MAP531: Homework

You are asked to provided answers to all these exercises as both Rmd and pdf files. The two files should be uploaded on Moodle on the 11th of December (23h59 Paris time).

This homework should be done by groups of 2. Only one submission per group on moodle, with both names indicated in the file.

This homework is composed of 2 independent problems.

One question is a bit more technical: its is marked by a * and is optional.

Part 1: Estimating parameters of a Poisson distribution to model the number of goals scored in football

We recall that the Poisson distribution with parameter $\theta > 0$ has a pdf given by $(p(\theta, k), k \in \mathbb{N})$ w.r.t the counting measure on $\mathbb{N}$:

$$p(\theta, k) = \exp(-\theta)\frac{\theta^k}{k!}\,.$$

Question 1

Is it a discrete or continuous distribution? Can you give 3 examples of phenomenons that could be modeled by such a distribution in statistics?

Question 2

Compute the mean and the variance of this distribution as a function of λ.

Remark: that if X_1 and X_2 are two independent random variables following a Poisson distribution with respective parameters $\lambda_1 > 0$ and $\lambda_2 > 0$, then $X_1 + X_2$ has a Poisson distribution of parameter $\lambda_1 + \lambda_2$. You do not need to prove this result.

We are provided with n independent observations of a Poisson random variable of parameter $\theta \in \Theta = \mathbb{R}_+^*$.

Question 3

- What are our observations? What distribution do they follow?
- Write the corresponding statistical model.
- What parameter are we trying to estimate?

Question 4

- What is the likelihood function?
- Compute the Maximum Likelihood Estimator $\hat{\theta}_{ML}$.

Question 5

Prove that $\sqrt{n}(\hat{\theta}_{ML} - \theta)$ converges in distribution as $n \to \infty$.

Question 6

- Prove that $\sqrt{n}\frac{\hat{\theta}_{ML}-\theta}{\sqrt{\theta_{ML}}}$ converges in distribution as $n \to \infty$.
- On R, verify that the distribution of the random variable $\sqrt{n}\frac{\hat{\theta}_{ML}-\theta}{\sqrt{\theta_{ML}}}$ is what you found theoretically, through a histogram and a QQ-plot (compute $Nattempts = 1000$ times the random variable $\sqrt{n}_{sample}\frac{\hat{\theta}_{ML}-\theta}{\sqrt{\theta_{ML}}}$ from a sample of size n_{sample} of simulated Poisson data, with $\theta = 3$, like in PC2).

Question 7

For $\alpha \in (0, 1)$, give an asymptotic confidence interval of level α, that is an interval $[a_n(\alpha, (X_i)_{i\in\{1\dots,n\}}); b_n(\alpha, (X_i)_{i\in\{1\dots,n\}})]$, such that:

$$\lim_{n\to+\infty} \mathbb{P}\left(\theta \in \left[a_n\left(\alpha, (X_i)_{i\in\{1\dots,n\}}\right); b_n\left(\alpha, (X_i)_{i\in\{1\dots,n\}}\right)\right]\right) \geq 1 - \alpha.$$

Question 8

- Propose two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ based on the first and second moments of a Poisson distribution.
- What can you say about $\hat{\theta}_1$?

Question 9

Compute the Bias, the Variance, and the quadratic risk of $\hat{\theta}_{ML}$.

Question 10

Let $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, with $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Show that:

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 - (\theta - \bar{X}_n)^2.$$

Question 11

- Compute $\mathbb{E}(\theta - \bar{X}_n)^2$.
- Prove that $\hat{\theta}_2$ is an biased estimator of θ and give the bias. How can we get an unbiased estimator?

Part 2: Application to Premier League scores.

We want to model the number of goals during a soccer game by a Poisson distribution. We first model the number of goals of the local team and of the visiting team are independent Poisson, resp. with parameter $\lambda > 0$ and $\mu > 0$.

Question 1

- Load the season-1718_csv file and describe what it contains.
- What do the variables FTHG, FTAG, FTR correspond to?

Question 2

Exploring the dataset:

- Compute the number of points over the season of each team (victory = 3 points, draw=1 point), the number of points in "home" matches, the number of points in "away" matches. How many points did Arsenal score and what was its rank?
- Compare the histogram of the total number of points at home and away.
- Fit a density to those histograms.

Question 3

- Write the statistical model associated to the observation of n match results. Do you think it is a realistic model?
- Propose a method to estimate λ and μ.

Question 4

- Compute the empirical mean and variance of the number of goals of 1)the visiting team 2) the home team.
- Compute the MLE estimators (of λ, μ) for the Poisson model.
- Does the Poisson assumption look correct?

Question 5

- Compute the confidence intervals for λ derived at question 7, and a similar confidence interval for μ.
- Do you think the distribution of the number of goals scored by the home team and the visiting team is the same?

Question 6

- What would be the best approach to answer the previous question? Formalize the problem as a testing problem.
- Use a t.test to give a more precise answer.
- Comment on the assumptions of such a test. Are they valid, "nearly valid", or problematic?
- What is the p-value of the test? What does that mean?
- If you want a test of level $\alpha = 0.05$, do you accept or reject the null hypothesis?

We now only focus on the number of goals scored by the home team.

Question 7

- Compute the 2 confidence intervals for λ derived at question 7, 9. Which one would you rather use?

Question 8*

The two favourite for the title of champion of Premier League this year are Liverpool and Manchester City. Which one was the best in 2017/2018? To answer that question, we will compare the offence of both teams.

- Create two vectors ManCity and Liverpool containing the goals scored during the season, both away and home.
- Formalise the previous question as a testing problem and use a t.test to answer it.
- If you want a test of level $\alpha = 0.05$, do you accept or reject the null hypothesis? How would you conlude on which team has the best offence for the season 2017/2018?