

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

1. Per contrastar si un dau està trucat o no es llença 60 vegades, obtenint els resultats següents:

x	1	2	3	4	5	6
n	7	12	10	11	8	12

Amb un nivell de significació $\alpha = 0.05$, pots acceptar que el dau no està trucat?

2. Les dades següents corresponen al diàmetre de la closca (mm) i a la separació entre espirals (μm) de 20 cargols. Contrasta amb un nivell de significació $\alpha = 0.05$ la hipòtesi de normalitat de les 2 variables.

Cargol	Diam	Separ
1	17	123
2	15	115
3	18	132
4	19	129
5	16	122
6	16	110
7	17	129
8	15	97
9	13	88
10	14	95
11	21	135
12	15	121
13	15	127
14	18	132
15	17	119
16	19	128
17	16	128
18	17	130
19	16	120
20	17	123

3. El CAP de l'Eixample vol estudiar si és necessari contractar un altre infermer per donar suport a les extraccions de sang de primera hora del mati (de 8 a 8 : 15). Se sap que la mitjana històrica de pacients (durant els últims 10 anys) que han anat a les extraccions en aquesta franja és de 5 pacients. Durant 2 mesos (61 dies), la directora del servei anota el nombre de pacients X que hi va. Observa que $X \in \{3, 4, 5, \dots\}$. El número de pacients d'un dia és independent del següent. Si O_j és el nombre de dies en que j pacients van anar a les extraccions de sang, les dades que obté són les següents:

j	O_j
3	3
4	9
5	18
6	23
7	5
≥ 8	3

La directora afirma que que la distribució de X segueix una Poisson amb mitjana 5. Demana a un estadístic que contrasti la hipòtesi

$$H_0 : X \sim \text{Poisson}(5) \quad \text{vs} \quad H_1 : X \not\sim \text{Poisson}(5)$$

Té raó la cap d'infermeres? La distribució del número de pacients segueix una Poisson(5)?

4. La Laia i en Josep agafen cada dia el metro a les 13:30 a la parada de Palau Reial. Cansats d'haver d'esperar tanta estona decideixen anotar durant 100 dies consecutius quant de temps triga a passar el metro un cop ells arriben a l'estació. Els resultats obtinguts es troben al document .RData amb el nom de 'metro'.

Proposa un model que s'ajusti a les dades i comprovar-ho amb $\alpha = 0.05$.

5. Volem esbrinar si estudiar (succés A) i aprovar (succés B) estan relacionats. Amb una mostra de 100 individus, la taula de contingències resultant és

		B		
		No	Sí	Total
A	Cada dia	2	8	10
	Cada setmana	10	40	50
	Abans de l'examen	28	12	40
	Total	40	60	100

Podem dir que hi ha relació entre estudiar i aprovar?

6. Volem saber si el tipus de música escoltada té una distribució diferent entre nois i noies. Per això, es seleccionen 60 nois i 60 noies i se'ls hi pregunta quin dels 3 tipus de música (rock, pop o house) prefereixen. S'obtenen els resultats següents:

	Rock	Pop	House	Total
Noies	20	15	25	60
Noies	19	22	19	60
Total	39	37	44	120

Les distribucions de la música escoltada són les mateixes per nois que per noies (és a dir, els nois i noies tenen el mateix comportament pel que fa a gustos musicals)?

7. En un estudi sobre creences religioses dels estudiants universitaris, es van recollir les següents dades

	Creients	Ateus	Agnòstics	Total
Ciències	10	45	45	100
Lletres	15	40	45	100
Total	25	85	90	200

És cert que les creences religioses són diferents entre els estudiants de ciències i els de lletres?

8. Un grup d'estudiants del grau d'Estadística ha decidit utilitzar el seu coneixement per apostar a la Quiniela. Una de les primeres coses que volen saber abans de començar a fer les seves apostes és veure si existeix relació entre el nombre de victòries, i el fet de jugar com a local o visitant. Per fer-ho han agafat les dades dels partits disputats a la temporada 13/14.

	Guanyat	Perdut	Total
Local	179	115	294
Visitant	115	179	294
Total	294	294	588

A partir d'aquestes dades, podem considerar que existeixen diferències entre el fet de jugar com a local o visitant?

9. Es vol estudiar si existeix relació entre el color dels ulls d'una persona i el seu nivell d'estudis. Per fer-ho hem realitzat una enquesta a 200 persones.

	Nivell d'estudis			
	Bàsic	Mig	Superior	Total
Marró	26	58	23	107
Blau	13	23	11	47
Altres	8	26	12	46
Total	47	107	46	200

Podem dir que existeix relació entre aquestes dues variables? Realitza els càlculs amb un nivell de significació $\alpha = 0.01$.

10. Es vol estudiar el comportament electoral entre les diferents circumscripcions catalanes. Per aquest motiu s'han agafat les dades dels cinc partits polítics més votats a les eleccions al Parlament de Catalunya de l'any 2012. A continuació es mostren els escons que va rebre cada partit a les diferents circumscripcions.

	Barcelona	Girona	Lleida	Tarragona	Total
CIU	26	9	8	7	50
ERC	12	3	3	3	21
PSC	14	2	1	3	20
PP	12	2	2	3	19
ICV-EUiA	10	1	1	1	13
Total	74	17	15	17	123

Ens plantegem si les distribucions dels escons atorgats a cada partit polític és el mateix a cada circumscripció o no.

Resolt el problema amb $\alpha = 0.05$.

Exercicis solucionats

- 1.
2. Analitzarem la variable Diam. Volem resoldre el contrast

$$H_0 : X \sim N(\mu, \sigma)$$

en front de l'alternativa

$$H_1 : X \not\sim N(\mu, \sigma)$$

Ens cal primer estimar els paràmetres. La mitjana mostral és $\bar{x} = 16.55$ i la desviació estandard és $s = 1.88$. Es tracta d'una variable continua que caldria discretitzar per aplicar el mètode χ^2 de bondat d'ajustament. Considerem 4 intervals equiprobables, és a dir, $p_j = 1/4$ amb $j = 1, \dots, 4$, per tant, les freqüències esperades seran $E_j = 20/4 = 5$. Amb $k = 4$ intervals la prova de la χ^2 tindrà $k - 2 - 1 = 4 - 3 = 1$ grau de llibertat.

Per construir els intervals I_j buscarem els percentils q_j de la distribució normal estandar i els construirem de tal manera que

$$p_j = P(X \in I_j | X \sim N(\mu, \sigma)) = 1/4$$

Calculem els percentils que necessitem

$P(Z \leq q_j)$	1/4	2/4	3/4
q_j	-0.67	0	0.67

Definim $a_j = \bar{x} + q_j s = 16.55 + q_j 1.88$ i construïm els intervals $I_j = (a_{j-1}, a_j]$ per $j = 1, \dots, 4$. Obtenim $I_1 = (-\infty, 15.29]$; $I_2 = (15.29, 16.55]$; $I_3 = (16.55, 17.81]$; $I_4 = (17.81, \infty)$. Ara ja podem construir la taula

interval	f. obs.	f. esp.	$(O_j - E_j)^2$	$(O_j - E_j)^2 / E_j$
$(-\infty, 15.29]$	6	5	1	0.2
$(15.29, 16.55]$	4	5	1	0.2
$(16.55, 17.81]$	5	5	0	0
$(17.81, \infty)$	5	5	0	0
TOTAL	20	20		0.4

Per tant, el valor experimental de l'estadístic és:

$$X^2 = 0.2 + 0.2 + 0 + 0 = 0.4.$$

Sota H_0 la distribució de l'estadístic serà una χ^2 amb $4 - 1 - 2 = 1$ grau de llibertat. Tenint en compte que treballem amb $\alpha = 0.05$, amb les taules de la χ^2 podem obtenir la regió de rebuig.

$$R = \{\mathbf{X} : X^2 \geq \chi_1^2(0.05)\} = \{\mathbf{X} : X^2 \geq 3.841\}$$

i com que hem obtingut $X^2 = 0.4$ no podem rebutjar H_0 i, per tant, les dades de la variable Diam s'ajusten a la llei Normal.

3. Escrivim la taula de les dades com

classe I_j	j	O_j
$I_3 = \{X = 3\}$	3	3
$I_4 = \{X = 4\}$	4	9
$I_5 = \{X = 5\}$	5	18
$I_6 = \{X = 6\}$	6	23
$I_7 = \{X = 7\}$	7	5
$I_8 = \{X \geq 8\}$	≥ 8	3

Calculem primer la probabilitat, sota H_0 , que el model assigna a cada classe. En aquest cas, com que tenim una distribució Poisson,

$$p_j = P\{X \in I_j | X \sim \text{Poisson}(5)\} = P(X = j | X \sim \text{Poisson}(5)) = e^{-5} \frac{5^j}{j!}$$

i, per l'última classe,

$$p_8 = P\{X \in I_8 | X \sim \text{Poisson}(5)\} = P(X \geq 8 | X \sim \text{Poisson}(5)) = 1 - \sum_{i=3}^7 e^{-5} \frac{5^i}{i!}$$

Denotem per $E_j = np_j$ la freqüència esperada en la mostra de la classe j d'acord amb el model Poisson(5) i ja podem construir la taula que ens servirà per calcular la mesura de discrepància

classe I_j	j	O_j	$p_j = P(X = j)$	$E_j = nP(X = j)$
$I_3 = \{X = 3\}$	3	3	0.14	8.54
$I_4 = \{X = 4\}$	4	9	0.17	10.37
$I_5 = \{X = 5\}$	5	18	0.17	10.37
$I_6 = \{X = 6\}$	6	23	0.15	9.15
$I_7 = \{X = 7\}$	7	5	0.11	6.71
$I_8 = \{X \geq 8\}$	≥ 8	3	0.26	15.86

Ara ja podem calcular la discrepància entre les freqüències observades i les previstes pel model:

$$X^2 = \sum_{j=3}^8 \frac{(O_j - E_j)^2}{E_j},$$

que segueix una distribució χ^2 amb tants graus de llibertat ν com classes que tenim menys 1. En el nostre cas, $\nu = 6 - 1 = 5$. Per tant, $X^2 \sim \chi_5^2$.

Calculem la discrepància

$$X^2 = \frac{(3 - 8.54)^2}{8.54} + \frac{(4 - 10.37)^2}{10.37} + \frac{(5 - 10.37)^2}{10.37} + \frac{(6 - 9.15)^2}{9.15} + \frac{(7 - 6.71)^2}{6.71} + \frac{(8 - 15.86)^2}{15.86} = 41.22.$$

Amb un nivell de significació de $\alpha = 0.05$, la regió de rebuig és

$$R = \{X | X^2 \geq \chi_{\nu}(0.05)\} = \{X | X^2 \geq \chi_5^2(0.05)\},$$

i, calculant el valor de $\chi_5^2 = 11.07$, tenim

$$R = \{X | X^2 \geq 11.07\}.$$

Com que hem observat $X^2 = 41.22$, rebutgem H_0 i no podem acceptar la distribució Poisson(5) com a model pels clients que van a les extraccions de sang.

4.

5. La prova d'hipòtesi en aquest cas és

H_0 : Estudiar i aprovar no estan relacionats (és a dir, A i B independents)

H_1 : Estudiar i aprovar estan relacionats (A i B no independents)

Utilitzem la notació vista en les transparències i escrivim la taula de contingència com

		B		
		No	Sí	Total
A	Cada dia	2	8	$10 = O_{1+}$
	Cada setmana	10	40	$50 = O_{2+}$
	Abans de l'examen	28	12	$40 = O_{3+}$
	Total	$40 = O_{+1}$	$60 = O_{+2}$	100

Construïm la taula de valors esperats sota H_0 sabent que $E_{ij} = \frac{O_{i+}O_{+j}}{100}$.

		B		
		Sí	No	Total
A	Cada dia	$\frac{10 \cdot 40}{100} = 4$	$\frac{10 \cdot 60}{100} = 6$	10
	Cada setmana	$\frac{50 \cdot 40}{100} = 20$	$\frac{50 \cdot 60}{100} = 30$	50
	Abans de l'examen	$\frac{40 \cdot 40}{100} = 16$	$\frac{40 \cdot 60}{100} = 24$	40
	Total	60	40	100

Tenint en compte que en aquest cas, com que R són els nivells de A, tenim $R = 3$ i $C = 2$ (nivells de B), la discrepància és

$$X^2 = \sum_{i=1}^3 \sum_{j=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

que segueix una distribució $\chi_{(3-1)(2-1)}^2(0.05) = \chi_2^2(0.05)$. Per tant, la regió de rebuig és

$$R = \{X | X^2 > \chi_2^2(0.05)\} = \{X | X^2 > 5.99\}.$$

Calculem la discrepància en el nostre cas

$$X^2 = \frac{(2-4)^2}{4} + \frac{(10-20)^2}{20} + \frac{(28-16)^2}{16} + \frac{(8-6)^2}{6} + \frac{(40-30)^2}{30} + \frac{(12-24)^2}{24} = 25$$

Per tant, com que $25 > 5.99$, rebutgem H_0 i podem afirmar que estudiar i aprovar estan relacionats.

6.

7. Utilitzem la notació vista en les transparències i escrivim la taula de contingències com

	Creients	Ateus	Agnòstics	Total
Ciències	10	45	45	$100 = O_{1+}$
Lletres	15	40	45	$100 = O_{2+}$
Total	$25 = O_{+1}$	$85 = O_{+2}$	$90 = O_{+3}$	200

Anàlogament a les transparències de classe, construïm la taula de valors esperats sota H_0 sabent que $E_{ij} = \frac{O_{i+}O_{+j}}{120}$

	Creients	Ateus	Agnòstics
Ciències	$\frac{100 \cdot 25}{200} = 12.5$	$\frac{100 \cdot 85}{200} = 42.5$	$\frac{100 \cdot 90}{200} = 45$
Lletres	$\frac{100 \cdot 25}{200} = 12.5$	$\frac{100 \cdot 85}{200} = 42.5$	$\frac{100 \cdot 90}{200} = 45$

Com que tenim 2 poblacions ($R = 2$) i 3 categories ($C = 3$), la discrepància és

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

que segueix una distribució $\chi_{(2-1)(3-1)}^2(0.05) = \chi_2^2(0.05)$. Per tant, la regió de rebuig és

$$R = \{X|X^2 > \chi_2^2(0.05)\} = \{X|X^2 > 5.99\}.$$

Calculem la discrepància en el nostre cas

$$X^2 = \frac{(10 - 12.5)^2}{12.5} + \frac{(15 - 12.5)^2}{12.5} + \frac{(45 - 42.5)^2}{42.5} + \frac{(40 - 42.5)^2}{42.5} + \frac{(45 - 45)^2}{45} + \frac{(45 - 45)^2}{45} = 1.29$$

Per tant, com que $1.29 < 5.99$, no tenim evidència per rebutjar H_0 , és a dir, no podem dir que les creences religioses no són homogènies entre els estudiants de ciències i els de lletres.

8.

9.

10.