ensia

The National School of Artificial Intelligence	**2023-2024**	**Inferential Statistics Semestre 2**

Name:.......................... .... **Group:**......................

Midterm Exam
(1h30)

Exercise 1 (6,5points) *The inventory manager at a pharmaceutical laboratory wants to know how many doses of vaccine he needs to keep in stock.*

It therefore tracks sales of this vaccine over the last 100 days, which are assumed to be representative:

Number of doses sold	0	1	2	3	4	5	6
Number of days	13	27	26	18	9	5	2

Can we say that vaccine sales are distributed according to a Poisson distribution? Propose two different tests for $\alpha = 0.05$.

We test the hypothesis

H_0: "The vaccine sales follows Poisson distribution"; 0,5

H_1: "The vaccine sale doesn't follow Poisson distribution"

For that we will use Khi2 test:

So we calculate $\chi_c^2 = \sum_{i=1}^{k} \frac{(n_i - c_i)^2}{c_i}$ 0,25

where $c_i = n\,\mathbb{P}(X=k)$ and $X \sim \mathcal{P}(\lambda)$ then $\mathbb{P}(X=k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$

First we have the following table

k	n_i	$n\,\mathbb{P}(X=k)$
0	13	12,75
1	27	26,26
2	26	27,04
3	18	18,57
4	9	9,56
5	5	3,94
6	2	1,35

1,75

Since 3,94 and 1,35 are less than 5 0,25 we will ignore the two last classes

k	n_i	$n \cdot \mathbb{P}(X=k)$	$(n_i - c_i)^2/c_i$
0	13	12,75	0,0051
1	27	26,26	0,0211
2	26	27,04	0,0402
3	18	18,57	0,0175
4	9	9,56	0,0332
5	7	5,29	0,5506
Total	100		0,6677

$\chi_c^2 = 0{,}6677$ and $ddl = k - 1 - r = 6 - 1 - 1 = 4$ 0,5

$\Rightarrow \chi_{0,05}^2(4) = 9{,}4877$

So $\chi_c^2 < \chi_{0,05}^2(4)$ 0,25 then we accept H_0

The second test we will use is the Kolmogorov test

So we calculate $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$ 0,5

where F is the cumulative function of the Poisson distribution

k	n_i	$F_i^c / = F_n$	F	$F_n - F$
0	13	0,13	0,127	0,003
1	27	0,40	0,390	0,01
2	26	0,66	0,660	0
3	18	0,84	0,846	0,006
4	9	0,93	0,942	0,012
5	5	0,98	0,995	0,015
6	2	1	1,008	0,008

1,75

Then $D_n = 0{,}015$ 0,25

For $n = 100$ and $\alpha = 0{,}05$ we have from the Kolmogorov table $c = 0{,}1340$ 0,25

* So $\frac{c}{\sqrt{n}} = \frac{0{,}1340}{10} = 0{,}0134$.

Then $D_n = 0{,}015 > \frac{c}{\sqrt{n}}$ thus we reject H_0 0,25

Exercise 2 (7 **points**) *Let $n \in \mathbb{N}^*$ and $x_1, \cdots, x_n$ n be observations following the exponential law $\mathcal{E}\left(\frac{1}{\theta}\right)$ with $\theta > 0$:*

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{1}{\theta}x} & \text{if } x \geq 0 \\ 0 & \text{elswhere} \end{cases}$$

The parameter θ is unknown. Let $(X_1, \cdots, X_n)$ be an $n-$sample of X.

1. *Determine a maximum likelihood estimator $\widehat{\theta}_n$ of θ:*

$L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{1}{\theta}x_i} = \frac{1}{\theta^n} \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right\}$

$\Rightarrow \ln L(x_1, \ldots, x_n; \theta) = \ln\left(\frac{1}{\theta^n} \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right\}\right)$

$= -n \ln \theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i$ 1

$\hat{\theta}_n$ is the solution of $\frac{\partial}{\partial \theta} \ln L(x_1, \ldots, x_n; \theta) = 0$

$\frac{\partial}{\partial \theta} \ln L(x_1, \ldots, x_n; \theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = 0 \Rightarrow \theta = \frac{1}{n}\sum_{i=1}^{n} x_i$

where it is a maximum of $\ln L(x_1, \ldots, x_n; \theta)$

Then $\boxed{\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i}$ 1

2. *Is $\widehat{\theta}_n$ an unbiased, consistent and effective (efficace) estimator?*

* We have $\mathbb{E}[X_1] = \frac{1}{1/\theta} = \theta$ and since the r.v $X_1, X_2, \dots, X_n$ are iid

$\Rightarrow \mathbb{E}[\hat{\theta}_n] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\, n\, \mathbb{E}(X_i) = \theta$ 1

then $\hat{\theta}_n$ is an unbiased estimator of θ

* We have also $Var(X_1) = \frac{1}{(1/\theta)^2} = \theta^2$. 1

then $Var(\hat{\theta}_n) = Var(\frac{1}{n}\sum X_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\theta^2}{n}$

Since $\hat{\theta}_n$ is an unbiased estimator of θ and $Var(\hat{\theta}_n) \xrightarrow[n\to\infty]{} 0$ then $\hat{\theta}_n$ is effective. 0,5

* By the weak law of large numbers 1

$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{} \mathbb{E}(X_1) = \theta$ then $\hat{\theta}_n$ is consistent.

3. *Determine* $\lim_{n\longrightarrow\infty} \mathbb{P}\left(\widehat{\theta}_n - 1.96\frac{\widehat{\theta}_n}{\sqrt{n}} \leq \theta \leq \widehat{\theta}_n + 1.96\frac{\widehat{\theta}_n}{\sqrt{n}}\right)$.

$\mathbb{P}\left(\hat{\theta}_n - 1{,}96\frac{\hat{\theta}_n}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + 1{,}96\frac{\hat{\theta}_n}{\sqrt{n}}\right) = \mathbb{P}\left(-1{,}96 \leq \frac{\sqrt{n}}{\hat{\theta}_n}(\theta - \hat{\theta}_n) \leq 1{,}96\right)$

$= \mathbb{P}\left(\left|\frac{\sqrt{n}}{\hat{\theta}_n}(\theta - \hat{\theta}_n)\right| \leq 1{,}96\right)$ 0,5

We have $\frac{\sqrt{n}}{\hat{\theta}_n}(\hat{\theta}_n - \theta) = \frac{\theta}{\hat{\theta}_n}\,\frac{\hat{\theta}_n - \theta}{\theta/\sqrt{n}}$ 0,5 Since $\hat{\theta}_n \to \theta$ (consistent)

and $\theta/\sqrt{n} = \sqrt{Var\,\hat{\theta}_n}$ then by the central limit theorem

$\frac{\hat{\theta}_n - \theta}{\hat{\theta}_n/\sqrt{n}} = Z \sim \mathcal{N}(0,1)$ so $\mathbb{P}(|Z| \leq 1{,}96) = 0{,}95$ 0,5

then $\boxed{\mathbb{P}\left(\hat{\theta}_n - 1{,}96\frac{\hat{\theta}_n}{\sqrt{n}} \leq \theta \leq \hat{\theta} + 1{,}96\frac{\hat{\theta}_n}{\sqrt{n}}\right) = 0{,}95}$

xercise 3 (6,5 points) *Let us consider the following marks obtained by two student in Analysis and atistics*

Analysis' marks X	2	3	5	6	6	7	8	11	12	12	15	18	20
Statistics' marks Y	6	4	9	11	9	8	12	7	10	8	10	17	16

there a link between the Analysis' marks (X) and Statistics' marks (Y), for $\alpha = 0.05$?

1. *If the variables are assumed to be normal.*

0,25

We use in this case the correlation test, so we have the hypothesis $H_0: \rho = 0$ versus $H_1: \rho \neq 0$

Under H_0 the statistics $T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim \mathcal{T}(n-2)$. 0,5

We calculate first the correlation coefficient r.

We have

$\overline{X} = 9,6154$ (0,25) ; $\sigma_X \simeq 5,4$ (0,25)

$\overline{Y} = 9,7692$ (0,25) ; $\sigma_Y \simeq 3,5116$ (0,25) and $Cov(X,Y) \simeq 14,5266$ (0,5)

Then $r = 0,7661$ and $t_c = \frac{0,7661 \cdot \sqrt{11}}{\sqrt{1-0,7661^2}} \simeq \boxed{3,9529}$ 0,5

For $n-2 = 11$ degree of freedom and $\alpha = 0,05$ we have $t_\alpha = 2,2010$ 0,5

then $t_c \notin]-2,2010 ; 2,2010[$ thus we reject H_0.

0,5

The marks X and Y are linked.

2. *If we have no information on the distribution of the two variables.*

Then we use the Spearman test, so

x_i	2	3	5	6	6	7	8	11	12	12	15	18	20	
y_i	6	4	9	11	9	8	12	7	10	8	10	17	16	
x'_i	1	2	3	4,5	4,5	6	7	8	9,5	9,5	11	12	13	0,5
y'_i	2	1	6,5	10	6,5	4,5	11	3	8,5	4,5	8,5	13	12	0,5
$(x'_i - y'_i)^2$	1	1	12,25	30,25	4	2,25	16	25	1	25	6,25	1	1	126 (0,25)

We calculate the rank correlation coefficient of Spearman:

$$r_S = 1 - \frac{6 \sum (x'_i - y'_i)^2}{n(n^2-1)} = 1 - \frac{6 \cdot 126}{13(169-1)}$$

0,5

$\Rightarrow \boxed{r_S \simeq 0,6538}$

Since $n \leq 13$ we use the Spearman table to determine $r_{0,05}$ which gives $r_{0,05} = 0,5602$ 0,5

We have $r_S = 0,6538 > r_{0,05}$ 0,5 then we reject H_0 and the marks X and Y are linked.