MSc. Data Sciences & Business Analytics
Prof. F. Pascal
Advanced statistical methods

CentraleSupélec / ESSEC
December, 2017
This text has 3 pages

Exam of statistics

Possible softwares for simulations : R, Python, Matlab.

Exercise 1. Ones observes 200 persons that eat either groundnut oil or olive oil. Among them :

- 80 have eaten groundnut oil
- 20 have eaten olive oil and then, had cardiovascular problems
- 70 have eaten groundnut oil and had no problem.

One wants to test the independence between the consumed oil and cardiovascular problems.

a) Write a contingence table, thanks to previous values.

b) Derive the χ^2 test associated to this problem.
At a level $\alpha = 5\%$, what would be your conclusion ? Now what is the conclusion if $\alpha = 10^{-3}$ or $\alpha = 20\%$? Comments on previous results.

Exercise 2. A paracetamol concentration greater than 150 mg per kilogram is considered to be dangerous ; e.g. the limit for a person with a weight of 75 kg is 11.25g. Measures of paracetamol in the blood are modelled by a random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. The standard deviation associated to the measurement procedure is supposed to be known and $\sigma = 5$. For security purposes, 4 tests are done and are supposed to be independent realisations of the same Gaussian distribution.

a) Write the hypotheses of the test for testing if a patient has a risk from the 4 experiments. Write the the critical region for the test at level $\alpha = 5\%$ (you are a wise doctor).

b) For a given patient, the A experiments have given the following paracetamol concentrations : 141, 150, 144, 142. Compute the p-value of the previous test. Is this patient in danger ?

Problem 1. Let us consider the following PDF :

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

where $\theta > 0$ is the parameter to estimate.

One observes a n-sample $(X_1, \ldots, X_n)$ i.i.d. with PDF f_θ and we will denote $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Q1. What is the distribution of X_i ?

Q2. Show that the model belongs to the exponential family and exhibits a sufficient statistic S.

Q3. Prove that S is complete.

Q4. The model is regular. Why ?

Q5. Compute the Fisher information $I_1(\theta)$ for $n = 1$.

Q6. Compute $\mathrm{E}_\theta[X_1]$. Deduce an estimator $\tilde{\theta}_n$ thanks to the method of moment. Is this estimator unbiased ?

Q7. Show that $\bar{\theta}_n = \frac{(2n-1)}{n} \frac{1}{\bar{X}_n}$ is an unbiased estimator of θ.

Q8. Is $\bar{\theta}_n$ optimal in the class of unbiased estimators ? Is-it efficient ?

Q9. Write the likelihood function and find the Maximum Likelihood estimator $\hat{\theta}_n$.

Q10. Show that $\hat{\theta}_n$ is asymptotically efficient.

Q11. By writing $\bar{\theta}_n$ with $\hat{\theta}_n$, show that $\bar{\theta}_n$ is asymptotically efficient.

Q12. Let us now consider the test with the null hypothesis $H_0 : \{\theta = \theta_0\}$ versus the alternative hypothesis $H_1 : \{\theta > \theta_0\}$.

1. Show that $\bar{X}_n$ follows a Gamma distribution and give the parameters of this distribution.
2. Propose an UMP test at level α for testing H_0 versus H_1 (be careful at the sens of the inequality).
3. Derive the Wald test for $H_0 : \{\theta = \theta_0\}$ versus $\{H_1 : \theta \neq \theta_0\}$.

Q13. Simulations and numerical applications Choose a value for θ.

1. Propose a way of simulating a n-sample $(X_1, \ldots, X_n)$ i.i.d. with PDF $f_\theta(.)$.
2. **Estimation :** Given this sample, compute the three estimators $\bar{\theta}_n$, $\tilde{\theta}_n$ and $\hat{\theta}_n$.
3. **Monte-Carlo simulations :** Evaluate the numerical performance of previous estimators by plotting their MSEs as well as te CRB (on the same graph). *Of course, it should be done for different values of n and for an appropriate number of Monte Carlo trials.*

4. Comment previous plot with regards to the theoretical results.
5. **Hypothesis testing :** Fix a value for θ_0, α and n. Simulate $\bar{X}_n$. What is the conclusion of the test ? Evaluate the performance of this test with Monte Carlo simulations.
6. Keep previous values for the parameters. What is the conclusion of the Wald test ? Evaluate the performance of this test with Monte Carlo simulations.
7. Find a scenario that highlights the better performance of the Neyman-Pearson approach, compared to the asymptotic approach (e.g., Wald test).