Advanced Statistics

Exercice 1: Is there a correlation between consumed oil and cardiovascular problems?

1.1 We will perform a χ^2 test for independance between cardiovascular problem and consumed oil. We are given information for 200 persons.

Let X be a random variable modelling the consumed oil. It takes value in { Olive, Groundnut }

Let Y be a random variable modelling the cardivascular status. It takes value in { HasProblem, NoProblem }

The contingence table is given below : $N_{ij}{}_{1 \le i \le 2, 1 \le j \le 2}$

y \ X	Olive	Groundnut	
HasProblem	20 (given) $= N_{11}$	10 $= N_{21}$	$\rightarrow N_{\cdot 1} = N_{11} + N_{21} = 30$
NoProblem	100 $= N_{12}$	70 (given) $= N_{22}$	$\rightarrow N_{\cdot 2} = N_{12} + N_{22} = 170$
	$N_{1\cdot} = 200 - N_{2\cdot} = 120$	$N_{2\cdot} = 80$ (given)	

1.2. The hypothesis of our χ^2 test are below:

$H_0 = \{$ X and Y are independant $\}$

$H_1 = \{$ X and Y are not independant $\}$

Under H_0, the statistic $\xi_n = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(N_{ij} - \frac{N_{i\cdot} N_{\cdot j}}{n}\right)^2}{\frac{N_{i\cdot} N_{\cdot j}}{n}}$

converges in distribution towards a $\chi^2(1)$

The test is defined by $W_n = \{ \xi_{200} > q_1(1-\alpha) \}$

with $q_1(1-\alpha)$ the quantile of order $1-\alpha$ of $\chi^2(1)$

Let compute ξ_{200}:

$$\xi_{200} = \frac{\left[20 - \frac{120\times 30}{200}\right]^2}{\frac{120\times 30}{200}} + \frac{\left[100 - \frac{120\times 170}{200}\right]^2}{\frac{120\times 170}{200}} + \frac{\left[10 - \frac{80\times 30}{200}\right]^2}{\frac{80\times 30}{200}} + \frac{\left[70 - \frac{80\times 170}{200}\right]^2}{\frac{80\times 170}{200}}$$

$$= 0.65$$

Let know write the critical regions:

Case $\alpha = 5\%$, we have $P(\chi^2_1 \leq 3.841) = 95\%$ | $W^1 = \{\xi_n > 3.841\}$

ξ_{200} is par below 3.841 $\Rightarrow$ We do not reject H_0.

Case $\alpha = 0.001$, we have $P(\chi^2_1 \leq 10.828) = 99{,}9\%$ | $W^2 = \{\xi_n > 10.828\}$

same conclusion in the case $\alpha = 5\%$

Let know test with $\alpha = 20\%$.

Case $\alpha = 20\%$ $P(\chi^2_1 \leq 1{,}64) = 80\%$ | $W^3 = \{\xi_n > 1.64\}$

Even with $\alpha = 20\%$, we cannot reject the hypothesis of independance between consumed oil and cardiovascular problems. So in conclusion, we can say that these two variables are indeed independant.

To ensure that conclusion we can compute the p-value of the test.

$$\text{P-value} = P(\chi^2_1 \geq 0.65) = 1 - P(\chi^2_1 \leq 0.65)$$

We observed on the χ^2 quantile table that

$$P(\chi^2_1 < 0.7) = 0.6 \quad \text{and} \quad P(\chi^2_1 < 0.455) = 0.5$$

$\Rightarrow$ $\frac{0.6 - 0.5}{0.7 - 0.455} = 0.40$ added probability for 0.01

$\Rightarrow$ $P(\chi^2_1 \leq 0.65) = P(\chi^2_1 < 0.7) - 0.05 \times 0.40 = 0.6 - 0.02 = 0.58$

Thus P-value $= 1 - 0.58 = 0.42$

Conclusion: The p-value of the χ^2 test is very large compared to all $\alpha(s)$, then we cannot reject H_0 even when α is 40%.

Exercice 2:

A paracetamol concentration greater than 150 mg per kilogram is considered to be dangerous. We want to build a test that says if a patient has a risk from 4 experiments that follow the same Gaussian distribution.

a) Measure of blood concentration $X \sim N(\mu, \sigma^2)$

σ is 5.

We will perform a one sided test hypothesis.

$H_0 = \{$ the parameter μ of the random variable that models the blood concentration is greater than 150 mg/kg : $\mu > \mu_0 = 150 \}$

The alternative hypothesis is that $\{$ the blood is lower than $\mu_0 = 150$, $\mu < 150 \} = H_1$.

Let write the likelihood function:

$$L(x, \mu) = \prod_{i=1}^{n} f_\mu(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

for $\mu'' > \mu'$.

$$\sum_{i=1}^{n}\left(\frac{x_i - \mu''}{\sigma}\right)^2 < \sum_{i=1}^{n}\left(\frac{x_i - \mu'}{\sigma}\right)^2$$

$$\Rightarrow \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu''}{\sigma}\right)^2\right] > \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu'}{\sigma}\right)^2\right]$$

$$\Rightarrow L(x, \mu'') > L(x, \mu')$$

Let now compute the likelihood ratio:

$$LR = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu'')^2 - (x_i - \mu')^2\right] = \exp\left[\frac{-1}{\sigma 2}\sum_{i=1}^{n}(x_i - \mu'' - x_i + \mu')(x_i - \mu'' + x_i - \mu')\right]$$

$$= \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(\mu' - \mu'')(2x_i - \mu'' - \mu')\right]$$

$$= \exp\left[\frac{(\mu'' - \mu')}{\sigma^2}\sum x_i + n(\mu'^2 - \mu''^2)\right]$$

$\Rightarrow$ LR is an increasing function of $\sum_{i=1}^{n} x_i = S(x)$ because the exponential is increasing.

The Lehman theorem gives us a UMP test at level α for testing $H_0 = \{\mu \geq \mu_0\}$ versus $\{\mu < \mu_0\}$ because the likelihood ratio is increasing. The test is defined as follow:

$$\begin{cases} \varphi(x) = 1 & \text{if } S(x) = \sum_{i=1}^{n} x_i < c \\ \varphi(x) = \gamma & \text{if } S(x) = \sum_{i=1}^{n} x_i = c \\ \varphi(x) = 0 & \text{if } S(x) = \sum_{i=1}^{n} x_i > c \end{cases}$$

The critical region is $W = \{x \in \mathbb{R}^n \mid S(x) < c\}$

We need to compute c.

To do so, let find the distribution of $S(x)$

$$\bar{X} \sim N(\mu, \sigma^2)$$

$$\sum_{i=1}^{n} x_i \sim N(\mu n, n^2\sigma^2)$$

In our case, we have

$$X \sim N(150, 5)$$

$$S(x) = \sum_{i=1}^{4} x_i \sim N(600, 4 \times 5^2)$$

Under μ_0: we can compute c

$$P(S(x) < c) = \alpha$$

Since the table of the quantile of the normal distribution exist in $\sim N(0,1)$.

$$\Rightarrow P\left(\frac{S(x) - 4\mu_0}{2\sigma} < \frac{c - 4\mu_0}{2\sigma}\right) = \alpha$$

If $\alpha = 0.05$

$$\Rightarrow -1,645 = \frac{c - 4\mu_0}{2\sigma}$$

$$\Rightarrow c = 583.55$$

$\Rightarrow W = \{S(x) < 583.55\}$ is the critical region at level $\alpha = 0.05$

b- We are given the results of the experiments of a patient

$X_1 = 141$; $X_2 = 150$; $X_3 = 144$; $X_4 = 142$

$S(X) = \sum_{i=1}^{4} X_i = 577$

The test return 0 i.e. the hypothesis H_0 is rejected. So the patient is not in danger.

Let compute the p-value of the test

$$p\text{-value} = P(S < 577) = P\left(\frac{S - 4\mu_0}{2\sigma} < \frac{577 - 600}{10}\right)$$

$$= P\left(\frac{S - 4\mu_0}{2\sigma} < -2.3\right)$$

Since the normal distribution is symmetric

$$p\text{-value} = 1 - P\left(\frac{S - 4\mu_0}{2\sigma} < 2.3\right) = 1 - 0.9893$$

$$\boxed{p\text{-value} = 0.0107}$$

The p-value is below 5% $\Rightarrow$ we reject the null hypothesis

Conclusion: The patient is not ~~is~~ not in danger.

Problem :

Let us consider the following PDF

$$f_\theta(x) = \theta^2 x e^{-\theta x} \mathbb{1}_{[0,+\infty[}(x)$$

One observe a n-sample $(X_1 \dots X_n)$ iid

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Q.1 : this is a gamma distribution with $p = 2$

$$f_\theta(x) = \frac{\theta^2}{\Gamma(2)} x^{2-1} e^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x)$$

$$\Gamma(2) = (2-1)! = 1!$$

$$\Rightarrow f_\theta(x) = \theta^2 x e^{-\theta x} \mathbb{1}_{[0,+\infty[}(x)$$

Q 2: Is the model belongs to the exponential family

the likelihood function can be written as

$$L(x,\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta^2 x_i e^{-\theta x_i} \mathbb{1}_{[0,+\infty[}(x_i)$$

$$= \theta^{2n} \mathbb{1}_{[0,+\infty[}(\min(x_i)\ i \in [1,n]) \left(\prod_{i=1}^{n} x_i\right) e^{-\sum_{i=1}^{n} x_i \theta}$$

can be written as

$$L(x,\theta) = h(x)\, \phi(\theta) \exp\left(\sum_{i=1}^{r} Q_i(\theta)\, S_i(\theta)\right)$$

with

$$h(x) = \mathbb{1}_{\mathbb{R}_+}(\min(x_i)\ i \in [1,n]) \prod_{i=1}^{n} x_i$$

$$\phi(\theta) = \theta^{2n}$$

$$Q(\theta) = -\theta$$

$$s(x) = \sum_{i=1}^{n} x_i \quad \text{is the canonical statistic.}$$

$$r = 1$$

Thanks to the Factorisation criterion the likelihood function is

$$L(x,\theta) = \psi(s(x),\theta)\, h(x)$$

with $\psi(S(x),\theta) = \phi(\theta)\exp\left(Q(\theta)S(x)\right)$

and $h(x) = \left(\prod_{i=1}^{n} x_i\right) \mathbb{1}_{\mathbb{R}^+}\left(\min(x_i)_{i\in[1,n]}\right)$

⇒ thus the Factorisation criterion theorem says that the statistic is sufficient.

Q3. Is S complete?

$Q(\Theta) = \mathbb{R}^-$ as $\theta \in \mathbb{R}_*^+$

⇒ since $\mathbb{R}^-$ is a non-empty set of $\mathbb{R}$, the canonical statistic is complete thanks to the theorem relative to exponential family completeness (slide 10, Part B).

Q4: The likelihood function of the model can be written as

$$L(x,\theta) = K(\theta)h(x)\exp\left(\sum_{j=1}^{r}\theta_j S_j(x)\right)$$

with $K(\theta) = \theta^{2n}$

$h(x)$ unchanged

$\theta_1 = -\theta$

$S(x) = \sum_{i=1}^{n} x_i$

$r = 1$

⇒ $Q(\theta) = \theta_1 \subset \mathbb{R}$ is a non-empty set

⇒ the theorem of regularity of exponential model says that our model is regular.

Q5: And Also the score function is square integrable

⇒ The fisher information for $n = 1$

$$I(\theta) = -E_\theta\left[\frac{\partial^2 \ln L(x,\theta)}{\partial^2\theta}\right]$$

$$= -E_\theta\left[\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta}\left(\log\theta^2 + \log(x) - \theta x\right)\right] \qquad x > 0$$

$$= -E_\theta\left[\frac{\partial^2}{\partial\theta}\left(2\log\theta - \theta x + \log x\right)\right]$$

$$= -E_\theta\left[\frac{\partial}{\partial\theta}\left(\frac{2}{\theta} - x\right)\right] = E_\theta\left[\frac{2}{\theta^2}\right]$$

$$= \int_0^{+\infty} \theta^2 x e^{-\theta x} \times \frac{2}{\theta^2}\, dx$$

$$= 2\int_0^{+\infty} x e^{-\theta x} dx = 2\left[-\frac{1}{\theta}e^{-\theta x}\right]_0^{+\infty} + 2\int_0^{+\infty} \frac{1}{\theta} e^{-\theta x} dx$$

$$I_1(\theta) = \frac{1}{\theta}\left[-\frac{1}{\theta}e^{-\theta x}\right]_0^{+\infty} = \frac{2}{\theta^2}$$

Q.6.

$$E_\theta[X_1] = \int_0^{+\infty} x^2 \theta^2 e^{-\theta x} dx$$

$$= \left[-\theta x^2 e^{-\theta x}\right]_0^{+\infty} + \int_0^{+\infty} 2x\theta e^{-\theta x} dx$$

$$= \int_0^{+\infty} 2x\theta e^{-\theta x} dx = 2\left[-x e^{-\theta x}\right]_0^{+\infty} + \int_0^{+\infty} 2e^{-\theta x} dx$$

$$= \left[-\frac{2}{\theta} e^{-\theta x}\right]_0^{+\infty} = \frac{2}{\theta}$$

$$E[X_1] = \frac{1}{\theta}$$

Thanks to the SLLN since X_1 is integrable

$$\bar{X}_n \xrightarrow{a.s} E[X_1] = \frac{1}{\theta}$$

$$\rightsquigarrow \boxed{\tilde{\theta}_n = \frac{1}{\bar{X}_n}} \text{ an estimator for } \theta.$$

ID et unbiaised ?

$$E\left[\frac{1}{x}\right] = \int \frac{x^p \theta^p}{\Gamma(p)} e^{-\theta x} dx \quad \text{with } y = \theta x$$

$$= \int \frac{y^p e^{-y}}{\Gamma(p)\theta} dy = \frac{1}{\theta^2}\frac{\Gamma(p+1)}{\Gamma(p)} = \frac{1}{\theta^2}\frac{p!}{(p-1)!}$$

$\rightsquigarrow$ $E[\tilde{\theta}_n] = \frac{1}{\theta^2} \times 2$ in this case because $p = 1$

$\rightsquigarrow$ estimator is biaised

$\bar{X}_n \sim \Gamma(2n, n\theta)$ since the X_i are iid and

$\Gamma(p_1, \theta) + \Gamma(p_2, \theta) = \Gamma(p_1 + p_2, \theta)$

and $a\,\Gamma(p, \theta) \sim \Gamma(p, \frac{\theta}{a})$

$$E[\hat{\theta}_n] = E\left[\frac{1}{\bar{X}_n}\right]$$

let say $y = \bar{X}_n$

$$E\left[\frac{1}{y}\right] = \int \frac{1}{y} f_{2n}^{\theta n}(y)\, dy$$

with $f_{2n}^{\theta n}(y)$ the density function of a gamma $(2n, \theta n)$

We can write $f_{2n}^{\theta n}(t) = \frac{(n\theta)^{2n}}{\Gamma(2n)} t^{2n-1} e^{-\theta n t} \mathbb{1}_{\mathbb{R}^+}(t)$

$$= \frac{\theta n}{2n-1} \frac{(\theta n)^{2n-1}}{\Gamma(2n-1)} t \times t^{2n-2} e^{-\theta n t} \mathbb{1}_{\mathbb{R}^+}(t)$$

$$\Rightarrow \frac{f_{2n}^{n\theta}(t)}{t} = \frac{\theta n}{2n-1} f_{2n-1}^{n\theta}(t)$$

$$\Rightarrow E\left[\frac{1}{y}\right] = \frac{\theta n}{2n-1} E\left[f_{2n-1}^{n\theta}(y)\right]$$

$$= \frac{\theta n}{2n-1}$$ because $E\left[f_{2n-1}^{n\theta}(t)\right] = 1$

$\Rightarrow$ the estimator is biaised.

Q-7 $\bar{\theta}_n = \frac{2n-1}{n} \frac{1}{\bar{X}_n}$.

Let show that $\bar{\theta}_n$ is an unbiaised estimator.

$$E[\bar{\theta}_n] = E\left[\frac{2n-1}{n} \frac{1}{\bar{X}_n}\right] = \frac{2n-1}{n} E\left[\frac{1}{\bar{X}_n}\right]$$

using question 6 $E[\bar{\theta}_n] = \frac{2n-1}{n} \times \frac{\theta n}{2n-1} = \theta$

$\Rightarrow$ $\bar{\theta}_n$ is unbiaised.

8. The statistic S is complete and sufficient (question 2)

$\bar{\theta}_n$ is also unbiaised.

⇒ Thanks to the Lehman Scheffé theorem $\bar{\theta}_n$ is optimal.

$$E\left[\bar{\theta}_n \mid \sum_1^n x_i = \sum_1^n x_i\right] = \frac{2n-1}{n}\frac{n}{\sum x_i}$$

The estimator is its own Rao Blackwellization

⇒ It is optimal.

Q9. Likelihood function

$$L(x,\theta) = \theta^{2n} \prod_{[0,+\infty[} (\min(x_i)_{i\in[1,n]}) \left(\prod_i^n x_i\right) e^{-\sum x_i \theta}$$

Let find the MLE

$$\frac{\partial}{\partial\theta} \log L(x,\theta) = \frac{\partial}{\partial\theta}\left[2n\log\theta + \log\prod_i^n x_i - \theta\sum x_i\right] = \frac{2n}{\theta} - \sum_{i=1}^n x_i$$

$$\boxed{\hat{\theta}_n = \frac{2n}{\sum_{i=1}^n x_i}}$$

since $\frac{\partial^2 l(x,\theta)}{\partial^2\theta} = -\frac{2n}{\theta^2} < 0$

$\hat{\theta}_n$ is the maximum likelihood estimator

Q10.

$\Theta(\theta) = \mathbb{R}^-$ is a non empty set of $\mathbb{R}$

the fisher information is inversible $\forall\ \theta$

⇒ Thanks to the theorem (part B - slide 32) the MLE is unique, strongly consistent and asymptotically efficient.

Q11.

$$\bar{\theta}_n = \frac{2n-1}{n}\frac{1}{\bar{X}_n} = \frac{2}{n\bar{X}_n} - \frac{1}{n\bar{X}_n} = \hat{\theta}_n - \frac{\hat{\theta}_n}{2n}$$

$\hat{\theta}_n$ is asymptotically efficient estimator of θ

⇒ $\hat{\theta}_n\left[1-\frac{1}{2n}\right]$ is an efficient estimator of $\theta\left[1-\frac{1}{2n}\right]$

$\bar{\theta}_n \xrightarrow[n\to\infty]{} \theta$ then $\bar{\theta}$ is asymptotically efficient.

Q12 :

$H_0 = \{\theta = \theta_0\}$ $\quad$ $H_1 = \{\theta > \theta_0\}$

1°) $\bar{X}_n$ follows a gamma distribution

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$X_i \sim \Gamma(\alpha, \theta)$

$\sum_{i=1}^{n} X_i \sim \Gamma(2n, \theta)$ because les X_i are iid

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\Rightarrow \quad \bar{X}_n \sim \Gamma(2n, \theta/_{1/n}) = \Gamma(2n, n\theta)$$

2. The test deriv from the Lehman theorem in case of a monotone likelihood ratio function is also the UMP for testing $H_0 = \{\theta = \theta_0\}$ versus $H_1 = \{\theta > \theta_0\}$

$$\begin{cases} \varphi(x) = 1 & \sum_{i}^{n} x_i > c \\ \varphi(x) = \gamma & \sum_{i}^{n} x_i = c \\ \varphi(x) = 0 & \sum x_i < c \end{cases}$$

c, γ are obtained with $E_{\theta_0}[\varphi] = \alpha$

3. Wald test $H_0 = \{\theta = \theta_0\}$ $\quad$ $H_1 = \{\theta \neq \theta_0\}$

The wald statistic is $W = \dfrac{[\hat{\theta}_n - \theta_0]^2}{1/I_n(\hat{\theta})} = I_n(\hat{\theta}_n)[\hat{\theta}_n - \theta_0]^2$

$\hat{\theta}_n$: the MLE

$I_n(\hat{\theta})$: the fisher information evaluated at $\hat{\theta}_n$

$$W \sim \chi^2(1)$$

With the critical region $\{W > q_1(1-\alpha)\}$

$q_1(1-\alpha)$ is the quantile of order $1-\alpha$ of $\chi^2(1)$

Q 13.

1. Voir code

2. Voir code

3. Voir code

4. The MSE of the estimators $\tilde{\theta}_n$ is high comparatively to the two others $\bar{\theta}_n$ and $\hat{\theta}_n$ which are decreasing as the size of the sample increases, and tends to zero asymptotically. The reason why $\tilde{\theta}_n$ is so high is because it is a biaised estimator.