M1 Master's Degree in Cognitive Science 2024–2025 | October 10^{th}
TA: Hugo Ninou (hugoninou@gmail.com)

MATHEMATICAL TOOLS FOR DATA SCIENCE, MACHINE LEARNING AND STATISTICAL MODELLING

TD1: STATISTICS

Click here to access the notebook for this tutorial.

You should copy the notebook from Google colab (go to file->save a copy in drive) in order to be able to save your changes.

1 Exercise 1

1.1 Sampling from a uniform distribution

Let $(X_1, X_2, \ldots, X_n)$ be i.i.d. random variables with $(X_1, X_2, \ldots, X_n) \sim \mathcal{U}(0, \theta)$.

1. Given a sampling $(X_1, X_2, \ldots, X_n)$, what is the range of possible values for θ, the parameter of the uniform distribution ?
2. What is the likelihood of generating a sample $(X_1, X_2, \ldots, X_n)$ from this distribution ? For what value $\hat{\theta}$ of θ is this likelihood maximal ? $\hat{\theta}$ is called the maximum likelihood estimator (MLE) of θ.
3. Give an expression of the cumulative density function (cdf) and the probability density function (pdf) of $\hat{\theta}$ as a function of θ
4. Evaluate the bias, standard error and mean-squared error of $\hat{\theta}$. Is this estimator consistent ?
5. Derive an unbiased estimator $\tilde{\theta}$ for θ.

Note: For a discrete uniform distribution, this estimator becomes $\tilde{\theta} - 1$.

1.2 Antique coin collection

You are a historian analyzing ancient coins from a recently discovered collection. Only a limited number of coins were produced, and each coin was minted with its serial number. You've acquired a random sample from a larger collection, and your task is to estimate how many coins in total were originally minted.
We will derive the probability $p(N = n|M = m, K = k)$ that has for parameters (N, M, K) respectively : the number of coins originally produced, the highest serial number observed and the number of coins collected.

1. Using Bayes' formula, give an expression of $p(n|m, k)$ as a function of $p(m, n|k)$
2. Give an expression of $p(m|n, k)$ as a function of n, m, k.

3. Give an expression of $p(m|k)$ as a function of $p(m|n,k)$

4. What's the probability $p(n|m,k)$ as a function of n, m, k ? You'll need this equation

$$\sum_{n=m}^{+\infty} \frac{1}{\binom{n}{k}} = \frac{k}{k-1} \frac{1}{\binom{m-1}{k-1}} \text{ for } k \geq 2$$

5. For what value of n is $p(n|m,k)$ maximal ? How is this value called ?

6. What is the expected value μ of m ?

   The quantities found in the tutorial ($m = max(X_1, ..., X_k)$, $\tilde{\theta} = m(1 + 1/K) - 1$ and $\mu = (m-1)\frac{k-1}{k-2}$) are three estimators for the total number of coins originally produced. Your goal in the python notebook will be to sample these estimators and compare their empirical distributions for given parameters of the problem N and k, the number of collected coins.

7. First consider $N = 1000$ coins originally produced with serial numbers ranging from 1 to N and $k = 10$ collected coins. Write a python program that would sample the serial number of each collected coin using the function np.random.choice(). From these serial numbers, compute the corresponding values of m, $\tilde{\theta}$ and μ.

8. Repeat this experiment 10000 times and store the results in the empyy numpy arrays in the notebook. Fill in the code for plotting the empirical distributions and means of the estimators.

9. We would now like to vizualize how these empirical estimator distributions change with regard to the problem parameter k. Repeat the previous experiment for 20 values of k ranging from 3 to 200 that you will store in the array "k_list", store the results in the empty numpy arrays.

10. Plot the results, what do you see as k increases ? Based on these empirical distributions, what can you say about the consistency and bias of the three estimators ? Is there one we should prefer to the others?

2 Analyzing knowledge retrieval impairments associated with Alzheimer's disease (AD)

In this exercise, we are going to use semantic networks as a model of how memories are encoded. In a semantic network, concepts are represented by nodes and semantic similarity is represented by edges that connects pairs of nodes.
Among the most commonly used tasks to diagnose semantic memory impairment is the semantic fluency task, in which participants list as many items from a category (e.g. animals) as they can in a short period of time. In this exercise, we apply a random walk model of semantic memory retrieval to a longitudinal dataset (from UCSD Shiley-Marcos Alzheimer's Disease Research Center) of semantic fluency data from AD patients (PAD) and healthy controls (NC) in order to estimate semantic network representations of individuals and investigate mechanisms responsible for impaired performance due to AD. Our goal will be to retrieve the underlying semantic networks from the fluency lists and compare how their properties differ between AD and control patients.

- **Average shortest path length:** let $d(v_1, v_2)$ denote the shortest distance between nodes v_1 and v_2. The average shortest path length is the average distance $d(v_i, v_j)$ for $i \neq j$.

$$\text{aspl} = \frac{1}{n(n-1)} \sum_{i \neq j} d(v_i, v_j)$$

- **diameter:** The diameter of a graph is the greatest length of all of its shortest path lengths.

$$\text{diameter} = \max_{i \neq j}(d(v_i, v_j))$$

1. Compute the total number of patients, lists and items for both groups, comment the results.
2. Compute and store in a list the adjacency matrix for each of the patients of both groups.
3. For each of semantic network associated to a given patient, compute and store the average shortest path lengths, diameters, number of nodes and number of edges and plot their distributions and means. What difference do you see between the NC and PAD groups ?

   Participants with more fluency lists (and longer fluency lists) will typically have semantic networks that have more nodes and more edges. This is confounding because most network properties (such as diameter or average shortest path length) are affected by the number of nodes and edges in a network. This makes it difficult to draw inferences from a direct comparison of NC and PAD networks, as the networks vary in the amount of data used to generate them. To alleviate this problem, we are going to analyze each network by comparing it to its own set of mock networks generated by the following procedure: For each participant, we will generate a random permutation of each fluency list so that the order of the words in each list is arbitrary.
4. What kind of test seen in the lecture does this procedure refer to ?
5. Write a python code for randomly permuting elements of each fluency list thanks to the function np.random.randint()
6. Show the permuted lists for the first patient in the PAD group and compare it to the original data.
7. For each patient, compute and store the average shortest path lengths (ASPL), diameters, number of nodes and number of edges for 50 iterations of the permutation procedure.
8. For both groups, compare the mean of the difference between the average shortest path lengths (ASPL) of the real data and that of the mock networks.
9. We want to test whether this difference is significant, design a statistical test for the hypothesis H_0 :"ASPL(PAD) > ASPL(NC)". What is the associated p-value ?
10. How do you interpret these results as for the way memories are encoded in AD patients ?