STNUM 2025-26
Exercises Weeks 1–2
(Homework 1)

Exercises marked with a star (*) are **homework** and must be submitted by **Friday, December 19, at 19h00**. Please submit the homework in groups of 2 or 3 (homework groups must be within your small class groups). See Educnet for further instructions.

- It is recommended that you use Python for the coding portions; other languages may be accepted depending on your group instructor.
- You are welcome to use any standard tools (e.g., functions in `numpy` or `scipy`, including those for statistics), but you are responsible for knowing what they do.
- The more qualitative questions (e.g., "discuss…" or "comment…") could have multiple possible good answers! Any thoughtful response is acceptable.

Summary statistics and consistency

1. (*) Three data files (`hw1data1.csv`, `hw1data2.csv`, and `hw1data3.csv`) are provided on Educnet. Each of these contains 10^5 numbers. For each file, we will compute summary statistics on various "sample" sizes from the full "population" of 10^5 data.

(a) Write code that randomly samples n points from the dataset without replacement (e.g., this `numpy` function may be useful).

(b) For each dataset (file), for (independent) samples of sizes $n = 10^k$ for each $k \in \{1, 2, 3, 4, 5\}$, calculate and plot together (vs. n) the sample **mean**, **median**, and **variance** (it may be useful to make one or both plot axes logarithmic). You should, in the end, have three plots (one for each file) with three lines each. Of course, when $n = 10^5$, we are using the entire set.

(c) What differences do you see between the three plots (files)? For each dataset, which estimators appear to be *consistent* (i.e., converging to some true mean/median/variance)?

(d) Plot the histograms of the three (full) datasets and comment on the differences in distribution and what impact we expect on the summary statistics we computed. To get an easily readable histogram, you may need to play around with the plot settings (number of bins, log scale, etc.).

Distributions and estimators

2. (*) The gamma distribution $\gamma(p, \theta)$ for parameters $p, \theta > 0$ is a continuous distribution with density

$$f_{p,\theta}(x) = \begin{cases} \frac{\theta^p}{\Gamma(p)} e^{-\theta x} x^{p-1} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0, \end{cases}$$

where Γ is the gamma function. This is a versatile distribution that includes others such as the exponential distribution and the χ^2 distributions as special cases. We have, if $X \sim \gamma(p, \theta)$,

$$\mathbf{E}\, X = \frac{p}{\theta}, \qquad \text{and} \qquad \operatorname{var}(X) = \frac{p}{\theta^2}.$$

(a) Assuming $p > 0$ (the "shape" parameter) is known, what is the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ for θ given i.i.d. samples $X_1, \ldots, X_n \geq 0$? Hint: your answer should be a function of $\overline{X}$, the sample mean.

(b) For large n, what do we expect to be the (approximate) distribution of $\hat{\theta}_{\text{MLE}}$? Express your answer in terms of p, θ, and n.

Hint: the central limit theorem gives you an asymptotic distribution for $\overline{X}$; you may then use the fact that, for any differentiable function g with derivative g', $g(a+h)$ is well-approximated by the linear function $g(a) + g'(a)h$ for **small** values of $h \in \mathbf{R}$.

3. The Poisson distribution Poisson(λ) with parameter $\lambda \geq 0$ is a discrete distribution with probability mass function

$$\mathbf{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for all integers } k \geq 0.$$

(a) Given independent data $X_1, \ldots, X_n \sim \text{Poisson}(\lambda_*)$, what is the maximum likelihood estimator $\hat{\lambda}$ of λ?

(b) Show that, if X and Y are **independent** Poisson-distributed random variables (say, with parameters λ_X and λ_Y), then their sum also has a Poisson distribution. You may use the formula

$$\mathbf{P}(X + Y = k) = \sum_{\substack{\ell, m \\ \ell + m = k}} \mathbf{P}(X = \ell, Y = m).$$

The binomial formula may also be useful.

(c) The Poisson distribution is often used to model counts of independent discrete events (customers entering a supermarket, photons hitting a detector... See also problem 7.) Comment on why this fits well with the additivity discussed in part (b).

Simulation and error distributions

Collecting real data takes a long time, and storing it in an intelligent way and then writing code to import (and possibly further clean) it also takes effort. However, to practice coding estimators and to explore numerically the properties of different distributions and estimators, there is a far easier option: your computer's random number generator (RNG)! Most software libraries for numerics (and especially for statistics) have built-in (pseudo[1]-)random number generation with a wide variety of standard distributions. In Python, the standard module is `numpy.random`. More probability tools and many more exotic distributions can also be found in `scipy.stats`.

4. **(*)** Consider estimating the variance $\sigma^2 \geq 0$ of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ from i.i.d. samples $X_1, \ldots, X_n$ (for simplicity, we assume that μ is known).

The maximum likelihood estimate of σ^2 is simply the sample variance (with known μ):

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Because the normal likelihood is smooth in σ^2, we know that, as $n \to \infty$, the error distribution of the MLE converges to a normal distribution in the sense that

$$\frac{\hat{\sigma}^2_{\text{MLE}} - \sigma^2}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, v(\mu, \sigma^2)),$$

where v is some function of the true distribution parameters.

(a) Write code that generates n i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$ for $\mu = 10$ and $\sigma^2 = 1.5$ and compute the sample variance (recall that μ is known).

[1] Most libraries default to producing "random-looking" but in fact repeatable *pseudorandom* numbers. We will not worry about this distinction here, but, if you are getting the same results again and again, this is something to consider (probably the random seed is being reset to the same value each time).

(b) For the choices $n = 4$, $n = 10$, and $n = 100$, calculate $\hat{\sigma}^2_{\text{MLE}}$ for 400 separate trials and plot a histogram of the values of the normalized error $\frac{\hat{\sigma}^2_{\text{MLE}} - \sigma^2}{\sqrt{n}}$. You should have 3 histograms, each with 400 values.

(c) Comment on the appearance of the histograms and how this matches our expectation of asymptotic normality of the MLE.

5. Consider estimating the (upper) bound $a > 0$ of the uniform distribution $\mathcal{U}([0,a])$ with density

$$f_a(x) = \begin{cases} \frac{1}{a} & \text{if } 0 \leq x \leq a \\ 0 & \text{otherwise} \end{cases}$$

from independent samples $X_1, \ldots, X_n \geq 0$.

(a) What is the maximum likelihood estimator $\hat{a}_{\text{MLE}}$ of a?

(b) Similarly to problem 4, for a fixed value of a (say, $a = 5$), write code to generate n i.i.d. samples from $\mathcal{U}(0, [0,a])$ and compute $\hat{a}_{\text{MLE}}$. For various values of n, plot a histogram of $\frac{\hat{a}_{\text{MLE}} - a}{\sqrt{n}}$ over 400 trials. Include in your submission a histogram for $n = 100$. Does the normalized error appear normally distributed for large n? Discuss this in light of what we know about asymptotic normality of the MLE.

The sample mean $\overline{X}$ is not a good estimator of a, as, if $X \sim \mathcal{U}([0,a])$, then its mean is $\mathbf{E}\, X = \frac{a}{2}$. Therefore, $\mathbf{E}\, \overline{X} = \frac{a}{2}$. However, this suggests that an unbiased estimator is $\hat{a}_2 = 2\overline{X}$.

(c) Repeat part (b) for this estimator (this is not an MLE, but the central limit theorem still applies).

(d) Which estimator ($\hat{a}_{\text{MLE}}$ or $\hat{a}_2$) appears to have less error? You can judge this qualitatively from the histograms, but you can also calculate it quantitatively: for each estimator (with 400 independent trials with $n = 100$), compute the empirical mean squared error, that is,

$$\widehat{\text{MSE}} = \frac{1}{400} \sum_{k=1}^{400} (\hat{a}(k) - a)^2,$$

where $\hat{a}(k)$ is the estimator from the kth trial.

Confidence intervals 1

6. **(*)** Suppose that, as part of a construction project, we need to inspect a particular type of critical bolts (boulons) provided by a supplier to verify that they meet the required strength standards. You test 1000 bolts (that takes a while...) and discover that 28 of those bolts failed to meet the standards. Estimate an *upper* bound on the defect rate p with 95% confidence (i.e., this is a one-sided confidence interval). You may use the normal approximation from the central limit theorem.

7. The department Seine-et-Marne has 4420 km of roads[2]; suppose that we want to estimate the rate of pothole formation in this road network. Specifically, we want to estimate the rate r of average potholes formed **per month, per kilometer**.

A common model for such a problem is a Poisson distribution (see problem 3), that is, on L kilometers of road, the number of potholes that form follows the distribution Poisson(Lr).

To estimate r, you choose 20km of roads that you then inspect carefully at the beginning and end of a given month. You observe that, on these roads, the total number of potholes that form in this month is 65.

(a) Give a 90% (symmetric) confidence interval for r. You may use the normal approximation from the central limit theorem. Recall that, if $X \sim \text{Poisson}(\lambda)$, $\text{var}(X) = \lambda$ (hence the standard deviation is $\sqrt{\lambda}$).

(b) Comment: What are some limitations of our model? What are some possible sources of bias in our measurements?

[2] Source: https://www.seine-et-marne.fr/fr/infos-routes-77