Exam : October 2019.

duration : 3h

Only authorized document : one handwritten A4 sheet.
Any other material (exercise sheets, lecture notes, electronic devices) is forbidden.

1 Estimation of a default rate

In the context of a graphics card market research, we would like to know the proportion of defects among the cards produced by a rival company. We suppose that the total quantity produced each year is equal to an unknown constant number $k \in \mathbb{N}$. For each year $i \in \{1 \dots n\}$, we observe the number $X_i \in \mathbb{N}$ of defective cards produced in the year. The probability for a given card to be defective is unknown and denoted by p. Defectiveness or not of the cards produced during a year are independent. Observations X_i are independent. The parameter of the model is $\theta = (k, p)$.

1. What is the the distribution (law) of the X_is for a fixed θ ? From now on, we denote by X a random variable with the same distribution as the X_is.
2. Using the method of moments, based on the two first moments $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$, propose an estimator $\widehat{\theta} = (\widehat{k}, \widehat{p})$, as a function of
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
and
$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2,$$
assuming that $\bar{X} > \widehat{\sigma^2}$.
3. What can happen when the empirical variance of the X_is is of the same order as their mean value ?(answer in one sentence)

2 Estimation of the variance of a Gaussian sample

Consider an independent Gaussian sample $X_i, i = 1, \dots, n$ with mean μ and variance σ^2 both unknown. Consider an estimator of σ^2 given by
$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. In this exercise, you may use the properties of the Gaussians provided in the appendix.

1. Show that S is an unbiased estimator of σ^2.
Hint : write S as a function of the variables $Z_i = X_i - \mu$ then expand the expression.
2. Is the estimator S efficient in the sense of Cramer-Rao as an estimator of σ^2 ? For the answer, we will admit that for all (μ, σ^2),
$$Var(S) = \frac{2\sigma^2}{n-1}.$$
3. Consider the class of estimators of the variance : $\mathcal{S} = \{bS, b > 0\}$. Show that there exists an estimator with uniformly minimal quadratic risk in this class which means that for a certain b_0, for all θ, for all b, $R(b, \theta) \geq R(b_0, \theta)$, where $R(b, \theta)$ is the quadratic risk of the estimator bS, when the data are generated with the parameter θ.
4. Is the estimator $b_0 S$ efficient in the sense of Cramer-Rao ?

3 Poll for an election

During a presidential election, two candidates A and B are competing in the second ballot. To simplify the exercise, we consider that the outcome of each vote X_i ($i \in \{1, \dots, N\}$) follows a Bernoulli law with an unknown parameter $\theta \in [0, 1]$ and that there are no blank votes. The votes are independent. By convention, $X_i = 1$ if the elector i votes for A and $X_i = 0$ if the elector i votes for B. Hence

$$\mathbb{P}_\theta(X_i = 1) = \theta = 1 - \mathbb{P}_\theta(X_i = 0).$$

We denote by $V = \frac{1}{N} \sum_{i=1}^{N} X_i$ the outcome of the vote. Therefore, the candidate A wins the election if $V > 0.5$. We conduct a poll before the election : we observe independent $Y_i, i \in \{1, \dots, n\}$ which are independent from X_i and with the same distribution as X_i. We denote by $S = \frac{1}{n} \sum_{i=1}^{n} Y_i$ the result of the poll. In this section, we will take $N = 49 \times 10^6$ and $n = 10^4$.

A. Limit value of θ for which the outcome of the election is almost certain.

1. Using a Gaussian approximation given by the central limit theorem to model the law of V, give the value of the deviation ϵ such that

$$\max_{\theta \in [0,1]} \mathbb{P}_\theta(V > \theta + \epsilon) = 10^{-3}.$$

   Use the fact that $\max_{\theta \in [0,1]} \theta(1-\theta) = 1/4$. Give the result as a function of N and the quantile of the standard normal distribution, then an approximate value at the precision 10^{-5} using the probabilities and quantiles tables provided in the appendix.

2. Using the result of the previous question, give a limit value $\theta_0 \leq 0.5$ as large as possible (at precision 10^{-5}) such that

$$\forall \theta \leq \theta_0, \quad \mathbb{P}_\theta(\text{ A wins }) \leq 10^{-3}.$$

   Give the result as a function of ϵ then give a numerical value at precision 10^{-5}.

B. Classical analysis of the poll

The outcome of the poll is $S = 0.49$. We wonder if the candidate A has still a chance to be elected. Suppose that the Gaussian approximation is still valid for the chosen sample size $n = 10^4$.

3. At which level of confidence can you reject the null hypothesis $\theta \geq \theta_0$? You may choose the upper confidence bounds or the tests to answer (in the latter, consider a unilateral test).
4. Summarize in one sentence the conclusion from the poll, in terms of the probabilities of A to be elected and the confidence level.

C. Bayesian approach

We adopt now the Bayesian point of view on the poll problem. Consider the prior $\pi(\theta) = \mathcal{B}eta(1, 1)$ on $[0, 1]$. Recall that the density of the Beta distribution with parameters (a, b) on $[0, 1]$ is :

$$f_{a,b}(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}.$$

5. In the case of the poll of section 1.B ($n = 10^4, S = s = 0.49$), what is the posterior probability that $\theta \geq 0.5$? (Give the details of the computations). Write the result using the cumulative distribution function of a suitable Beta law, then give the numerical value using the tables in the appendix.

D. Neyman-Pearson approach under a double constraint (level and power)

We propose to draw a new poll sample $Z_1, \dots, Z_{n'}$ of size n' (which we will determine in the following), independent from the previous sample.

6. Consider θ_0 as obtained in section (A) and an arbitrary $\theta_1 < \theta_0$. Construct a unifomly most powerful (U.M.P.) test at level $\alpha = 0.001$ in the Neyman-Pearson setting for the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ with the data $Z_i, i \leq n'$. We will **not** use the Gaussian approximation for this. Show that this test can be written as follows :

$$\delta(Z_1, \dots, Z_{n'}) = \mathbb{1}\{S' < c\}$$

   with $S' = \frac{1}{n'} \sum_1^{n'} Z_i$, for a threshold c (you do not need to compute c for the moment).

7. Determine c as a function of n' and a quantile of the standard normal distribution (use the approximation $\theta_0(1 - \theta_0) \approx 1/4$ assuming that the Gaussian approximation is valid for this sample size).
8. Show that the test above is also a test at level α for the hypothesis $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$.
9. For which limit value of θ is the type II error (probability to wrongly accept H_0) maximal ? (justify your answer). What is the corresponding value of the type II error ?
10. For the test δ constructed above, determine the minimal sample size n' for which the type II error for $\theta = 0.49$ is smaller than 10^{-3} (use the approximations $\sqrt{0.49 \times 0.51} \approx 1/2$ and $\theta_0 \approx 1/2$)

Appendix

Appendix 1 : Properties of Gaussian random variables

1. **Moments :** If $U \sim \mathcal{N}(0,1)$, then $\mathbb{E}(X^{2p+1}) = 0$ and $\mathbb{E}(X^{2p}) = (2p-1)(2p-2)\cdots \times 3 \times 1 = \prod_{j=1}^{p}(2(p-j)+1)$.
2. **Fisher information :** Setting $\theta = (\mu, \sigma^2)$, the Fisher information of a Gaussian random variable $\mathcal{N}(\mu, \sigma^2)$ is
$$I(\theta) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}.$$

Some values of the cumulative distribution function F of the standard normal distribution and its inverse $F^{-1}(\alpha) = \{x : F(x) = \alpha\}$ are given below.

level α	0.950	0.975	0.990	0.999
quantile $F^{-1}(\alpha)$	1.645	1.960	2.326	3.090

Table 1 – Table of quantiles of the standard normal distribution

x	-4.00000	-3.00000	-2.00000	-1.96000	-1.64000
F(x)	0.00003	0.00135	0.02275	0.02500	0.05050

Table 2 – Cumulative distribution function of the standard normal distribution

Appendix 2 : Cumulative distribution functions of some Beta distributions

a	3900.000	4901.000	5000.000	5100.000
b	4100.000	5101.000	5301.000	5401.000
$F_{a,b}(0.5)$	0.987	0.977	0.998	0.998

Table 3 – Cumulative distribution functions at point $\theta = 0.5$ of different distributions Beta(a,b)