EPFL, Spring 2023
Guillaume Obozinski

EE-209 Eléments de Statistiques pour les Data Sciences

Feuille d'exercices 4

Exercise 4.1 Overbooking

In the language of civil aviation, overbooking is the action of reserving more seats than there are available on an aircraft. All airlines overbook their planes in order to take advantage of reserved but unused seats. AirEPFL charters a plane with 328 seats and decides to accept 400 reservations. The probability that a traveler with a reservation will show up at the airport is 0.8 (independently of other travelers). Let S be the random variable that counts the number of ticketed customers who show up for boarding.

For convenience, we provide the following table of values of the cdf Φ of a standard normal variable.

x	-3	-2	-1	0	1	2	3
$\Phi(x)$	0.001	0.022	0.158	0.5	0.841	0.979	0.999

(a) What is the distribution of S?

(b) How can we write the event $\mathcal{E}$ ="the number of passengers unable to board is greater than or equal to k" as an inequality on S? How can we calculate its probability?

(c) Compute $\mathbb{E}[S]$ and $\text{Var}(S)$.

(d) Using the central limit theorem, propose a Gaussian approximation for the distribution of S.

(e) Express an approximate value of the probability of this event using a reduced centered normal random variable. More precisely, write $\mathbb{P}(\mathcal{E}) \approx \mathbb{P}(Z \geq z)$ for Z standard normal variable related to S and z a value that depends on k, on the total number of sold tickets and the number of seats in the plane.

(f) Estimate the probability that all passengers present at the boarding gate will be able to board the plane.

Exercise 4.2 Estimation of the size of a bird population

To estimate the size m of a bird population, we capture respectively $X_1, \ldots, X_n$ birds of a given species on each of n successive days and release them in the evening. We suppose that each bird has the same unknown probability p to be captured each day and this independently of the other birds and independently of whether or not it was captured on previous days.

(a) What distribution does X_i follow? What is the distribution of $(X_1, \ldots, X_n)$?

(b) Compute $\mathbb{E}[X_1]$ and $\mathbb{E}[X_1^2]$ and deduce moment estimators $\widehat{m}$ and $\widehat{p}$ for m and p.

(c) The protocol was used for $n = 20$ days and the following number of birds were caught: $159, 154, 163, 159, 162, 160, 148, 147, 163, 161, 156, 162, 147, 154, 186, 151, 145, 153, 154, 181$. After calculation this leads to moment estimates: $\overline{x} = 158.25$ and $\overline{x^2} = 25145.35$. Deduce from these an estimate of m.

Exercise 4.3 Empirical variance

We have seen that there are two ways to express the variance of a random variable X, namely:

$$\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

The goal of this exercise is to prove the counterpart of this result for the so-called empirical variance given by

$$v_n := \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2,$$

where $x_1, \ldots, x_n$ is a collection of observed values and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is their empirical mean.

(a) Express the empirical variance v_n as a function of $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

Exercise 4.4 Higgs boson

The Higgs boson is an elementary particle produced by the quantum excitation of the Higgs field. To produce Higgs bosons, two beams of particles are accelerated to very high energies and allowed to collide within a particle detector. Occasionally, although rarely, a Higgs boson will be created fleetingly as part of the collision byproducts.

Particle physicists measure the number of events X given by their detector. In the experiment, the number of events reported by the detector can be due to a true signal or to background noise. The number of true events and the number of events due to background noise can be both modeled by a Poisson distribution with respective parameters θ_S and θ_B. We end up with

$$X = X_S + X_B,$$

where the random variables X_S and X_B are independent and where X_S (resp. X_B) is distributed as a Poisson random variable with parameter $\theta_S > 0$ (resp. $\theta_B > 0$).

(a) What is the distribution of the random variable X?

(b) Physicists expect $\theta_B = 1.8$ background events. They observe $x_{obs} = 5$ events during their experiment.

To announce to the world that they have been able to detect the Higgs boson, the physicists need to be sure that their observation is very unlikely if one assumes that there is no signal (i.e. if $\theta_S = 0$). More precisely, they need to check that

$$\mathbb{P}_0(X \geq x_{obs}) \leq 10^{-4}, \tag{1}$$

where $\mathbb{P}_0$ is the distribution of X when $\theta_S = 0$ and $\theta_B = 1.8$.

Compute $\mathbb{P}_0(X \geq x_{obs})$. Deduce if the physicists can make public their discovery (we say in that case that the result of the experiment is *statistically significant*).