

# Statistical Inference

**ESSEC**

Olga Klopp

Home work 1

**This Homework is to be done by teams of two students. Write your names, homework number, course title and date. A script of your code in pdf or html should be submitted on the Moodle.**

**Deadline: March 15th**

**1. Simulation, Asymptotic behavior (4pts).**

For  $n = 10, 100, 1000$ , answer the following questions.

- (a) In R, simulate a  $n$ -sample of from the Poisson distribution with the rate 5. Represent its histogram, and the Q-Q plot comparing its quantiles to the theoretical quantiles of Poisson distribution.
- (b) Simulate 100 samples in the same way. For each sample, compute the empirical mean. Represent the histogram of the 100 empirical means. What distribution (what type of distribution, and what parameters) are we expecting the histogram to be close to?
- (c) What do you observe when  $n$  increases? By what theoretical result is this justified?

**2. Statistical approach and model (4pts).**

Devore, in his book *Probability and statistics* writes the following:

‘ An article in the New York times (Jan.27, 1987) reported that heart attack risk could be reduced by taking aspirin. This conclusion was based on a designed experiment involving both a control group of individuals that took a placebo (...) and a treatment group that took aspirin (...). Of the 11,034 individuals in the control group, 189 subsequently experienced heart attacks, whereas only 104 of the 11,037 in the aspirin group had heart attack. The incidence rate of heart attacks in the treatment group was only about half that in the control group. One possible explanation for this result is chance variation—that aspirin really doesn’t have the desired effect and the observed difference is just typical variation in the same way that tossing two identical coins would usually produce different numbers of heads. However, in this case, inferential methods suggest that chance variation by itself cannot adequately explain the magnitude of the observed difference.’

Can you formalize statistically this description?

**3. Algorithm for simulation (4pts).**

Let  $F : (a, b) \rightarrow (0, 1)$  a one to one cdf with  $-\infty \leq a \leq b \leq +\infty$ , and  $U$  a random variable uniformly distributed on  $(0, 1)$ , that is  $U \sim \mathcal{U}(0, 1)$ .

- (a) What is the distribution of  $F^{-1}(U)$ , where  $F^{-1}$  is the inverse function of  $F$ ?
- (b) Propose an algorithm to sample from the Exponential distribution  $\text{Exp}(\lambda)$  using a sample from the uniform distribution  $\mathcal{U}(0, 1)$ .
- (c) Implement this algorithm in R, and use it to generate a 1000 -sample distributed according to  $\text{Exp}(\lambda)$ . Compare the distribution of the sample to  $\text{Exp}(\lambda)$  using a graph.
- (d) Propose an algorithm to sample from a Cauchy distribution using a sample of the uniform distribution  $U(0, 1)$ . The density function of a Cauchy distribution with parameters  $x_0 \in \mathbb{R}$  and  $\gamma > 0$  is:

$$f(x) = \frac{1}{\pi\gamma} \frac{\gamma^2}{(x - x_0)^2 + \gamma^2}.$$

#### 4. Estimation of a agricultural area (4pts).

A farmer has a square field and wants to estimate its area. His friend, a statistician told him that when he measures one side of the field, the experimental error is a random variable  $X$  distributed according to a centered normal distribution with variance  $\sigma^2$ . The first measurement of the side is  $x_1 = 510$  meters. From this first measurement, the farmer deduces that the area  $s_1 = 26.01$  hectares. He measures the side a second time and finds  $x_2 = 490$  meters and a value of the surface  $s_2 = 24.01$  hectares. He abandons his measurements and thinks: Which area estimator should he choose :  $s_1$ ,  $s_2$ , or another estimator combining the two measures? For example :

$$s_3 = x_1 x_2 = 24.99;$$

$$s_4 = \frac{s_1 + s_2}{2} = 25.01;$$

$$s_5 = \left( \frac{x_1 + x_2}{2} \right)^2 = 25.$$

Should he continue his measurements until he finds two identical results, or intelligently combine the  $n$  measures to construct an estimator similar to  $s_4$  or  $s_5$  (generalized to  $n$  measures)? This exercise tries to answer this problem.

- (a) Specify the statistical model. Is the model identifiable?
- (b) Write the quantity to be estimated (the area of the field) in the form of a regular function  $g$  applied to  $X$ . Suggest an estimator for the area of the field.
- (c) Generalize this estimator to the case of  $n$  measurements  $x_1, \dots, x_n$ . Study the asymptotic properties of this estimator: is it strongly consistent? Asymptotically normal? What is its asymptotic variance?

#### 5. Descriptive statistics (4pts).

In this exercise, we try to describe a dataset in R using descriptive statistics and graphs, the dataset `cider.csv` is available on Moodle. This dataset is the result of an experiment focusing on the balance of flavors in ciders and more specifically on the relationship between the sweet, sour, bitter and astringent flavors, depending on the type of cider (raw, half dry or soft). For this, we have the evaluation of these four flavors, provided by a sensory jury (average score of 24 judges, scores from 1 to 10) for each of the 90 ciders divided between brut, semi-dry and sweet

- (a) Load the dataset `cider.csv` in R.
- (b) Represent the distribution of ciders between brut, semi-dry and sweet using a barplot.
- (c) Represent the distribution of the sweet flavor using a boxplot, for each type of cider.
- (d) Represent the sweet flavor according to the bitter flavor and add the regression line, for each type of cider. Hint: use the functions `geom_point` and `geom_smooth`.

(e) Comment your results.