Non-parametric tests
Solutions

Statistics II

Szabolcs Sebestyén

szsebe@fcee.ucp.pt

Faculdade de Ciências Económicas e Empresariais
Universidade Católica Portuguesa

Exercise 7.1

(Exame final Junho 2007) Tem-se observado o número de defeitos produzidos por uma máquina ao longo de 100 semanas. Os resultados estão na seguinte tabela:

Número de defeitos	0	1	2	3	4
Semanas	73	18	6	2	1

Teste a hipótese nula ($\alpha = 5\%$) do número de defeitos produzidos pela máquina numa semana seguir uma distribuição de Poisson com média λ.

Exercise 7.1 Solution

λ is unknown $\Longrightarrow$ we need to estimate it
The MLE of λ is $\hat{\lambda} = \overline{X}$

$$\overline{X} = \frac{0 \cdot 73 + 1 \cdot 18 + 2 \cdot 6 + 3 \cdot 2 + 4 \cdot 1}{100} = 0.4$$

$$\begin{aligned}
\mathbb{P}(X = 0) &= e^{-0.4}\frac{0.4^0}{0!} = 0.6703 \Longrightarrow \widehat{E}_0 = 100 \cdot 0.6703 = 67.03 \\
\mathbb{P}(X = 1) &= e^{-0.4}\frac{0.4^1}{1!} = 0.2681 \Longrightarrow \widehat{E}_1 = 100 \cdot 0.2681 = 26.81
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}(X \geq 2) &= 1 - 0.6703 - 0.2681 = 0.0616 \\
\Longrightarrow \hat{E}_{\geq 2} &= 100 \cdot 0.0616 = 6.16 > 5
\end{aligned}$$

Exercise 7.1 Solution

Número de defeitos	0	1	≥ 2
O_i	73	18	9
$\widehat{E}_i$	67.03	26.81	6.16

The test statistic is

$$\chi^2 = \sum_{i=1}^{K} \frac{\left(O_i - \widehat{E}_i\right)^2}{\widehat{E}_i} = 4.736$$

The critical value is $\chi^2_{3-1-1;0.05} = 3.84 < 4.736 \implies$ we can reject H_0

Exercise 7.2

(2° Teste Junho 2007) Um professor afirma que os seus alunos estudam pouco e que se realizasse um teste com 5 perguntas de verdadeiro/falso, o número de respostas certas dadas por um aluno seria uma variável aleatória com distribuição binomial $b\,(5, 0.5)$. Para comprovar a sua afirmação, o professor faz um exame com estas características a 240 alunos, exigindo-lhes que respondam a todas as perguntas. Os resultados do exame estão na tabela seguinte:

Respostas acertadas	0	1	2	3	4	5
Número de alunos	5	10	78	111	28	8

Teste a hipótese nula da afirmação do professor estar certa. Acha que os resultados obtidos são melhores ou piores do que se esperaria se os alunos respondassem aleatoriamente?

Exercise 7.2 Solution

$X \sim b\,(5, 0.5)$

$$\begin{aligned}
\mathbb{P}\,(X = 0) &= 0.03125 \implies E_0 = 240 \cdot 0.03125 = 7.5 \\
\mathbb{P}\,(X = 1) &= 0.15625 \implies E_1 = 240 \cdot 0.15625 = 37.5 \\
\mathbb{P}\,(X = 2) &= 0.31250 \implies E_2 = 240 \cdot 0.3125 = 75 \\
\mathbb{P}\,(X = 3) &= 0.31250 \implies E_3 = 240 \cdot 0.3125 = 75 \\
\mathbb{P}\,(X = 4) &= 0.15625 \implies E_4 = 240 \cdot 0.15625 = 37.5 \\
\mathbb{P}\,(X = 5) &= 0.03125 \implies E_5 = 240 \cdot 0.03125 = 7.5
\end{aligned}$$

Exercise 7.2 Solution

Respostas acertadas	0	1	2	3	4	5
O_i	5	10	78	111	28	8
E_i	7.5	37.5	75	75	37.5	7.5

The test statistic is

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = 40.84$$

The critical value is $\chi^2_{6-1;0.05} = 11.07 < 40.84 \Longrightarrow$ we can reject H_0

Exercise 7.3

Num banco regista-se o tempo (em minutos) que leva a atender um cliente. Os resultados obtidos para uma amostra aleatória de 6 clientes são:

$$3.4 \quad 2.1 \quad 7.8 \quad 4.5 \quad 9 \quad 6.4$$

Usando o teste de Kolgomorov-Smirnov, teste a hipótese nula do tempo que leva a atender um cliente ser uma variável aleatória com função densidade de probabilidade

$$f_0(x) = \begin{cases} \frac{1}{5} e^{-x/5} & \text{se } x > 0 \\ 0 & \text{outros } x \end{cases}$$

Exercise 7.3 Solution

The null hypothesis is that X follows an exponential distribution with parameter $1/5$
$\Longrightarrow$ the c.d.f. of the exponential distribution is

$$F_0(x) = \begin{cases} 1 - e^{-x/5} & \text{se } x > 0 \\ 0 & \text{outros } x \end{cases}$$

Exercise 7.3 Solution

The last step is to find the maximum value in the last column

x_k	$F_N(x) = \frac{k}{N}$	$F_0(x)$	$\lvert F_N(x) - F_0(x) \rvert$
2.1	0.17	0.34	0.17
3.4	0.33	0.49	0.16
4.5	0.50	0.59	0.09
6.4	0.67	0.72	0.05
7.8	0.83	0.79	0.04
9	1	0.83	0.17

The critical value at the 5% significance level, and for sample size of 6 is $d_{6;0.05} = 0.52 > 0.17 \Longrightarrow$ the null hypothesis cannot be rejected, that is, the sample observations can be thought of as realisations from an exponential distribution

Exercise 7.4

Quer-se analizar se a variável $X =$ "logaritmo dos rendimentos laborais mensais das famílias duma população" segue uma distribuição normal. Para fazer este estudo, extraiu-se uma amostra aleatória de 12 famílias, e obtiveram-se os seguintes resultados para X:

6.41	6.83	7.68	8.31	7.76	6.87
6.02	8.21	6.43	6.54	7.11	7.02

$$\overline{X} = 7.0992 \quad S = 0.7389$$

Um investigador quer testar a hipótese nula da distribuição de X ser normal com média 7 e variância 1. Para isso elaborou a tabela anexa, na qual compara a função de distribuição empírica das observações (F_N) com a função de distribuição duma variável aleatória $\mathcal{N}(7, 1)$.

Exercise 7.4

x_k	$F_N(x_k)$	$F_0(x_k)$	$\lvert F_N(x_k) - F_0(x_k) \rvert$
6.02	0.0833	0.1635	0.0802
6.41	0.1667	0.2776	0.1109
6.43	0.2500	0.2843	0.0343
6.54	0.3333	0.3228	0.0105
6.83	0.4167	0.4325	0.0158
6.87	0.5000	0.4483	0.0517
7.02	0.5833	0.5080	0.0753
7.11	y_1	y_2	y_3
7.68	0.7500	0.7517	0.0017
7.76	0.8333	0.7764	0.0569
8.21	0.9167	0.8869	0.0298
8.31	1.0000	0.9049	0.0951

Exercise 7.4

(a) Quais são os valores y_1, y_2 e y_3 que faltam na tabela?

The null hypothesis is $H_0 : X \sim \mathcal{N}(7, 1)$

y_1 can be calculated as

$$y_1 = F_N(7.11) = \frac{8}{12} = 0.6667$$

y_2 can be calculated as

$$y_2 = F_0(7.11) = \mathbb{P}\left(Z \leq \frac{7.11 - 7}{1}\right) = 0.5438$$

y_3 is simply the absolute difference of y_1 and y_2

$$y_3 = |F_N(7.11) - F_0(7.11)| = |y_1 - y_2| = 0.1229$$

Exercise 7.4

b. Há evidência suficiente para rejeitar a hipótese nula com um nível de significância de 10%?

x_k	$F_N(x_k)$	$F_0(x_k)$	$\lvert F_N(x_k) - F_0(x_k) \rvert$
6.02	0.0833	0.1635	0.0802
6.41	0.1667	0.2776	0.1109
6.43	0.2500	0.2843	0.0343
6.54	0.3333	0.3228	0.0105
6.83	0.4167	0.4325	0.0158
6.87	0.5000	0.4483	0.0517
7.02	0.5833	0.5080	0.0753
7.11	0.6667	0.5438	0.1229
7.68	0.7500	0.7517	0.0017
7.76	0.8333	0.7764	0.0569
8.21	0.9167	0.8869	0.0298
8.31	1.0000	0.9049	0.0951

Exercise 7.4

The Kolgomorov-Smirnov statistic is

$$D_N = \sup_x \left[\left| F_N(x) - F_0(x) \right| \right] = 0.1229$$

The critical value at 10% with $N = 12$ is $d_{12;0.1} = 0.34 > 0.1229 \implies$ we cannot reject the null hypothesis

Exercise 7.5

Realizou-se um estudo para analizar as possíveis relações entre o nível educativo dos indivíduos (estudos obrigatórios, estudos médios, estudos universitários) e o nível de consumo de tabaco (consumo nulo, consumo baixo, consumo médio, consumo alto). Entre os 232 indivíduos inquiridos com estudos obrigatórios, 78 não consomem tabaco, 29 têm nível de consumo baixo, 70 têm nível de consumo médio e 55 têm nível de consumo alto. Entre os 116 indivíduos inquiridos com nível de estudos médios, 47 não consomem tabaco, 18 têm nível de consumo baixo, 30 têm nível de consumo médio e 21 têm nível de consumo alto. Entre os 52 indivíduos inquiridos com estudos universitários, 35 não consomem tabaco, 5 têm nível de consumo baixo, 6 têm nível de consumo médio e 6 têm nível de consumo alto.

Exercise 7.5 Solution

a. Construa a tabela das frequências absolutas conjuntas e marginais.

X / Y	Nulo	Baixo	Médio	Alto	*Marginal X*
Est. obr.	78	29	70	55	232
Est. Médios	47	18	30	21	116
Est. Univ.	35	5	6	6	52
Marginal Y	160	52	106	82	400

Exercise 7.5 Solution

- b) Construa a tabela das frequências relativas conjuntas e marginais.

$X\,/\,Y$	Nulo	Baixo	Médio	Alto	*Marginal X*
Est. obr.	0.1950	0.0725	0.1750	0.1375	0.5800
Est. Médios	0.1175	0.0450	0.0750	0.0525	0.2900
Est. Univ.	0.0875	0.0125	0.0150	0.0150	0.1300
Marginal Y	0.4000	0.1300	0.2650	0.2050	1

Exercise 7.5 Solution

- c Construa a tabela das frequências relativas do nível de consumo de tabaco, condicionadas em cada um dos três níveis educativos considerados.

Nível de consumo	$f_{i\|\text{Est.obr.}}$
Nulo	0.3362
Baixo	0.1250
Médio	0.3017
Alto	0.2371

Nível de consumo	$f_{i\|\text{Est.med.}}$
Nulo	0.4052
Baixo	0.1552
Médio	0.2586
Alto	0.1810

Nível de consumo	$f_{i\|\text{Est.uni.}}$
Nulo	0.6731
Baixo	0.0962
Médio	0.1154
Alto	0.1154

Exercise 7.5 Solution

d. Há evidência suficiente para rejeitar a hipótese do nível de consumo de tabaco dum individuo se distribuir igualmente entre os três níveis educativos considerados?

X / Y	Nulo	Baixo	Médio	Alto	*Marginal X*
Est. obr.	92.80	30.16	61.48	47.56	232
Est. Médios	46.40	15.08	30.74	23.78	116
Est. Univ.	20.80	6.76	13.78	10.66	52
Marginal Y	160	52	106	82	400

The test statistic is

$$\chi^2 = \frac{(78 - 92.80)^2}{92.80} + \cdots + \frac{(6 - 10.66)^2}{10.66} = 22.25$$

The critical value at the 5% significance level is $\chi^2_{6;0.05} = 12.592 < 22.25$
$\Longrightarrow$ the null hypothesis can be rejected