EPFL, Spring 2023
Guillaume Obozinski

EE-209 Eléments de Statistiques pour les Data Sciences

Feuille d'exercices 7

Exercise 7.1 A pivot to build a confidence intervals for S^2

In this exercise, we consider n independent and identically distributed random variables $X_i \sim \mathcal{N}(\mu, \sigma^2)$. The goal of this exercise is to obtain a confidence interval for σ^2. μ is not assumed to be known.

(a) Under the above assumptions, can you think of a pivot combining $S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$, (where $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$) and σ^2 ?

(b) Let us denote by $\chi^2_{m,\beta}$ the quantile of level $\beta \in (0,1)$ of Z, a χ^2 random variable with m degrees of freedom. Using the definition of quantiles, give the probability of the event

$$\mathcal{E} := \left\{\chi^2_{m,\alpha/2} \leq Z \leq \chi^2_{m,1-\alpha/2}\right\}.$$

(c) Deduce a confidence interval for σ^2 with confidence level $1-\alpha$.

Exercise 7.2 Overbooking

With your newly acquired statistical expertise, you are now ready to help AirEPFL, an airline company already encountered in Exercise 1 of Week 4.

Remember that AirEPFL charters a plane with 328 seats and decides to accept n reservations. The probability that a traveler with a reservation will show up at the airport is 0.8 (independently of other travelers). We denote by S_n the random variable that counts the number of ticketed customers who show up for boarding.

(a) For any $i = 1...n$, we denote by X_i the random variable equal to 1 if the i-th customer with a ticket shows up for boarding and 0 otherwise. What is the distribution of X_i? Can we consider that the X_is are independent?

(b) Express S_n as a function of the X_is and feduce the distribution of S_n.

(c) Using the central limit theorem, specify the asymptotic distribution of $\frac{S_n - np}{\sqrt{np(1-p)}}$.

(d) Using the approximation provided by the CLT show that the probability that all the passengers showing up at the airport can board the plane can be approximated by the probability that a standard normal variable exceeds a threshold to specify.

(e) Deduce what is the maximum number of reservations that AirEPFL can make such that each person with a reservation can board the plane with a probability greater than 0.98? (You can use that the quantile of level 0.02 of a standard normal Gaussian is $z_{0.02} = -2.054$).

Exercise 7.3 Benchmark study in research.

A scientist submits a research paper where he claims that his new algorithm A_1 to detect frauds is better that the method considered so far as the state of the art in the field A_2. To evaluate and compare fraud detection algorithms, the research community is using a benchmark dataset and requires to compute the accuracy, which is defined as the proportion of entries in the dataset for which the algorithm predicts the correct label (namely if there is a fraud or not).

The dataset contains $n = 993$ entries. A_1 has an estimated accuracy of $\widehat{p}_1 = 96.7777\%$ while A_2 has an estimated accuracy $\widehat{p}_2 = 96.2739\%$.

The scientist claim is in his opinion supported by the fact that $\widehat{p}_1$ is larger than $\widehat{p}_2$ by more than 0.5%.

One can consider that A_1 (resp. A_2) predicts for each new input the correct label with probability p_1 (resp. with probability p_2), which are estimated on the benchmark dataset by $\widehat{p}_1$ and $\widehat{p}_2$.

(a) Compute confidence intervals for p_1 and p_2 with confidence level $1 - \alpha = 95\%$, first keeping two significant digits of the margin of error, and then keeping a single significant digit of the margin of error.

(b) Conclude on the claim of the researcher.