CSDS 313/413: Introduction to Data Analysis Fall 2025

Homework 2: Statistical Inference

Task 1: German Tank Problem

Recall the German Tank (or the "Rocket Science") problem from Exercise #7 in the class. You know that the Nazi Army is assigning sequential integer IDs to their tanks. In other words, if they have M tanks, there is a tank with ID i for all $1 \leq i \leq M$. The allied forces capture n tanks and the IDs of these tanks are $1 \leq x_1 \leq x_2 \leq ... \leq x_n$. Our goal is to estimate M, i.e., the number of tanks that the Nazis have, based on the assumption that the tanks were captured uniformly at random from these M tanks.

(a) Let $\hat{M}_{MLE}$ be the maximum likelihood estimator for M. Compute $\hat{M}_{MLE}$.

(b) Let $\hat{M}_{MEAN} = 2(\sum_{i=1}^{n} x_i/n) - 1$ and $\hat{M}_{MVU} = x_n(\frac{n+1}{n}) - 1$ be two additional candidate estimators for M. Theoretically characterize the expected value of each of the three estimators ($\hat{M}_{MLE}$, $\hat{M}_{MEAN}$ $\hat{M}_{MVU}$) for a given M and N. Comment on the bias of each estimator.

(c) Using different values of M and n, simulate different instances of the German Tank Problem, visualize the mean and variance of these three estimators ($\hat{M}_{MLE}$, $\hat{M}_{MEAN}$, $\hat{M}_{MVU}$) as a function of M and n (i.e., fix M to, say 100, 1000, 10K etc., and for each M, plot the mean and variance of the estimators as a function of n - you can compute the mean and variance by repeating the simulation multiple times for each combination of M and n). Discuss the differences between the estimators in terms of their bias and variance.

Task 2: The Distribution of the Sample Mean

In this task, you will we gain insight about the Central Limit Theorem. Consider three random variables:

(a) A **Uniform distribution** on $[0.25, 1.25]$

(b) A **Power-Law distribution** (with parameter $\alpha = 3$)

Part A: Data Generation

- For each distribution, generate **replicates of data** with sizes 10, 100, and 1000.
- Each replicate should consist of 100 samples.
- Example: For "10 replicates," generate 10 separate datasets, each with 100 samples.
- For one replicate from each distribution, plot the distribution of the sample to ensure that it follows a uniform/power-law distribution.

Implementation notes:

- In **Python**, use `numpy.random.uniform` and `numpy.random.power`.
- In **R**, use `runif()`, `rplcon()` (for Power, with shape1 = α, shape2 = 2.5).

Part B: Sample Means and Visualization

- Compute the sample means for each set of replicates. You should have 6 of them, 3 for each distribution with varying sample sizes.
- For each distribution and replicate size, we want to view and compare their distributions. Which plot do you think is appropriate? Plot the appropriate plot for each replicate, please plot them separately.
- Do not confuse the distribution of sample mean with the distribution of the sample (which you plotted in Part A).

Part C: Normal Approximation via CLT

- By the Central Limit Theorem, the distributions of sample means should approximate normal distributions. Determine the parameters (mean and standard deviation) of the approximating normal distributions for both power law and uniform cases.
- Overlay the probability density functions (PDFs) of these normal distributions on the corresponding plot from part B.

Part D: Distribution Comparison with KS-Test

- Compare the empirical distributions of sample means (from power law vs. uniform) using the Kolmogorov–Smirnov (KS) test.

- Discuss whether the KS-test supports the theoretical expectation that both sample mean distributions should approximate normality under the CLT. And investigate how the power of the KS-test changes as the number of sample means increases (10, 100, 1000).

Task 3: Hypothesis Testing

Car sales people usually claim that Japanese cars have better mileage than US cars. As a critical thinker, you know not to immediately believe what somebody who is trying to sell you something says. However, as a data scientist, you are curious to see whether this claim has any statistical basis.

You found some data collected in “cars.csv”. The first column contains a sample of measured mileage per hour for different US cars and the second column contains a sample of measured mileage per hour for different Japanese Cars. Answer the following questions:

(a) How many US and how many Japanese cars does the dataset contain?

(b) State the hypothesis you are testing to scrutinize the sales person’s claim.

(c) What is the statistic you would compute to test this hypothesis and its theoretical distribution?

(d) What is the value of your statistic and what is its p-value?

(e) What is your conclusion about the sales person’s claim?

Note: Be sure to check any conditions or assumptions of the chosen test before computing statistics and drawing your conclusion.