombardeig a Londres (IV)

Si les probabilitats $p_i(\lambda)$ es calculen estimant λ , el nombre de graus de llibertat s'ha de reduir en 1 unitat. En aquest cas hem fet 6 classes, el nombre de graus de llibertat és $6 - 1 - 1 = 4$.

H_0 : les dades s'ajusten a la llei de Poisson.

Si H_0 és cert, la discrepància $\chi^2 = \sum_{j=0}^5 \frac{(O_j - E_j)^2}{E_j}$ es distribueix com una χ_4^2 amb $6 - 1 - 1 = 4$ graus de llibertat.

NOTA: Si les probabilitats $p_i(\theta)$ es calculen estimant r paràmetres θ desconeguts, el nombre de graus de llibertat és $\nu = k - r - 1$.

Bombardeig a Londres (V)

Per un nivell de significació $\alpha = 0.05 \Rightarrow \chi_4^2(0.05) = 9.49$ la regió de rebuig és:

$$R = \{\mathbf{X} : X^2 \geq \chi_4^2(0.05)\} = \{\mathbf{X} : X^2 \geq 9.49\}$$

Hem obtingut $X^2 = 1.27 < 9.49$

No rebutgem H_0 , és a dir, podem treballar suposant que les dades es comporten com una Poisson.

Fixem-nos que $\text{Prob}\{\chi_4^2 \geq 1.92\} = 0.750$ i

$\text{Prob}\{\chi_4^2 \geq 5.39\} = 0.250$ i per tant el p-valor

$\text{Prob}\{\chi_4^2 \geq 1.27\} > 0.250$.

La conclusió és que durant la Segona Guerra Mundial els bombardejos dels alemanys sobre la ciutat de Londres varen ser indiscriminats i no varen apuntar a objectius militars concrets.

TREBALL INDEPENDENT Funcionament bombetes

Mesurem el temps de funcionament ininterromput en mesos, X , amb una mostra de 100 bombetes, amb l'objectiu de contrastar

$$H_0 : X_j \sim \exp(\lambda) \quad H_1 : X_j \not\sim \exp(\lambda)$$

Els valors mostrals x_1, \dots, x_{100} tenen una mitjana $\bar{x} = 2.069$ mesos i els agrupem en mesos fins al 5^e mes, és a dir, definim $I_1 = [0, 1)$, $I_2 = [1, 2)$, $I_3 = [2, 3)$, $I_4 = [3, 4)$, $I_5 = [4, 5)$ i $I_6 = [5, \infty)$.

La taula de freqüències és:

I_j	O_j
$[0, 1)$	31
$[1, 2)$	30
$[2, 3)$	13
$[3, 4)$	10
$[4, 5)$	6
$[5, \infty)$	10

Resoldre el contrast per decidir si el temps de vida de les bombetes és exponencial

Exemple funcionament bombetes (II)

Estimem el paràmetre λ pel mètode dels moments, és a dir,

$$\hat{\lambda} = \bar{X} = 2.069$$

Si $X \sim \exp(\lambda)$ aleshores

$$\mathbf{P}(X \leq x) = 1 - e^{-x/\lambda} \quad i$$

$$\mathbf{P}(a \leq X \leq b) = \left(1 - e^{-a/\lambda}\right) - \left(1 - e^{-b/\lambda}\right)$$

valors, que, en el nostre cas concret, representem a la següent taula

I_j	$p_j = \mathbf{P}(a \leq X \leq b)$	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, ∞)	$1 - 0.9108 = 0.0892$	8.9

Exemple funcionament bombetes (III)

Per tant, tenim que l'estadístic de Pearson serà:

$$\chi^2 = \frac{(31 - 38.3)^2}{38.3} + \dots + \frac{(10 - 8.9)^2}{8.9} = 3.5650$$

La distribució de referència és una χ^2 amb $\nu = k - 1 - 1 = 4$ graus de llibertat (ja que hem estimat un paràmetre).

La regió de rebuig és la següent:

$$R = \{\mathbf{X} : \chi^2 \geq \chi_4^2(\alpha)\} \quad \text{on } \chi_4^2(\alpha) \text{ és tal que } \mathbf{P}(\chi_4^2 \geq \chi_4^2(\alpha)) = \alpha$$

treballant, igual que abans, amb $\alpha = 0.05$, tenim que $9.487 = \chi_4^2(0.05)$ i com que nosaltres hem obtingut $\chi^2 = 3.565 < 9.487 = \chi_4^2(0.05)$, no trobem evidències estadístiques significatives que indiquin que la distribució de referència no sigui una exponencial i acceptem H_0 .

6.2 La prova χ^2 per validar la distribució normal

Per a validar la distribució normal existeixen molts tipus de proves:

- **Prova de Shapiro i Wilks:** representa la mostra en paper probabilístic normal i rebutja la normalitat quan el plot s'allunya d'una recta.
- **Prova de Kolmogorov-Smirnov-Lilliefors:** compara la llunyania de la funció de distribució empírica a la teòrica mitjançant l'estadístic de Kolmogorov-Smirnov. Valors significativament grans d'aquest estadístic condueixen a rebutjar la normalitat.
- **Prova de χ^2 de Pearson:** Ja l'hem explicat i l'aplicarem amb detall al cas normal
- **Proves d'asimetria i curtosis:** calculen quan lluny de la simetria i de la curtosis (apuntament) són les dades mostrals dels valors de referència normals (asimetria=0 i curtosis=3).

Si l'hipòtesi de normalitat es rebutja i es volen fer servir proves basades en la normalitat, es poden fer les **transformacions Box-Cox** en les dades per aconseguir la normalitat.

Llegir Peña pàgines 473-475.

Prova de χ^2 de Pearson per validar que una mostra és normal

Volem validar si les dades X_1, X_2, \dots, X_n provenen d'una llei $\text{Normal}(\mu, \sigma^2)$ amb μ i σ^2 desconeguts mitjançant una prova χ^2 .

$$H_0 : X \sim \text{Normal}(\mu, \sigma^2) \quad H_1 : X \not\sim \text{Normal}(\mu, \sigma^2)$$

Només l'utilitzarem quan $n \geq 100$.

Pasos a seguir:

- 1 Estimarem (μ, σ^2) per (\bar{X}, S^2)
- 2 Dividirem l'eix d'abscisses en k intervals equiprobables, I_1, \dots, I_k , és a dir, per $1 \leq j \leq k$

$$p_j = \frac{1}{k}; \quad E_j = \frac{n}{k}$$

- 3 Amb k classes (intervals), la prova χ^2 té $k - 1 - 2 = k - 3$ g.ll.
- 4 Per a construir els intervals I_j buscarem els percentils q_j de la distribució normal estàndar i construirem de tal manera que $p_j = \mathbf{P}\{X \in I_j | X \sim \text{Normal}(\mu, \sigma^2)\} = 1/k$.
- 5 Definim $a_j = \bar{x} + q_j s$. Els intervals I_j es defineixen com $I_j = (a_{j-1}, a_j]$ per $1 \leq j \leq k$. $I_1 = (-\infty, a_1]$, $I_2 = (a_1, a_2]$, \dots , $I_{k-1} = (a_{k-2}, a_{k-1}]$, $I_k = (a_{k-1}, +\infty]$.
- 6 Calcularem $X^2 = \sum_{j=1}^k (O_j - E_j)^2 / E_j$
- 7 Sota H_0 la distribució de l'estadístic serà una χ^2 amb $k - 3$ graus de llibertat.
- 8 La regió de rebuig és: $R = \{\mathbf{X} : X^2 \geq \chi^2_2(\alpha)\}$

Exemple 12.6 de Peña (pàgines 474-475)

Comprovarem si les següents dades venen d'una distribució normal

107.9; 96.7; 91, 2; 79; 103.1; 88; 101.3; 106; 93.7; 86; 100.7; 99.4

104.6; 117.2; 112.2; 106.9; 93; 88.3; 101.9; 109.8

(malgrat amb $n = 20$ no és el més indicat però és més fàcil per explicar-ho)

Aplicant-lo a les dades de Peña

- 1 $(\bar{X}, S^2) = (99.35; 9.56^2)$
- 2 Dividim l'eix d'abscisses en 5 intervals equiprobables, és a dir, $k = 5$, $p_j = 1/5 = 0.2$ i $E_j = 20/5 = 4$ (efectius esperats) per $1 \leq j \leq 5$.
- 3 La prova χ^2 té $k - 1 - 2 = 5 - 3 = 2$ g.ll.

4 .

$P(Z \leq q_j)$	1/5	2/5	3/5	4/5
q_j	-0.84	-0.26	+0.26	+0.84

- 5 Definim $a_j = \bar{x} + q_j s = 99.35 + q_j 9.56$

I_j	O_j	E_j	$(O_j - E_j)^2 / E_j$
$(-\infty; 91.32]$	5	4	$(5 - 4)^2 / 4 = 0.25$
$(91.32; 96.87]$	3	4	$(3 - 4)^2 / 4 = 0.25$
$(96.87; 101.83]$	3	4	$(3 - 4)^2 / 4 = 0.25$
$(101.83; 107.36]$	5	4	$(5 - 4)^2 / 4 = 0.25$
$(107.36; +\infty)$	4	4	$(4 - 4)^2 / 4 = 0$
TOTAL	20	20	1

6

$$\chi^2 = 0.25 + 0.25 + 0.25 + 0.25 + 0 = 1$$

- 7 Sota H_0 la distribució de l'estadístic serà una χ^2 amb $5 - 1 - 2 = 2$ graus de llibertat.
- 8 Treballant amb $\alpha = 0.05 \Rightarrow \chi^2_2(0.05) = 5.99$ i la regió de rebuig és:

$$R = \{\mathbf{X} : X^2 \geq \chi^2_2(0.05)\} = \{\mathbf{X} : X^2 \geq 5.99\}$$

i com que hem obtingut $X^2 = 1$ no hi ha cap dubte per dubtar de l'hipòtesi, és a dir, ens hem plantejat la prova següent:

$$H_0 : X \sim \text{Normal}(\mu, \sigma^2) \quad H_1 : X \not\sim \text{Normal}(\mu, \sigma^2)$$

i no tenim raons de pes per rebutjar-ho: l'ajust de les dades a la llei Normal és molt bo.

Observació: les dades s'havien generat d'una $N(100, 10^2)$.

Il·lustració definint l_1, \dots, l_8

Sigui $Z \sim N(0, 1)$. Si suposem $k = 8$, els percentils per Z

$\mathbf{P}(Z \leq q_j)$	1/8	2/8	3/8	4/8	5/8	6/8	7/8
q_j	-1.15	-0.68	-0.32	0	0.32	0.68	1.15

Aleshores els intervals $l_j = (a_{j-1}, a_j]$ per $1 \leq j \leq 8$ es construeixen a partir de q_j i de (\bar{x}, s^2) com $a_j = \bar{x} + q_j s$. Obtenim:

$$l_1 = (-\infty; \bar{x} - 1.15s], l_2 = (\bar{x} - 1.15s; \bar{x} - 0.68s]$$

$$l_3 = (\bar{x} - 0.68s; \bar{x} - 0.32s], l_4 = (\bar{x} - 0.32s; \bar{x}]$$

$$l_5 = (\bar{x}; \bar{x} + 0.32s], l_6 = (\bar{x} + 0.32s; \bar{x} + 0.68s]$$

$$l_7 = (\bar{x} + 0.68s; \bar{x} + 1.15s], l_8 = (\bar{x} + 1.15s; +\infty)$$

EXERCICI VOLUNTARI 5: Comprovant la normalitat

L'objectiu d'aquest exercici és practicar la comprovació de la normalitat d'un conjunt de dades.

Simulareu 4 conjunts de dades amb R:

- 1 Genereu 100 observacions Normals de mitjana $\mu = 30$ i variància $\sigma^2 = 16$.
- 2 Genereu 100 observacions Exponencials de mitjana $\mu = 30$.
- 3 Genereu 1000 observacions Normals de mitjana $\mu = 30$ i variància $\sigma^2 = 16$.
- 4 Genereu 1000 observacions Exponencials de mitjana $\mu = 30$.

Per a cada conjunt de dades heu de fer 3 proves d'ajustament per comprovar la normalitat. Una de les 3 proves és la χ^2 i per un cas l'heu de fer a mà. Les altres 2 proves les podeu escollir vosaltres a partir de les que ofereix R i heu d'explicar mínimament en què consisteix la prova.

Com a conclusió discutiu l'influència del canvi de grandària mostral -de $n = 100$ a $n = 1000$ - i l'influència d'haver generat exponencial o normal en els resultats obtinguts.

RESOLGUEU I LLIUREU EL DIMECRES 25 MAIG

6.3 Proves χ^2 d'independència entre variables categòriques

Siguin A i B dues variables de tipus categòric mesurades en una població.

Hi ha relació entre A i B, el coneixement del valor d'una d'elles (per exemple A) fa canviar les probabilitats de l'altre?

Exemple:

- Població: Estudiants d'un institut
- Variable categòrica A: Activitats extraescolars amb categories: 1=cap, 2=esportives, 3=artístiques, 4=complements.
- Variable categòrica B: Qualificació final amb categories: 1=suspès, 2=aprovat, 3=notable, 4=excel.lent.
- Es seleccionen a l'atzar 60 alumnes i per cadascun s'anota el valor d'A i de B
- Prova d'independència: Hi ha relació (associació) entre el fet de fer activitats extraescolars i les qualificacions?

Notació i taula de contingència

- A i B variables categòriques
- La variable A pren R valors $\{A_1, \dots, A_R\}$
- La variable B pren C valors $\{B_1, \dots, B_C\}$
- Hi ha relació entre A i B? Són A i B independents?.
- Seleccionem una mostra de grandària N i per a cada individu mesurem A i B. Comptem el nombre d'individus amb cada combinació A_i i $B_j \Rightarrow O_{ij}$ (freqüències observades) nombre d'individus amb $A = A_i$ i $B = B_j$ i ho resumim en forma d'una taula 2x2 anomenada **taula de contingència**.

	B_1	\dots	B_C	$TOTAL$
A_1	O_{11}	\dots	O_{1C}	O_{1+}
\cdot	\cdot		\cdot	\cdot
\cdot	\cdot		\cdot	\cdot
\cdot	\cdot		\cdot	\cdot
A_R	O_{R1}	\dots	O_{RC}	O_{R+}
$TOTAL$	O_{+1}	\dots	O_{+C}	N

Procediment teòric

- Les sumes de les files i de les columnes, tant pels valors teòrics com pels observats, s'anomenen **valors marginals**
- Sigui $P_{ij} = \mathbf{P}(A = A_i \text{ i } B = B_j)$

- Les marginals són $P_{i+} = \mathbf{P}(A = A_i) = \sum_{j=1}^C P_{ij}$, i

$$P_{+j} = \mathbf{P}(B = B_j) = \sum_{i=1}^R P_{ij}$$

- Observem que

$$\sum_{j=1}^C \sum_{i=1}^R P_{ij} = 1.$$

- Si A i B fossin independents es verificaria

$$P_{ij} = P_{i+} \cdot P_{+j}.$$

- La prova d'independència estableix, de forma equivalent,
 $H_0 : A \text{ i } B \text{ són independents} \Leftrightarrow P_{ij} = P_{i+} \cdot P_{+j}$ versus
 $H_1 : A \text{ i } B \text{ NO són independents} \Leftrightarrow P_{ij} \neq P_{i+} \cdot P_{+j}$

Taula de valors (efectius) esperats

- El número esperat d'individus amb $(A = A_i, B = B_j)$ és $E_{ij} = N \cdot P_{ij}$
- Sota $H_0 \Rightarrow E_{ij} = N(P_{i+} \cdot P_{+j})$.
- Els estimadors de P_{i+} i de P_{+j} són, respectivament:

$$\hat{P}_{i+} = \frac{\sum_{j=1}^C O_{ij}}{N} \quad \hat{P}_{+j} = \frac{\sum_{i=1}^R O_{ij}}{N}$$

- La taula de valors esperats serà:

	B_1	\dots	B_C
A_1	$(O_{1+} \cdot O_{+1})/N$	\dots	$(O_{1+} \cdot O_{+C})/N$
\cdot	\cdot		\cdot
\cdot	\cdot		\cdot
\cdot	\cdot		\cdot
A_R	$(O_{R+} \cdot O_{+1})/N$	\dots	$(O_{R+} \cdot O_{+C})/N$



$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^R \sum_{j=1}^C \frac{\left(O_{ij} - \frac{O_{i+} \cdot O_{+j}}{N}\right)^2}{\frac{O_{i+} \cdot O_{+j}}{N}}$$

- Sota $H_0 \Rightarrow \chi^2$ segueix una χ^2 amb $(R - 1) \times (C - 1)$ graus de llibertat (el nombre de paràmetres estimats són $R - 1$ per les probabilitats de A i $C - 1$ per les probabilitats associades a B)
- Regió de Rebuig

$$R = \{\mathbf{X} : \chi^2 > \chi^2_{(R-1)(C-1)}(\alpha)\}$$

on $\chi^2_{(R-1)(C-1)}(\alpha)$ és tal que $\mathbf{P}(\mathbf{X} \in R | H_0) = \alpha$.

Relació entre fumar i trastorns de la son

Volem establir si hi ha relació entre l'hàbit de fumar (A) i el fet de patir trastorns de la son (B). Amb una mostra de 150 individus la taula de contingència resultant és:

A	B		TOTAL
	NO	SI	
MOLT	7	30	37
MODERADAMENT	12	36	48
GENS	29	36	65
	48	102	150

La prova d'hipòtesi seria en aquest cas:

H_0 : L'hàbit de fumar i patir trastorns de la son NO estan relacionats

H_1 : L'hàbit de fumar i patir trastorns de la son estan relacionats
o equivalentment:

H_0 : A i B són independents H_1 : A i B NO són independents

Taules efectius observats i esperats per la relació entre fumar i trastorns de la son

A	B		TOTAL
	NO	SI	
MOLT	7	30	37 = O_{1+}
MODERADAMENT	12	36	48 = O_{2+}
GENS	29	36	65 = O_{3+}
	48 = O_{+1}	102 = O_{+2}	150

La taula esperada sota H_0 : $E_{ij} = \frac{O_{i+}O_{+j}}{150}$ és:

A	B		TOTAL
	NO	SI	
MOLT	$\frac{37 \times 48}{150} = 11.84$	$\frac{37 \times 102}{150} = 25.16$	37
MODERAD.	$\frac{48 \times 48}{150} = 15.36$	$\frac{48 \times 102}{150} = 32.6$	48
GENS	$\frac{65 \times 48}{150} = 20.8$	$\frac{65 \times 102}{150} = 4.42$	65
TOTAL	48	102	150

CONCLUSIÓ: Relació entre fumar i trastorns de la son

B

A	NO(Observats)	SI(Observats)	TOTAL
MOLT	7	30	37 = O_{1+}
MODERAD.	12	36	48 = O_{2+}
GENS	29	36	65 = O_{3+}
A	NO(Esperats)	SI(Esperats)	
MOLT	$\frac{37 \times 48}{150} = 11.84$	$\frac{37 \times 102}{150} = 25.16$	37 = O_{1+}
MODERAD.	$\frac{48 \times 48}{150} = 15.36$	$\frac{48 \times 102}{150} = 32.6$	48 = O_{2+}
GENS	$\frac{65 \times 48}{150} = 20.8$	$\frac{65 \times 102}{150} = 4.42$	65 = O_{3+}
TOTAL	48 = O_{+1}	102 = O_{+2}	150

CONCLUSIÓ: Relació entre fumar i trastorns de la son


B

A	NO(Observats)	SI(Observats)	TOTAL
MOLT	7	30	37 = O_{1+}
MODERAD.	12	36	48 = O_{2+}
GENS	29	36	65 = O_{3+}

A	NO(Esperats)	SI(Esperats)	
MOLT	$\frac{37 \times 48}{150} = 11.84$	$\frac{37 \times 120}{150} = 29.16$	37 = O_{1+}
MODERAD.	$\frac{48 \times 48}{150} = 15.36$	$\frac{48 \times 102}{150} = 32.6$	48 = O_{2+}
GENS	$\frac{65 \times 48}{150} = 20.8$	$\frac{65 \times 102}{150} = 44.2$	65 = O_{3+}
TOTAL	48 = O_{+1}	102 = O_{+2}	150

$$X^2 \sim \chi^2_{(3-1) \times (2-1)} = \chi^2_2 \text{ i } \chi^2_2(0.05) = 5.99, R = \{\mathbf{X} : X^2 > 5.99\}$$

$$X^2 = \frac{(7 - 11.84)^2}{11.84} + \dots + \frac{(36 - 44.2)^2}{44.2} = 8.744$$

Com que hem observat $X^2 = 8.744$ que és > 5.99 , rebutgem H_0 i concluïm que, amb una significació del 5%, hi ha alguna relació entre el fet de fumar i els trastorns de la son. 

Lectura Independent Depenen les notes de mates del professor?

Example of Chi-squared independence test

The headmaster of a large IB school is concerned that the maths results are dependent on the maths teacher. There are 3 SL teachers and the results for each class have been shown below.

These are the observed values.

Test at the 5% level of significance to see if the grades are independent of the teacher.

	1	2	3	4	5	6	7	Total
Mr. P	2	3	5	4	3	1	0	18
Ms. Q	1	2	5	6	4	1	1	20
Mrs. R	0	1	2	5	5	1	2	16
Total	3	6	12	15	12	3	3	54

Make your hypotheses:

H_0 : the grade at maths SL is independent of the teacher.

H_1 : the grade at maths SL is not independent of the teacher.

Make a table of expected values.

To do this take each row total \times column total and divide by the grand total.

This is shown opposite.

Find the expected number of grade 2s that Mr. P gets.

$$\frac{6 \times 18}{54} = 2$$

$$\frac{3 \times 18}{54} = 1$$

This value is the expected

Depenen les notes de mates del professor? (II)

continued

Observed

	1	2	3	4	5	6	7	Total
Mr. P	2	3	5	4	3	1	0	18
Ms. Q	1	2	5	6	4	1	1	20
Mrs. R	0	1	2	5	5	1	2	16
Total	3	6	12	15	12	3	3	54

Expected

	1	2	3	4	5	6	7	Total
Mr. P	1	2	4	5	4	1	1	18
Ms. Q	1.11	2.22	4.44	5.56	4.44	1.11	1.11	20
Mrs. R	0.89	1.78	3.56	4.44	3.56	0.89	0.89	16
Total	3	6	12	15	12	3	3	54

Calculate the chi squared test statistic:

$$\chi^2_{test} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2_{test} = 7.37 \quad \text{the } p \text{ value}$$

Find the critical value from your tables.

$$v = (7-1)(3-1) = 12$$

Critical value = 21.026

Make your conclusion:

Do not reject the null hypothesis. At the 5% level of significance there is no evidence to suggest that the choice of teacher influences the grade achieved.

6.4 Proves χ^2 d'homogeneïtat per relacionar dades categòriques

Sigui A una variable de tipus categòric que es pot mesurar en k poblacions diferents.

Prova d'homogeneïtat de les poblacions respecte a la característica que mesurem, és a dir, la distribució d'aquestes categories és la mateixa per a totes les poblacions?.

Exemple

- Poblacions: Habitants de tres països: un europeu (França), un asiàtic (Xina) i un africà (Senegal).
- Variable categòrica A: Grup sanguini amb categories 0, A, B, AB.
- Es seleccionen a l'atzar 20 habitants de cada país i per cadascun s'anota el valor d'A
- **Prova d'homogeneïtat** per saber si la freqüència dels diferents grups sanguinis és la mateixa pels habitants francesos, xinesos i senegalesos.

Es comporta de forma homogènia la variable categòrica B (que té k categories) en R poblacions diferents?

- R poblacions diferents, $i = 1, \dots, R$.
- Seleccionem una mostra aleatòria de grandària n_i ($i = 1, \dots, R$) de cada una de les R poblacions. $\sum_{i=1}^R n_i = N$
- Per a cada individu de cada una de les R poblacions mesurem el valor de la variable categòrica B que pren valors k valors diferents, $j = 1, \dots, k$

Són les probabilitats homogènies per a cada població, és a dir, $\mathbf{P}(B_j | \text{Pob. } 1) = \dots = \mathbf{P}(B_j | \text{Pob. } R) = \mathbf{P}(B_j)$ per a cada categoria j de B ?

Denotem per $p_{ij} = \mathbf{P}(B_j | \text{Població } i)$ ($i = 1, \dots, R, j = 1, \dots, k$) les probabilitats per cada una de les k categories en la població i . La prova d'hipòtesis es planteja equivalentment com:

$$\mathbf{H}_0 : \begin{array}{ccccccc} p_{11} & = & \dots & = & p_{R1} & = & \mathbf{P}(B_1) = p_1 \\ & \cdot & & & & & \cdot \\ & \cdot & & & & & \cdot \\ & \cdot & & & & & \cdot \\ p_{1k} & = & \dots & = & p_{Rk} & = & \mathbf{P}(B_k) = p_k \end{array}$$

versus $\mathbf{H}_1 : p_{ij} \neq p_j$ per algun $i = 1, \dots, R, j = 1, \dots, k$.

Taula de contingència

En una prova d'homogeneïtat cada fila de la taula de contingència correspon a una població, per exemple població i . Els k valors de la fila i es correspon amb les freqüències observades de les k categories de la variable B pels n_i individus la població i .

Notem que $n_i = O_{i+}$.

	B_1	\dots	B_k	$TOTAL$
Poblacio 1	O_{11}	\dots	O_{1k}	$n_1 = O_{1+}$
.	.		.	.
.	.		.	.
.	.		.	.
Poblacio R	O_{R1}	\dots	O_{Rk}	$n_R = O_{R+}$
$TOTAL$	O_{+1}	\dots	O_{+k}	N

Notem que els valors O_{ij} per $j = 1, \dots, k$ de la fila i són **independents** dels valors O_{hj} per $j = 1, \dots, k$ de la fila h , per qualsevol $i, h \in \{1, \dots, R\}$.

Taula d'efectius esperats

- Per cada categoria B_j ($j = 1, \dots, k$), si H_0 és certa, la millor estimació de la seva probabilitat $p_j = \mathbf{P}(B_j)$ és:

$$\hat{p}_j = \widehat{\mathbf{P}}(B_j) = \frac{O_{+j}}{N}$$

- La freqüència esperada E_{ij} de la categoria B_j dins de la població i ($i = 1, \dots, R$), sota H_0 , és $E_{ij} = n_i \frac{O_{+j}}{N} = O_{i+} \frac{O_{+j}}{N}$.
- La taula d'efectius esperats sota H_0 és:

	B_1	·	B_j	·	B_k	
Pob 1	$\frac{O_{1+} \times O_{+1}}{N}$	·	$\frac{O_{1+} \times O_{+j}}{N}$	·	$\frac{O_{1+} \times O_{+k}}{N}$	$n_1 = O_{1+}$
·	·	·	·	·	·	·
·	·	·	·	·	·	·
Pob i	$\frac{O_{i+} \times O_{+1}}{N}$	·	$\frac{O_{i+} \times O_{+j}}{N}$	·	$\frac{O_{i+} \times O_{+k}}{N}$	$n_i = O_{i+}$
·	·	·	·	·	·	·
·	·	·	·	·	·	·
Pob R	$\frac{O_{R+} \times O_{+1}}{N}$	·	$\frac{O_{R+} \times O_{+j}}{N}$	·	$\frac{O_{R+} \times O_{+k}}{N}$	$n_R = O_{R+}$
	O_{+1}	·	O_{+j}	·	O_{+k}	N

Mesura de discrepància

- La mesura de discrepància és l'estadístic de Pearson (com a les proves d'independència), és a dir,

$$X^2 = \sum_{i=1}^R \sum_{j=1}^k \frac{\left(O_{ij} - \frac{O_{i+} \cdot O_{+j}}{N}\right)^2}{\frac{O_{i+} \cdot O_{+j}}{N}}.$$

- Si H_0 és cert, $X^2 \sim \chi^2$ amb $(R-1) \times (k-1)$ graus de llibertat.
- Rebutgem H_0 si l'estadístic X^2 és prou gran. La regió de rebuig és

$$R = \{\mathbf{X} : X^2 > \chi^2_{(R-1)(k-1)}(\alpha)\}$$

on $\chi^2_{(R-1)(k-1)}(\alpha)$ és tal que $\mathbf{P}(\mathbf{X} \in R | H_0) = \alpha$.

Llegir Peña 12.4.4. pàgines 508-512. La notació que fem servir és diferent de la del Peña

Homogeneïtat dels sistemes operatius

Volem establir si la utilització de diferents sistemes operatius té una distribució diferent en dues facultats de la UPC. Es seleccionen 50 alumnes de cada centre i se'ls demana l'entorn informàtic que fan servir preferenment. Els resultats els presentem a la taula següent:

	WINDOWS	UNIX	MAC	
INFORMÀTICA	27	19	4	50
TELECOMUNICACIONS	20	22	8	50
	47	41	12	

Ens plantegem la següent prova d'hipòtesi: H_0 : Les distribucions d'ús són homogènies en els dos centres, és a dir, els estudiants d'informàtica i de telecomunicacions es comporten igual pel que fa a l'ús dels sistemes operatius

H_1 : Les distribucions no són homogènies

Taula d'efectius observats i esperats pels sistemes operatius

	WINDOWS	UNIX	MAC	
INFORMÀTICA	27	19	4	50
TELECOMUNICACIONS	20	22	8	50
	47	41	12	

La taula esperada sota la hipòtesi nul·la H_0 seria:

	WINDOWS	UNIX	MAC	
INFORM.	$\frac{47 \times 50}{100} = 23.5$	$\frac{41 \times 50}{100} = 20.5$	$\frac{12 \times 50}{100} = 6$	50
TELECOS	$\frac{47 \times 50}{100} = 23.5$	$\frac{41 \times 50}{100} = 20.5$	$\frac{12 \times 50}{100} = 6$	50
	47	41	12	

Homogeneïtat dels sistemes operatius: Conclusió

	WINDOWS (Obs)	UNIX (Obs)	MAC (Obs)	
INFORM.	27	19	4	50
TELECOS	20	22	8	50
	WINDOWS (Esp)	UNIX (Esp)	MAC (Esp)	
INFORM.	$\frac{47 \times 50}{100} = 23.5$	$\frac{41 \times 50}{100} = 20.5$	$\frac{12 \times 50}{100} = 6$	50
TELECOS	$\frac{47 \times 50}{100} = 23.5$	$\frac{41 \times 50}{100} = 20.5$	$\frac{12 \times 50}{100} = 6$	50
<i>TOTAL</i>	47	41	12	

$$\chi^2 = \frac{(27 - 23.5)^2}{23.5} + \frac{(19 - 20.5)^2}{20.5} + \frac{(4 - 6)^2}{6} + \frac{(20 - 23.5)^2}{23.5} + \frac{(22 - 20.5)^2}{20.5} + \frac{(8 - 6)^2}{6} = 2.595$$

$$\chi^2 \sim \chi^2_{(3-1) \times (2-1)} = \chi^2_2.$$

Homogeneïtat dels sistemes operatius: Conclusió

	WINDOWS (Obs)	UNIX (Obs)	MAC (Obs)	
INFORM.	27	19	4	50
TELECOS	20	22	8	50
	WINDOWS (Esp)	UNIX (Esp)	MAC (Esp)	
INFORM.	$\frac{47 \times 50}{100} = 23.5$	$\frac{41 \times 50}{100} = 20.5$	$\frac{12 \times 50}{100} = 6$	50
TELECOS	$\frac{47 \times 50}{100} = 23.5$	$\frac{41 \times 50}{100} = 20.5$	$\frac{12 \times 50}{100} = 6$	50
<i>TOTAL</i>	47	41	12	

$$\begin{aligned}
 \chi^2 = & \frac{(27 - 23.5)^2}{23.5} + \frac{(19 - 20.5)^2}{20.5} + \frac{(4 - 6)^2}{6} + \frac{(20 - 23.5)^2}{23.5} \\
 & + \frac{(22 - 20.5)^2}{20.5} + \frac{(8 - 6)^2}{6} = 2.595
 \end{aligned}$$

$$\chi^2 \sim \chi^2_{(3-1) \times (2-1)} = \chi^2_2.$$

Per $\alpha = 0.05$, $\chi^2_2(0.05) = 5.99$, i la regió crítica és

$R = \{\mathbf{X} : \chi^2 > 5.99\}$. Com que $\chi^2 = 2.595 < 5.99$

NO rebutgem H_0 : No tenim evidència per dir que les distribucions als dos centres NO són homogènies.