Exam_Vietnam_2024

Laurent DELSOL

28/12/2024

All the exercises are independent and may be done in the order you want. You may use R software at any time it may help you. You can use the result stated in a previous question even if you have not been able to prove it. The **hints** proof is not required. Please include all the commands you want to use (even the codes given during the lecture) in your script. At the end of the exam, send your script by email at laurent.delsol@univ-orleans.fr.

Exercise 1: Gold prospector

A gold-bearing river is known to have a distribution of gold nugget diameters following an $\mathcal{E}(1)$ distribution. A gold prospector is prospecting with a sieve with meshes of unknown diameter θ. The size of a nugget harvested, denoted X, then has the density

$$f_\theta(x) = e^{-(x-\theta)} 1_{\{x \geq \theta\}}.$$

The variable **size** in the dataset **Nuggets.RData** provides the sizes of the nuggets harvested (observed values of X).

1. Load the data **Nuggets.RData** and compute the sample size, the modalities **moda**, the frequencies **e**, the proportions **f**, and the cumulative proportions **Fc**. Then use the following commands to compare the empirical cumulative distribution function with the cumulative distribution function of $\mathcal{E}(1)$. Comment and try to explain the difference you observe.

```
tau=1
u=seq(min(moda)-tau,max(moda)+tau,len=10000)
plot(c(min(moda)-tau,moda,max(moda)+tau),c(0,Fc,1),type='s')
lines(u,pexp(u,1),col='red')
```

2. Show that $\mathbb{E}[X] = \theta + 1$. **Hint: you can use that $X - \theta$ has an exponential distribution with parameter 1.**
3. Deduce by the Method of Moments an estimator $\hat{\theta}_1$ of the parameter θ.
4. Compute the bias and variance of $\hat{\theta}_1$.
5. Give a sufficient statistics for θ.
6. Show that the maximum likelihood estimator is $\hat{\theta}_2 = \min_{1 \leq i \leq n} X_i$.
7. Prove that $\hat{\theta}_2$ has the density
$$g_{\theta,n}(x) = ne^{-n(x-\theta)} 1_{\{x \geq \theta\}}.$$
8. Calculate the bias and variance of $\hat{\theta}_2$. **Hint: the result stated in the previous question means that $\hat{\theta}_2 - \theta$ has an exponential distribution with parameter n.**
9. Construct an unbiased estimator $\hat{\theta}_3$ from $\hat{\theta}_2$. Calculate the variance of $\hat{\theta}_3$.
10. Compare the squared errors of $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$. Which estimator is the best?

11. Compare finally the empirical cumulative distribution function with the cumulative distribution function of $\mathcal{E}(\hat{\theta}_1), \mathcal{E}(\hat{\theta}_2)$, and $\mathcal{E}(\hat{\theta}_3)$. Comment. **Hint: use the command `lines(u,pexp(u-theta,1),col='red')` to add the cumulative distribution function of $\mathcal{E}(\theta)$.**

Exercise 2: Escape room

A new escape game has just opened in HCMC. The exit door only opens if the team has solved a series of random puzzles. At each try, one of the two paths (simple or complex) is chosen at random (with probabilities 1/4 and 3/4 respectively). The simple path has a success rate of θ while the complex path has a success rate of $\frac{\theta}{3}$. The probability of opening the door at each try is therefore $p = \frac{1}{4}\theta + \frac{3}{4}\frac{\theta}{3} = \frac{\theta}{2}$. At each failure, the game starts again from scratch and a new path is chosen randomly.
Consequently, the number of unsuccessful attempts before opening the door follows a $\mathcal{G}(\frac{\theta}{2})$ distribution given by:

$$\forall k \in \mathbb{N},\ \mathbb{P}(X = k) = \frac{\theta}{2}\left(1 - \frac{\theta}{2}\right)^k.$$

The aim is to estimate $\theta \in (0, 1)$.

Load the **Escape_Game.RData** working environment which contains X values, through the variable **unsuccessful_attempts**.

1. Precise the population, the type of variable and the sample size **n**.
2. Represent the empirical distribution through a relevant graphic. Give the value of the mode.
3. Let's define $Y = 1_{\{X=0\}}$ and for any $1 \leq i \leq n$, the random variable $Y_i = 1_{\{X_i=0\}}$. Prove $\mathbb{E}[Y] = \mathbb{P}(X = 0) = \frac{\theta}{2}$ and deduce an unbiased estimator from $Y_1, \ldots, Y_n$ using the Moments method. Compute its bias and variance.
4. Prove $\mathbb{E}[X] = \frac{2-\theta}{\theta}$. **Hint:** $\sum_{k=0}^{\infty} k(1 - \frac{\theta}{2})^{k-1} = \frac{4}{\theta^2}$.
5. Compute the Fisher information of the sample. Is $\hat{\theta}_1$ an efficient estimator of θ?
6. Prove the distribution of X belongs to the exponential family and explain why $\overline{X}$ is a sufficient and complete statistics.
7. Deduce from question 4 a new estimator $\hat{\theta}_2$ of θ from the Moments method.
8. Unfortunately, the bias and variance of $\hat{\theta}_2$ are very difficult to compute. Consider instead $\hat{\theta}_3 = 2\frac{1-\frac{1}{n}}{1-\frac{1}{n}+\overline{X}} = 2\frac{n-1}{n-1+n\overline{X}}$. Prove $\hat{\theta}_3$ is an unbiased estimator of θ. **Hint:** $\forall k \in \mathbb{N}$, $\mathbb{P}(X_1 + \cdots + X_n = k) = \frac{(n-1+k)!}{k!(n-1)!}\left(1 - \frac{\theta}{2}\right)^k \left(\frac{\theta}{2}\right)^n$, **and** $\mathbb{E}[\hat{\theta}_3] = \sum_{k=0}^{\infty} 2\frac{n-1}{n-1+k}\mathbb{P}(X_1 + \cdots + X_n = k)$.
9. Explain why there is no need to compute the variance of $\hat{\theta}_3$ to prove $\hat{\theta}_3$ is the best unbiased estimator of θ. **Hint: use questions 7 and 9.**
10. Compute the value of the three estimators $\hat{\theta}_1$, $\hat{\theta}_2$,and $\hat{\theta}_3$.

Exercise 3: Marathon

This exercise is devoted to the study of the **Marathon** dataset. For this, you have a **Marathon.RData** file containing the data.

- the variable **Time** indicates the travel time (in minutes) of the candidates
- the variable **Type** indicates the type of affiliation of the candidates (Professional, Club member, Amateur)

A. Study of the variable Time

1. What is the nature of the data? What is the sample size? Discuss in a few lines the interest of a non-parametric approach at this stage of the study.
2. Give the expression of the kernel estimator of the density. Demonstrate that if the kernel K is a density, then this is also true for the estimator.
3. Estimate the density of the variable **Time** using an Epanechnikov kernel and a smoothing parameter of 0.5,5, 50 and the one chosen by cross-validation. Comment on the results obtained.
4. Give an estimate of the 3 main modes. Can you make a guess about the affiliations of the 3 groups of athletes thus revealed.
5. Give an estimate of the 2 minima **m1** and **m2** between the modes obtained in question (4). Use them to define 3 clusters and compare them to the affilation (variable **Type**). **Hint: You can use the following commands**

```
cluster=rep(1,length(Time))
cluster[(m1>=Time) & (Time<m2)]=2
cluster[(Time>=m2)]=3
table(cluster,Type)
```

B. Supervised classification

1. Give the definition of the Bayes classifier. Can we use it directly here? Why? How can you obtain a classification method from kernel estimators?
2. Use the variables **Type** and **Time** and the function **Classif_NP** seen in class to perform a supervised classification of the runners' times (according to their affiliation). Comment on the results obtained (confusion matrix, error rate).
3. Here are the times obtained by 7 new competitors: 131.682, 171.873, 153.045, 188.223, 170.289, 154.824, 217.248. Estimate their probable affiliation. Are you confident about these estimated affiliations?

Exercise 4: Hidden object

An object of unknown shape and nature may be hidden inside a cake. You cannot access it completely or observe it but can in places measure how deep you can push a very fine needle. From these observations we expect to be able to estimate the depth for the whole cake, for every potential position (abscissa and ordinate), to identify the hidden object.

You can retrieve the observations collected in the **Hidden_Object.RData** dataset.

- **x1** and **x2** indicate the position (abscissa and ordinate)
- **d** indicates the depth

1. Indicate the nature of the data, the sample size and the nature of the problem considered.
2. Estimate the density of the variable **d** with a suitable smoothing parameter. From the result obtained, do you think there is any object hidden in the cake?
3. Estimate the regression function of the variable **d** on (**x1** , **x2**) with a suitable smoothing parameter (use options **nbins=0**, and **ngrid=100**). From the result obtained, identify the hidden object. **Hint:** save the result given by the **sm.regression** in a variable **res** and visualize the result obtained by applying the image function to the estimate matrix obtained: `image(res$estimate)`.