An introduction to Mathematical Statistics

Laurent DELSOL
University of Orleans

30 octobre 2019

Chapitre 1

Fondamental theorems

Exercise 1 Flip a coin
Flip a coin 20 times consecutively and write the obtained results (Heads and Tails) in the following way

$$HTTHHHTHTTHT...$$

Part A : Probability models

1. Which is the law of the result X_i of each flip ? Is there an unknown parameter ?
2. Let Y_{20} be the number of tails obtained over the 20 flips. What is the law of Y_{20} ?
3. Now let Z be the number of heads obtained before the first tail. What is the law of Z ?

Part B : Statistical experiments

1. Compare the proportion of tails you obtained with the one obtained by the other students. Why are them different ? Are them completely random ?
2. Do the same for the number of tails over the 20 flips and the number of heads before the first tail (gathering the values of all the students).
3. Compare the obtained values with the expectation of the variables X_i, Y_{20} and Z. What happens ?

Part C : Law of Large Numbers and Central Limit Theorem

1. Give estimations of the probability p of obtaining a tail (using the three approaches listed above).
2. Could you go further and propose a confidence interval for p ?

Exercise 2 Gaussian samples

Let $X_1, \ldots, X_n$ be a sample of independent and identically distributed random variables form a $\mathcal{N}(\mu, \sigma)$ distribution.

1. What do μ and σ represent for the sample distribution ?
2. What is the law of $Y = \frac{X-\mu}{\sigma}$?
3. What is the law of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?
4. Infer from this the law of $Z_n = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$.
5. When σ is unknown, one usually considers $T_n = \frac{\overline{X}-\mu}{\frac{S_{n-1}}{\sqrt{n}}}$. What is its law ?

6. Assume now the sample is not gaussian but its law admits a second order moment. Are the previous results still true ?

Exercise 3 Convergence of random variables

Prove the following statements

1. $(X_n \overset{\mathbb{P}}{\to} X) \Rightarrow (X_n \overset{\mathcal{L}}{\to} X)$
2. For all constant a, $(X_n \overset{\mathcal{L}}{\to} a) \Rightarrow (X_n \overset{\mathbb{P}}{\to} a)$
3. For all $\alpha > 0$, $(X_n \overset{\mathbb{L}^\alpha}{\to} X) \Rightarrow (X_n \overset{\mathbb{P}}{\to} X)$
4. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function and $X_n \overset{\mathcal{L}}{\to} X$. Then, $f(X_n) \overset{\mathcal{L}}{\to} f(X)$
5. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function and $X_n \overset{\mathbb{P}}{\to} X$. Then, $f(X_n) \overset{\mathbb{P}}{\to} f(X)$
6. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function and $X_n \overset{a.s.}{\to} X$. Then, $f(X_n) \overset{a.s.}{\to} f(X)$
7. $(X_n \overset{\mathcal{L}}{\to} X$ and $Y_n \overset{\mathcal{L}}{\to} y$ constant)$\Rightarrow (X_n, Y_n) \overset{\mathcal{L}}{\to} (X, y)$

Exercise 4 Counterexamples

Part 1 :
Let $(X_n)_{n\in\mathbb{N}}$ such that $\mathbb{P}(X_n = n) = \frac{1}{n}$ and $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$

1. Prove $X_n \overset{\mathcal{L}}{\to} 0$
2. Prove X_n does not tend to 0 in $\mathbb{L}^1$ (and hence in $\mathbb{L}^2$).

Part 2 :
Let X a centered gaussian random variable. Define $Y = -X$, and $X_n = X$.

1. Prove $X_n \overset{\mathcal{L}}{\to} Y$.
2. Prove X_n does not tend to Y in probability.

Part 3 :
Let $(X_n)_{n\in\mathbb{N}}$ such that $\mathbb{P}(X_n = 1) = \frac{1}{n}$ and $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$

1. Prove $X_n \overset{\mathcal{L}}{\to} 0$
2. Prove $X_n \overset{\mathbb{L}^2}{\to} 0$
3. Prove X_n does not tend to 0 in almost surely (use the converse Borel Cantelli Lemma).

Chapitre 2

Statistical models, estimators and their properties

Exercise 5
Prove the following inequality

$$\mathbb{E}_\theta\left[(\hat{\theta}-\theta)^2\right] = \left(\mathbb{E}_\theta[\hat{\theta}]-\theta\right)^2 + Var_\theta\left(\hat{\theta}\right).$$

Exercise 6
In a given population, the probability of an animal to be sick is p. Let X stands for the status (1 for sick, 0 for sane) of an animal chosen at random in this population.

Probabilities :

1. Give a probabilistic model for this experience
2. Give the distribution of X and compute its cumulative distribution function
3. Compute $\mathbb{E}[X]$ and $Var(X)$.

Statistics :
We consider now a n-sample $X_1, \dots, X_n \overset{i.i.d.}{\sim} X$ and denote $\overline{X} = \frac{X_1+\cdots+X_n}{n}$.

1. Give a statistical model for this experience and precise its natural parameter and its parameter of interest.
2. Compute $\mathbb{E}\left[\overline{X}\right]$ and $Var(\overline{X})$.
3. Infer from this an unbiased estimator of p.
4. Prove its quadratic mean convergence.

Exercise 7
Given an unbiased dice whose faces are numbered from one to six, we denote X the face obtained when the dice is thrown.

Probabilities :

1. Give a probabilistic model for this experience
2. Give the distribution of X and compute $\mathbb{E}[X]$ and $Var(X)$.
3. Could you do the same for a k faces unbiased dice ?

Statistics :
We consider now a biased 6 faces dice and note $p_1, \dots, p_n$ the unknown probabilities of obtaining each face. Given the results obtained by throwing this dice independently n times, the aim is to estimate the probability q of obtaining an even score.

1. Give a statistical model for this experience and precise its natural parameter and its parameter of interest.
2. Construct an unbiased estimator of p.
3. Prove its quadratic mean convergence.

Exercise 8
A statistical survey is made to estimate the proportion p of drug consumers in a population. Of course, if the question is posed directly, people will probably lie and the observations will be useless. An alternative is to include randomness in our survey to ensure to each participant that is its answer does not provide, in any way, the possibility to know if he consumes drugs or not.

For example, propose to each participant to throw a ball from a box containing a proportion $\theta \neq \frac{1}{2}$ of blue balls (away form the others) and answer to two different questions according to the obtained result.

— If the ball is blue, answer "Have you ever consumed drugs ?"
— If the ball is not blue, answer "Have you never consumed drugs ?"

Denote $X_1, \ldots, X_n$ the final answers obtained with this survey.

1. Give a statistical model for this experience and precise its natural parameter and its parameter of interest.
2. What is, as a function of p the probability q that an individual answers "YES".
3. Construct an unbiased estimator of q.
4. Infer from this an unbiased estimated of p.
5. Prove its quadratic mean convergence.

Exercise 9
Denote by X a random variable following a Poisson distribution with parameter $\theta \in \mathbb{R}_*^+$, what means

$$\forall n \in \mathbb{N}, \mathbb{P}\left(X = n\right) = C\frac{\theta^n}{n!}$$

Probabilities :

1. Compute the value of C.
2. Compute $\mathbb{E}\left[X\right]$ and $Var(X)$.
3. Given $X_1 \sim \mathcal{P}\left(\theta_1\right)$ and $X_2 \sim \mathcal{P}\left(\theta_2\right)$ two independent random variables, give the distribution of $X_1 + X_2$
4. Infer from this the distribution of $X_1 + \cdots + X_n$ where the X_i's are independent $\mathcal{P}\left(\theta_i\right)$ variables.

Statistics : The length of the queue at the entrance of an attraction may be represented by a random variable $X \sim \mathcal{P}(\theta)$, where θ is unknown. One is interested in the mean length of the queue.

1. Give a statistical model for this experience and precise its natural parameter and its parameter of interest.
2. Compute $\mathbb{E}[\overline{X}]$ and $Var(\overline{X})$ using two different methods
3. Infer from this an unbiased estimator of the mean length of the queue.
4. Prove its quadratic mean convergence.

Exercise 10 Exponential variables
Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{E}(\theta)$.

Recall that $\mathcal{E}(\theta) = \Gamma(1, \theta)$ and two properties of Γ distributions :

- If X_i has a $\Gamma(k_i, \theta)$ distribution for $i = 1, 2, ..., N$ (i.e., all distributions have the same scale parameter θ), then $\sum_{i=1}^{N} X_i \sim \Gamma\left(\sum_{i=1}^{N} k_i, \theta\right)$ provided all X_i are independent.
- If $X \sim \Gamma(k, \theta)$, then, for any $c > 0$, $cX \sim \Gamma(k, c\theta)$.

1. What is the distribution of $\overline{X}$?
2. Compute $\mathbb{E}[\overline{X}]$ and $Var(\overline{X})$.
3. Prove that $\mathbb{E}[\frac{1}{\overline{X}}] \overset{n\to+\infty}{\to} \theta$
4. Prove that $\hat{\theta} = \frac{1}{\overline{X}}$ tends to θ in probability and in $\mathbb{L}^2$
5. Give the asymptotic distribution of $\sqrt{n}\left(\hat{\theta} - \theta\right)$.

Exercise 11 Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.
On is interested in the estimation of σ^2.

- Prove $\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ converges almost surely to σ^2 and compute its bias and variance.
- Compute $\mathbb{E}[|X|]$ and infer an estimate $\hat{\sigma}_2$ of σ. Does it converges almost surely to σ^2 ? Compute its bias and variance.
- Which estmator would you recommend ?

Exercise 12 Linear regression
Let $(X_i)_{1\leq i\leq n}$ and $(\epsilon_i)_{1\leq i\leq n}$ be two independent sequences of i.i.d ; random variables. Assume that X_1 and ϵ_1 are square integrable, ϵ is centered and $Var(X_1) > 0$.

For $a \in \mathbb{R}$ define the regression model $Y_i = aX_i + \epsilon_i$.

1. Prove that $\hat{a} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$ converges almost surely to a.
2. Give the asymptotic distribution of $\sqrt{n}\,(\hat{a} - a)$.

Chapitre 3

How to construct interesting estimators ?

Exercise 13 Method of Moments and Maximum Likelihood
Construct using both methods estimators of the unknown parameters when the law of the n-sample of i.i.d. random variables is :

1. $Bin\left(N,\theta\right)$, where N is supposed to be known.
2. $\mathcal{N}\left(\mu,\sigma\right)$
3. $\mathcal{G}\left(p\right)$
4. a law with density $f : x \mapsto \exp(-(x-\theta))1_{x>\theta}$

Exercise 14 Poisson distribution
Let $x_1,\ldots,X_n \overset{i.i.d.}{\sim} \mathcal{P}\left(\theta\right)$.

Estimation of θ :

1. Propose an estimator of θ using the method of moments.
2. Compute its bias and variance.
3. Does the Maximum Likelihood give an other estimator ?

Estimation of $\lambda = e^{-\theta}$:

1. Link λ to the probability of an event.
2. Propose an estimator $\hat{\lambda}_1$ of λ using the method of moments.
3. Compute its bias and variance.
4. Do the same for the estimator $\hat{\lambda}_2$ obtained by Maximum Likelihood.
5. Compare $\hat{\lambda}_1$ and $\hat{\lambda}_2$ mean squared errors.

Exercise 15 Bernoulli distribution
Let $x_1,\ldots,X_n \overset{i.i.d.}{\sim} B\left(\theta\right), \theta \in \left]0;1\right[$.

Estimation of θ :

1. Propose an estimator $\hat{\theta}_1$ of θ using the method of moments or the maximum likelihood.
2. Compute its bias and variance.

Estimation of $\lambda = \theta\left(1-\theta\right)$:

1. Compute bias and variance of $\hat{\lambda}_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
2. Compute bias and variance of $\hat{\lambda}_2 = \overline{X}\left(1-\overline{X}\right)$.

3. Compare $\hat{\lambda}_1$ and $\hat{\lambda}_2$ mean squared errors.

Exercise 16 Uniform distribution
Let $(X_1, \ldots, X_n) \overset{i.i.d.}{\sim} X$, where $X \sim \mathcal{U}([0, \theta])$.

1. Give the density of X and compute $\mathbb{E}[X]$ and $Var(X)$.
2. Give the estimator $\hat{\theta}_1$ obtained by the method of moments.
3. Compute its bias and variance.
4. Give the estimator $\hat{\theta}_2$ obtained by maximum likelihood.
5. Compute its bias and variance.
6. We now look the best estimator, in terms of quadratic mean, of the form $\hat{\theta}_3 = c\hat{\theta}_2$, where c is a real constant.
7. Compute its bias and variance.
8. Compare $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$.

Exercise 17 Exponential distribution

Let $(X_1, \ldots, X_n) \overset{i.i.d.}{\sim} X$, where $X \sim \mathcal{E}\left(\frac{1}{\theta}\right)$.

1. Give the density of X and compute $\mathbb{E}[X]$ and $Var(X)$.
2. Give the estimator $\hat{\theta}_1$ obtained by the method of moments.
3. Compute its bias and variance.
4. Give the estimator $\hat{\theta}_2$ obtained by maximum likelihood.
5. Compute its bias and variance.
6. Give the distribution of $Z = min_{i=1,\ldots,n} X_i$.
7. Compute $\mathbb{E}[Z]$ and $Var(Z)$.
8. Infer from this an unbiased estimator $\hat{\theta}_3$
9. Compute its bias and variance.
10. Compare $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$.

Chapitre 4

Sufficiency, Likelihood, Exponential families, Cramer-Rao Lower Bound

Exercise 18 Let's consider a random sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$ following a Pareto distribution with parameters $\theta = (\alpha, \lambda)$ with $\alpha > 1$ and $\lambda > 0$ whose density is given by

$$f_X(x, \alpha, \theta) = Cx^{-\alpha} 1_{x>\lambda}$$

1. Compute the value of C.
2. Provide a sufficient statistics for θ.
3. Assume now λ value is known and compute the Fisher Information (on α) of the random sample.

Exercise 19 Consider a random sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$, where $X \sim \mathcal{U}([a; b])$ and $a < b$. Provide a sufficient statistics for $\theta = (a, b)$.

Exercise 20 Consider a random sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$, where $X \sim \mathcal{G}(\theta)$ and $0 < \theta < 1$. Provide a sufficient statistics S for θ and compute the Fisher information given by S.

Exercise 21 Consider a random sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$, where

$$f_X(x) = e^{-(x-\theta)} 1_{x>\theta}$$

Provide a sufficient statistics for θ and compute the Fisher information given by S.

Exercise 22 Consider a random sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$, where $X \sim \mathcal{N}(\theta, \theta)$ and $\theta > 0$. Compute the Fisher Information of the sample. Can you give an efficient estimator of θ ?

Exercise 23 For each set of probability distributions listed below, prove that it belongs to the exponential family and provide an exhaustive and complete statistics.

— $\mathcal{E}(\theta), \theta > 0$

— $\mathcal{N}(\mu, \sigma), \theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^* +$

— $\mathcal{P}(\theta), \theta > 0$

— $\mathcal{G}(\theta), 0 < \theta < 1$

Exercise 24 Loi Gamma
Recall the following formula :

$$\forall \theta > 0, \forall p \in \mathbb{N}, \int_0^{+\infty} e^{-\theta x} x^{p-1} dx = \frac{(p-1)!}{\theta^p} 0.$$

The time spent on a website may be modelized by $\Gamma(2, \theta)$ distribution whose density is given by

$$f_X(x, \theta) = \theta^2 e^{-\theta x} x 1_{x>0}.$$

1. Give the maximum likelihood estimator θ_{ML} of θ.
2. Is it biased ? Could you propose an unbiased estimator ?
3. For $n \geq 2$ compute the variance of this unbiased estimator and compare it to the Cramer Rao lower bound.

Exercise 25 The time spent to treat an administrative problem may be seen as a realisation of a variable X whose density is given by $f_X(x, \theta) = \frac{1}{\theta} x^{-1-\frac{1}{\theta}} 1_{x \geq 1}$ where $\theta > 0$.

1. Compute an estimator $\hat{\theta}_1$ by the method of moments
2. Compute an other estimator $\hat{\theta}_2$ by the maximum likelihood
3. Compute bias and variance of these estimators. Hint : $\ln(X) \sim \mathcal{E}\left(\frac{1}{\theta}\right)$.

Chapitre 5

Confidence and fluctuation intervals

Chapitre 6

Statistical tests