

An introduction to Mathematical Statistics

Exam 2025

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. 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**.

The given marking scheme may change a little. The total is greater than 20 to take into account the length of the exercises with respect to the exam duration. Every **BONUS** question gives 1 additional point.

Exercise 1: Catch mark release recapture

Our aim is to estimate the unknown number of fishes living in a lake, denoted θ hereafter. because we have to take care of the fishes we have one additional constraint, we can only catch fishes one by one. Hence we have to put into the lake before the next catch. However, we may add one mark on the fish.

The method is the following:

- **[CATCH and MARK]** First catch and mark a known number **M** of fishes
- **[RELEASE]** Then wait a few days to let time to the marked fishes to be uniformly distributed in the whole lake
- **[RECAPTURE]** Finally start fishing again and observe for every caught fish if it is marked or not. The probability to catch a marked fish is then $p_\theta = \frac{M}{\theta}$.

Start by loading the **CMR.RData workspace**. It contains two variables:

- **x**: the observations (of X) collected at the last **RECAPTURE** step (marked: 1, unmarked: 0).
- **y**: the observations corresponding to Y, the number of unmarked fishes caught before the next marked one, collected at the last **RECAPTURE** step.

Note that in that experiment we have set **M=20**.

Part A: X and Bernoulli [Total: 3.5 pts]

The variable **x** contains the observations (of X) collected at the last **RECAPTURE** step (marked: 1, unmarked: 0). Now clearly, $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Ber}(p_\theta)$, with $p_\theta = \frac{M}{\theta}$.

2. Construct an estimator $\hat{\theta}_1$ of θ from the method of moments. [0.75 pt]
3. The bias of $\hat{\theta}_1$ is very difficult to compute. Have a look to $\hat{\theta}_2 = \frac{M(n+1)}{n\bar{X}+1} = \frac{M(1+\frac{1}{n})}{\bar{X}+\frac{1}{n}}$. Prove $\mathbb{E}[\hat{\theta}_2] = \theta(1 - (1 - p_\theta)^n)$ and compute the bias of $\hat{\theta}_2$. Is $\hat{\theta}_2$ asymptotically unbiased? [2 pts]

Hint: $n\bar{X} \sim \text{Bin}(n, p_\theta)$ hence $\mathbb{E}[\hat{\theta}_2] = \sum_{k=0}^n \frac{M(n+1)}{k+1} \frac{n!}{k!(n-k)!} p_\theta^k (1 - p_\theta)^{n-k}$.

4. Compute the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ from the observations. [0.75 pt]

Part B: Y and geometric [Total: 3 pts]

The variable **y** contains the observations corresponding to Y, the number of unmarked fishes caught before the next marked one, collected at the last **RECAPTURE** step. Now clearly, $Y_1, \dots, Y_r \stackrel{i.i.d.}{\sim} \mathcal{G}(p_\theta)$, with $p_\theta = \frac{M}{\theta}$. **Hint:** if $Z \sim \mathcal{G}(p)$ then $\forall k \in \mathbb{N}, \mathbb{P}_p(Z = k) = (1 - p)^k p$, $\mathbb{E}_p[Z] = \frac{1-p}{p}$ and $\text{Var}_p(Z) = \frac{1-p}{p^2}$. [1.5 pt]

2. Construct an estimator $\hat{\theta}_3$ of θ from $Y_1, \dots, Y_r$ and the method of moments. [1 pt]
3. Compute the bias and variance of $\hat{\theta}_3$.
4. Compute the value of $\hat{\theta}_3$ from the observations. [0.5 pt]

Exercise : Archery competition

A great archery competition have organised between three legends of that sport: Rebel, Robin and Legolas.

Start by loading the **Archery2025_2.RData** workspace.

It contains three variables:

- **Archer**: shooter's name
- **X**: distance to the center of the target
- **Pos**: position (in 2 dimensions) of the arrow on the target
- **Pos0**: position (in 2 dimensions) of the 4 final arrows on the target

A. Distance to the center of the target

We observe first **X**, the distance between the arrows and the center of the target.

Part 1: Non parametric estimation: [Total: 3 pts +1pt BONUS]

1. Prove that $\hat{f}$ is a density if K is a density. [1 pt]
2. Give a nonparametric estimation of the density f of **X** using the kernel density estimator $\hat{f}$ using as smoothing parameter 0.1, 1, 10 and the value given by cross-validation criterion (hint: share the graphic window into four plots). Comment. [1.5 pts]
3. Give a nonparametric estimation of the mode of **X**. [0.5 pt]
4. [BONUS] Plot on the same graph the estimated densities obtained for the three archers and compare them. Do you expect supervised classification from **X** and **Archer** could be efficient?

Part 2: Potential parametric estimation: [Total: 7 pts +2pts BONUS]

If the position of the arrows follow a bidimensional gaussian distribution $\mathcal{N}((0,0)^T, \sqrt{\theta}I_2)$, then f should have the following expression :

$$f(x, \theta) = \frac{x}{\theta} e^{-\frac{x^2}{2\theta}} 1_{\{x \geq 0\}},$$

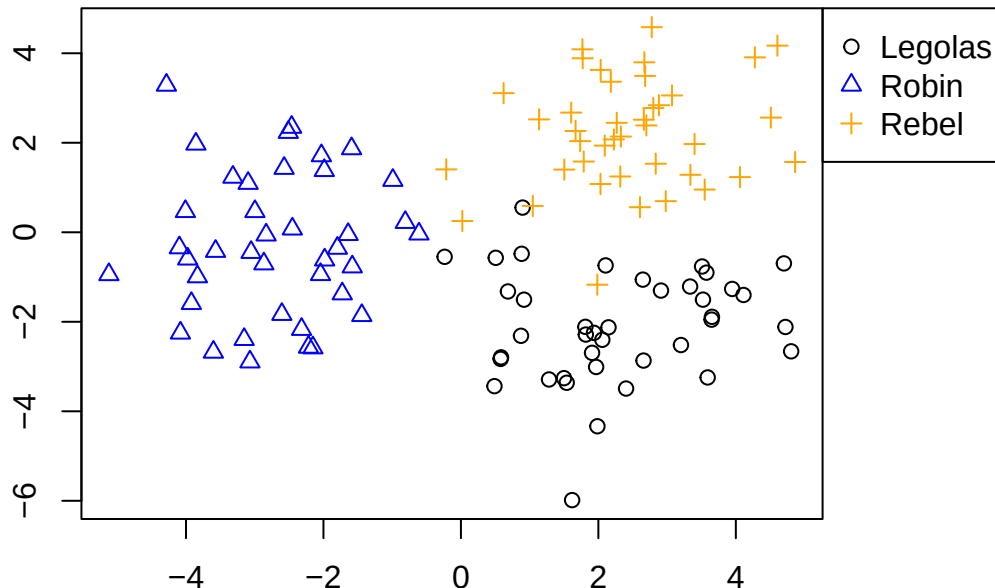
with $\theta > 0$.

1. Now use the parametric model to propose an unbiased estimator $\hat{\theta}$ of θ from maximum likelihood method. [1.5 pts]
2. Compute its bias and quadratic error (**hint**: $Y = \frac{X^2}{\theta} \sim \mathcal{E}(\frac{1}{2})$) [1 pt]
3. Prove that this parametric model defines an exponential family and deduce a sufficient and complete variable. What does this imply for $\hat{\theta}$? [1.75 pt]
4. Compute the Fisher Information of the sample. [1.25 pt]
5. Is $\hat{\theta}$ an efficient estimator of θ ? [0.5 pt]
6. Plot on the same graph the empirical distribution (as in Part 1, question 3 with the best choice of h) and the estimated parametric distribution. Comment the result. [1 pt]
7. [BONUS] Give a parametric estimation of the mode of x from the parametric model and $\hat{\theta}$. **Hint**: You should optimize the function $x \mapsto \frac{x}{\theta} e^{-\frac{x^2}{2\theta}} 1_{\{x \geq 0\}}$ with respect to x .
8. [BONUS] Compare the values of $\hat{\theta}$ obtained for the three archers. Do you expect parametric supervised classification from **X** and **Archer** could be efficient?

B. Discrimination from the position (in 2 dimensions) of the arrow on the target [Total: 3 pts]

The competitors fire their final arrows simultaneously, so it's impossible to know whose arrows they belong to. On the target, the four arrows land at positions (1.4, 1.3), (1.4, -1.2), (-2, 0), and (-0.1, 0.2). How could you identify the archers who fired them?

1. Compare the positions of the arrows for the different archers on the graph below. What do you observe? Why is this promising? [0.5 pt]



1. Use the **Classif_NP** function seen in class to put into practice the method explained in question 2 using the position and Archer variables. Comment on the result. This may take around one minute to run on your laptop. [1.5 pt]
2. Using what you did previously, for each of the last 4 arrows, can you correctly identify the archer who shot it? Explain why? [1 pt]

Exercise 3: Escape [Total: 4.5 pts]

Your friend Edmont Dantes has been unjustly imprisoned for a crime he didn't commit. You find a job as a prison gardener, hoping to help him escape by finding the location of the potential underground tunnel leading to the exit. Under the pretext of working on landscaping the courtyard, you manage, with the utmost discretion, to measure in places how deep you can sink your crowbar into the ground.

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

- **x1** and **x2** indicate the position (x-axis and y-axis)
 - **d** indicates the depth
1. Estimate the density of the variable **d** using a suitable smoothing parameter. How many clusters can you may expect from the estimated density? [1 pt]
 2. Give the two (local and global) modal values, the threshold (valley point between them) and the extract the clusters from the estimated density. Why is this encouraging regarding the hope of finding a hidden underground tunnel? [1.5 pt]
 3. Estimate the regression function of the variable **d** on (**x1** , **x2**) on the grid [1,2,..., 400]x[1,2,..., 400] with a suitable smoothing parameter. You will use the following parameters: **eval.points=cbind(g1,g2)** and **nbins=0**, with **g1=1:400** and **g2=1:400**. Name **R** the matrix

containing the estimated values of the regression function. This may take around one minute to run on your laptop. [1.5 pt]

4. Visualize the result obtained by applying the following commands. Which entry **E1**, **E2**, or **E3** will you indicate to your friend to allow his escape? [0.5 pt]

```
image(R,main='Freedom', xlab='Prison',yaxt="n",xaxt="n")
axis(1,at=c(0.22,0.53, 0.89), labels=c('E1','E2','E3'))
```