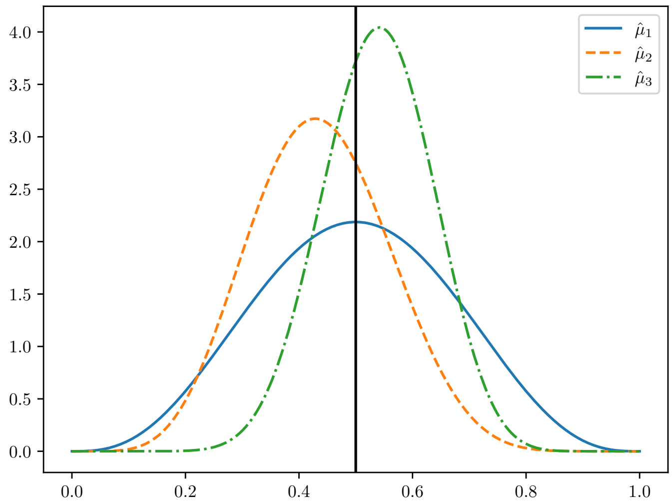
University of Lille
M1 Data Science - Stats 2

2023-2024
April 10th, 2024

Final Exam

Exercise 1 On the figure below the densities of the distribution of three estimators $\hat{\mu}_1 = h_1(X)$, $\hat{\mu}_2 = h_2(X)$ and $\hat{\mu}_3 = h_3(X)$ of a parameter μ are displayed, where X is generated from the distribution P_μ with $\mu = 0.5$. Discuss the relative merits of the three estimators.



Exercise 2 We perform a survey independently on n individuals in a population. The outcome of the survey can be summarized by a score X_i (that can be either positive or negative) observed for each individual. We define by p the probability that an individual has a positive experience. Letting

$$Y_i = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{if } X_i \leq 0 \end{cases}$$

we have $p = \mathbb{P}(Y_i = 1)$. We assume that the individual's scores $X_1, \ldots, X_n$ are iid from a $\mathcal{N}(\theta, 1)$ distribution, whose density is given by

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right), \quad \text{for all } \ x \in \mathbb{R} .$$

We recall that the maximum likelihood estimator of θ is given by $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

1. Compute the biais, the variance and the mean square error of $\widehat{\theta}_n$.
2. Prove that $p = \Phi(\theta)$ where Φ denotes the cdf of a standard normal distribution.

   Deduce a consistent estimator of p, denoted by $\widehat{p}_n$.

3. Using the Delta method, derive the asymptotic distribution of $\widehat{p}_n$.

4. We define $\widetilde{p}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Show that $\widetilde{p}_n$ is an unbiased estimator of p. What is its variance?

5. What is the asymptotic distribution of $\widetilde{p}_n$? Compare its asymptotic variance to that of $\widehat{p}_n$.

6. Now assume that the (X_i) are not really Gaussian but come from another distribution F with mean μ. Is $\widehat{p}_n$ still consistent? Same question for $\widetilde{p}_n$.

Exercise 3 We observe a n sample $X_1, \dots, X_n$ from a geometric distribution $\mathcal{G}(\theta)$, whose probability mass function is given by

$$f_\theta(k) = \theta(1-\theta)^{k-1} \quad \text{for all } k \in \mathbb{N}\backslash\{0\}$$

for some $\theta \in (0,1)$. Given $0 < \theta_0 < \theta_1 < 1$ we are first interested in testing

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta = \theta_1)$$

1. Prove that a Likelihood Ratio Test can be expressed as $D(X) = \mathbb{1}(\sum_{i=1}^n X_i < t)$ for some threshold $t \in \mathbb{N}$ with $t > n+1$.

2. For what values of $\alpha \in (0,1)$ can we find a LRT that is Uniformly More Powerful at level α?

Now we turn our attention to the composite hypothesis testing problem

$$\mathcal{H}_0 : (\theta = \theta_0) \quad \text{against} \quad \mathcal{H}_1 : (\theta \neq \theta_0)$$

where $\theta_0 \in (0,1)$.

3. Compute $\log \widetilde{\Lambda}(X)$, the (generalized) log-likelihood ratio statistic associated to the above test.

4. Propose a test that is asymptotically of level α.

Exercise 4 $X_1, \dots, X_n$ are iid from a distribution whose density is

$$f_a(x) = \frac{a}{x^{a+1}} \mathbb{1}_{[1,+\infty[}(x)$$

for some parameter $a > 0$.

1. Justify that the expectation of X_1 is finite if and only if $a > 1$.

2. Prove that the family of distributions $\{f_a, a > 0\}$ forms an exponential family.
   What is its canonical statistic?

3. Prove that the maximum likelihood estimator of the parameter a is $\widehat{a}_n = \frac{n}{\sum_{i=1}^n \log(X_i)}$

4. Let $Y = \log(X_1)$. Prove that its cdf is given by $F_Y(t) = (1 - e^{-at})$ for all $t > 0$.

We recognize an exponential distribution of parameter a, for which $\mathbb{E}[Y] = \frac{1}{a}$ and $\text{Var}[Y] = \frac{1}{a^2}$.

5. Using the Delta method, find the asymptotic distribution of $\widehat{a}_n$.

6. Prove that this estimator is asymptotically efficient.

7. Construct an asymptotic test of level α for $\mathcal{H}_0 : (a \leq 1)$ versus $\mathcal{H}_1 : (a > 1)$.