Exam

November 2020
Duration : 3h

No electronic device, no documents except one hand-written page, both sides.

Pareto distribution. A random variable X has a Pareto distribution with parameters $a > 0$ and $\theta > 0$ if its cumulative distribution function is :

$$\forall x \in \mathbb{R}, \quad P(X \leq x) = \begin{cases} 0 & \text{si } x \leq a \\ 1 - \left(\frac{a}{x}\right)^\theta & \text{si } x > a. \end{cases}$$

Exponential distribution. A random variable X has an exponential distribution with parameter $\lambda > 0$ if it has the density :

$$\forall x \in \mathbb{R}, \quad f_X(x) = \lambda e^{-\lambda x} \mathbb{1}_{\mathbb{R}_+}(x).$$

Beta distribution. A random variable X has a Beta distribution with parameters $a > 0$, $b > 0$, if it has the density :

$$\forall x \in \mathbb{R}, \quad f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \mathbb{1}_{[0,1]}(x),$$

where Γ is the Gamma function. We have :

$$\mathbb{E}(X) = \frac{a}{a+b}.$$

Geometric distribution. A random variable X has a geometric distribution with parameter $p > 0$ if :

$$\forall x \in \mathbb{N} \setminus \{0\}, \quad \mathbb{P}(X = x) = p(1-p)^{x-1}.$$

1 Maximum likelihood

We consider the statistical model $\{P_\theta, \theta > 0\}$ formed by the Pareto distribution, with known parameter $a > 0$ and unknown parameter θ. We have n i.i.d. observations $X_1, \ldots, X_n$ of the distribution P_θ.

1. Show that the distribution P_θ has density :

$$\forall x \in \mathbb{R}, \quad p_\theta(x) = \theta \frac{a^\theta}{x^{\theta+1}} \mathbb{1}_{\{x>a\}}.$$

2. Give the density of the vector $X = (X_1, \ldots, X_n)$.
3. Show that the maximum likelihood estimator $\widehat{\theta}_{ML}$ of θ is :

$$\forall x \in \mathbb{R}^n, \quad \widehat{\theta}_{ML}(x) = \frac{n}{\sum_{i=1}^n \log\left(\frac{x_i}{a}\right)}.$$

4. Give the almost sure limit of $\widehat{\theta}_{ML}(X)$ when $n \to +\infty$.

2 Quadratic risk

The exponential distribution is a standard model for continuous waiting times. We are interested in a variant where a sample waiting time is the sum of some fixed quantity $\theta > 0$ (the minimum waiting time) and a random variable with exponential distribution with parameter 1. We have n i.i.d. observations $X_1, \dots, X_n$ of this waiting time. We denote by P_θ the corresponding distribution.

1. Propose an estimator $\widehat{\theta}$ of θ based on the method of moments, using the expectation of the distribution P_θ.
2. Compute the quadratic risk of the estimator $\widehat{\theta}$.
3. We now consider the estimator $\tilde{\theta}$ defined by :

$$\forall x \in \mathbb{R}^n, \quad \tilde{\theta}(x) = \min_{1 \le i \le n} x_i.$$

   Show that the random variable $\tilde{\theta}(X)$, with $X = (X_1, \dots, X_n)$, is the sum of θ and an exponential random variable with some parameter to be determined.
4. Is the estimator $\tilde{\theta}$ biased ?
5. Compute the quadratic risk of the estimator $\tilde{\theta}$.
6. Given previous results, which estimator of θ would you recommend ?

3 Bayesian statistics

The geometric distribution is a standard model for discrete waiting times. We have n i.i.d. observations $X_1, \dots, X_n$ of the geometric distribution with parameter $\theta \in]0, 1[$, say P_θ. We denote by X the vector $(X_1, \dots, X_n)$. We consider a Bayesian setting where the parameter θ is random ; its prior π is supposed to be uniform over the interval $]0, 1[$.

1. Show that for all $x_1, \dots, x_n \in \mathbb{N} \setminus \{0\}$, the posterior distribution $\pi(\,\cdot\,|x)$ of θ given $X = x$, with $x = (x_1, \dots, x_n)$, is a Beta distribution with parameters to be determined.
2. What is the posterior expectation of θ given $X = x$? We denote by $\widehat{\theta}$ this estimator
3. What is the almost sure limit of $\widehat{\theta}$ when $n \to +\infty$?

4 Hypothesis testing

To be registred as "organic", a farmer must guarantee a percentage of Genetically Modified Organisms (GMO) less than 1%. She takes n samples and measures the percentage of GMO in each. We denote by X_i the logarithm base 10 of the percentage of GMO in sample i, for $= 1, \dots, n$. We assume that the random variables $X_1, \dots, X_n$ are i.i.d. with Gaussian distribution with unknown mean θ and known variance σ^2.

1. We want to test the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$, with $\theta_1 > \theta_0$. Show that the Neyman-Pearson test of level α is of the form :

$$\forall x \in \mathbb{R}^n, \quad \delta(x) = \mathbb{1}_{\{\bar{x} > c\}},$$

   where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the empirical mean of observations, and c is a constant to be determined with respect to θ_0 or θ_1, σ, n and a quantile of the standard normal distribution.
2. The farmer considers that the perrcentage of GMO is less than 1%, until proven otherwise. She wants to test the null hypothesis $H_0 : \theta \le 0$ against the alternative hypothesis $H_1 : \theta > 0$, with type I error rate less than 5%. Compute a constant c such that :

$$\sup_{\theta \le 0} \mathbb{P}_\theta(\bar{X} > c) = 5\%,$$

with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\sigma = \frac{1}{2}$ and $n = 25$. You can use the values of quantiles of the standard normal distribution given in the Appendix.

3. A "non-GMO" organization wants to make sure that there is no more than 1% GMO in organic food. In particular, it worries about the ability of the test to detect products with percentage of GMO higher by 50% than the allowed maximum. Show that the probability that the test of the farmer does *not* reject the null hypothesis H_0 when the percentage of GMO is 1.5% (i.e., $\theta \approx 0.176$) is approximately equal to 50%.
4. Given the previous result, the organization asks the farmer to prove that the percentage of GMO is less than 1%. It considers that the percentage of GMO is higher than 1% unless proved otherwise. Thus it considers the null hypothesis $H_0 : \theta > 0$ and the alternative hypothesis $H_1 : \theta \leq 0$. Propose a test of level α of H_0 against H_1.

5 Confidence interval

In the previous exercise, we now want to build a confidence interval for θ.

1. Propose a confidence interval at level 99% of the form $[\bar{X} - c, \bar{X} + c]$, where c is some constant to be determined with respect to σ, n and some quantile of the standard normal distribution. Give the interval obtained for $\sigma = \frac{1}{2}$, $\bar{X} = 0.8\%$ and $n = 25$.
2. Propose an upper confidence bound at level 99%.

Appendix

Quantiles of the standard normal distribution

The following table gives some approximate values of quantiles of the standard normal distribution.

x	0.0	0.25	0.52	0.84	1.28	1.64	2.33	2.58
$\mathbb{P}(X \leq x)$	0.5	0.6	0.7	0.8	0.9	0.95	0.99	0.995