Introduction to Data Science — Assignment 0

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh** TA(s): محمدجواد رنجبر، محمد مهدی ابراهیم سلطانی، حمید سالمی

Deadline: Esfand 24th

Introduction

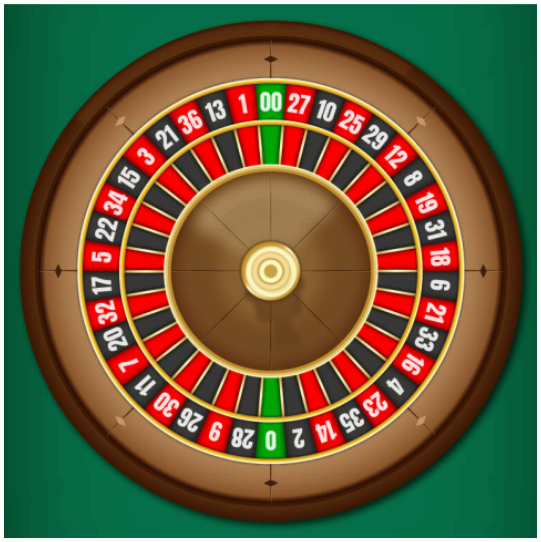
In this assignment, we will explore and implement key statistical concepts that are essential for analyzing data, making inferences, and drawing conclusions. These concepts play a critical role in research and real-world applications.

Task

Roulette Simulation and Profit Analysis

Roulette is a popular casino game played with a wheel that has numbered slots colored red, black, or green. In American roulette, the wheel has 38 slots: 18 red slots, 18 black slots, and 2 green slots labeled "0" and "00".

Players can place various types of bets, including betting on whether the outcome will be a red or black slot. In this exercise, we focus on a simple bet: betting on black.



If you place a bet on black and the outcome is indeed black, you win and double your money. However, if the outcome is red or green, you lose the amount you bet. For example, if you bet 1 dollar on black and win, you gain 1 dollar. If you lose, you forfeit your 1-dollar bet.

Because there are three colors, the chance of landing on the black isn't exactly $\frac{1}{2}$, it is less, specifically $\frac{18}{38} = \frac{9}{19}$.

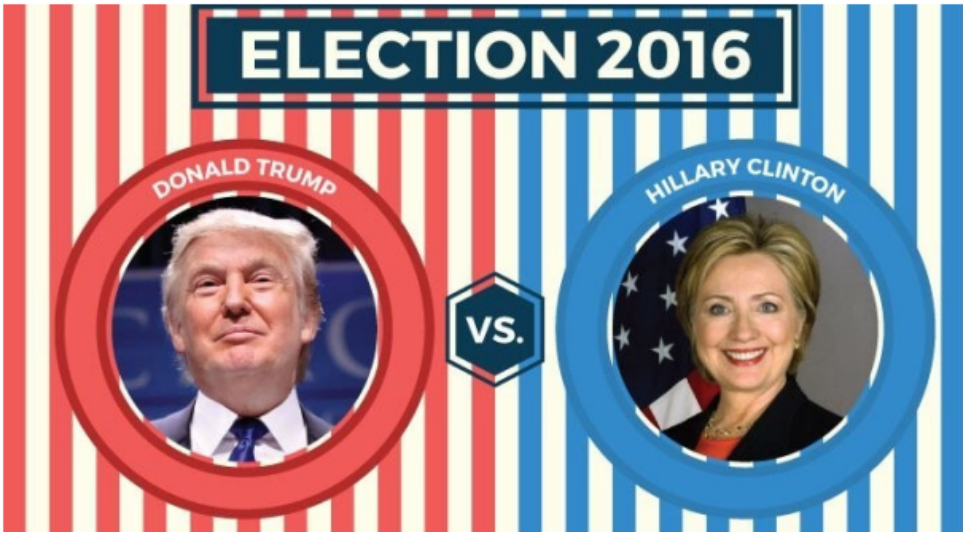
Consider the following tasks to simulate this game and analyze the expected outcomes of betting on black:

1. Write a function that simulates this game for N rounds, where each round consists of betting 1 dollar on black. The function should return your total earnings S_N after N rounds.
2. Use Monte Carlo simulation to study the distribution of total earnings S_N for N = 10, 25, 100, 1000. For each N, simulate 100,000 rounds and plot the distribution of total earnings. Analyze whether the distributions appear similar to a normal distribution and observe how the expected values and standard errors change with N.
3. Repeat the previous simulation but for the average winnings $\frac{S_N}{N}$ instead of S_N. For each N, plot the distribution of average winnings and examine the changes in expected values and standard errors with different values of N.
   ($N = 10, 25, 100, 1000$)
4. Calculate the theoretical expected values and standard errors of S_N for each N, and compare these theoretical values with your Monte Carlo simulation results. Report any differences between the theoretical and simulated values for each N.
5. Use the Central Limit Theorem (CLT) to approximate the probability that the casino loses money when you play $N = 25$ rounds, and verify this approximation using a Monte Carlo simulation.
6. Plot the probability that the casino loses money as a function of N for values N ranging from 25 to 1000. Discuss why casinos might encourage players to continue betting in light of these results.

Predicting the Outcome of the 2016 USA Presidential Election

In 2012, data scientists, including Nate Silver, accurately predicted the U.S. presidential election outcomes by aggregating data from multiple polls. By combining poll results, they provided more precise estimates than a single poll could achieve.

In this exercise, we aim to predict the result of the 2016 U.S. presidential election by analyzing polling data and aggregating results.



The data for this exercise is in a CSV file named `2016-general-election-trump-vs-clinton.csv`. Note that some rows may represent subgroups (e.g., voters affiliated with specific parties) and contain NaN values in the "Number of Observations" column. Exclude such rows from your calculations to avoid errors.

Now do the following tasks:

1. Let X_i be a random variable where:
   - $X_i = 1$ if the i-th voter supports the Democratic candidate.
   - $X_i = 0$ if the i-th voter supports the Republican candidate.

   With $i = 1, 2, \ldots, N$, the Central Limit Theorem (CLT) states that if N is large:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \hat{p} \approx N\left(p, \frac{\hat{p}(1-\hat{p})}{N}\right)$$

   where p is the true proportion of voters supporting the Democratic candidate. Based on the CLT result, derive and compute the 95% confidence interval (CI) for p.

2. Suppose the true population proportion $p = 0.47$. Perform a Monte Carlo simulation with $N = 30$ and 10^5 iterations to show that the CI derived in Question 1 captures the true proportion p approximately 95% of the time.
3. Load the data from the dataset into your coding workspace, and then make a data frame containing only the columns `Trump`, `Clinton`, `Pollster`, `Start Date`, `Number of Observations`, and `Mode`. Exclude any rows where the `Number of Observations` is missing.

4. Create a time-series plot of poll results showing support percentages for Trump and Clinton, using different colors for each candidate. Include a smooth trend line to visualize support trends over time.
5. Calculate the total number of voters observed by summing all poll observations in the dataset.
6. Calculate the estimated proportion of voters favoring Trump and Clinton. Display these estimates in a table.
7. Using the aggregated data, compute the 95% confidence intervals for Trump and Clinton support proportions.
8. For illustrative purposes, assume there are only two parties, and let p denote the proportion of voters supporting Clinton. Consequently, $1 - p$ represents the proportion supporting Trump. We define the spread as the difference in support between Clinton and Trump:

$$d = p - (1 - p) = 2p - 1$$

Using the aggregated poll data, we estimate p as $\hat{p}$. Therefore, the estimated spread d can be approximated as:

$$d \approx 2\hat{p} - 1$$

This also implies that the standard error for the spread is twice as large as the standard error for $\hat{p}$. So, our confidence interval for the spread d is:

$$CI\ for\ d = (2\hat{p} - 1) \pm 1.96 \times (2 \times SE_{\hat{p}})$$

where $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$ is the standard error of $\hat{p}$.

a) Calculate the 95% confidence interval for the spread d, using the formula provided above.

b) Conduct a hypothesis test to determine if the spread d is significantly different from zero by testing $H_0 : d = 0$ vs. $H_a : d \neq 0$. Provide the test statistic and p-value.

Drug Safety Test

The dataset comes from a randomized controlled drug trial conducted by a medical group and shared by Vanderbilt University Department of Biostatistics. The study evaluates drug safety by comparing a Drug and Placebo group while tracking adverse effects, vital signs, and lab measures .You will conduct hypothesis testing using t-tests to determine statistical differences between groups.

Columns Explanation:

- **`age`**: Age of the participant
- **`sex`**: Gender of the participant (`male` or `female`).
- **`trx`**: Treatment group:
  - `"Drug"` → Received the actual drug.
  - `"Placebo"` → Received a placebo (control group).
- **`week`**: Week number in the study.
- **`wbc`**: White blood cell count (WBC) measurement.
- **`rbc`**: Red blood cell count (RBC) measurement.
- **`adverse_effects`**: Presence of adverse effects (`Yes` or `No`).
- **`num_effects`**: Number of adverse effects experienced by the participant.

1. Load the `drug_safety.csv` into a Pandas DataFrame.
2. Remove samples that contain nan whenever it is needed.
3. Display basic statistics (e.g., mean, standard deviation) for numeric columns.
4. Group the dataset by `trx` (Drug vs. Placebo) and summarize key statistics for `wbc`, `rbc`, and `num_effects`.
5. Change **`adverse_effects`** column so that you can define **mean** for it.

6. For each metric below, determine if they differ significantly between the Drug and Placebo groups?

    a. mean white blood cell count

b. mean red blood cell count

c. mean num effects

d. mean adverse effect

- Formulate null and alternative hypotheses:
  - H_0: There is no significant difference
  - H_1: There is a significant difference
- perform an independent t-test.
- Interpret the p-value and state whether to reject or fail to reject the null hypothesis.
- If we set the p-value significant level to 0.05, which tests will fail? What about 0.1? What does this significant level mean?
- use `scipy` library for it and report:
  - What is the `alternative` argument and what did you choose for each metric, why?
  - What is the `equal_var` argument and what does it do?

Questions (10% Bonus)

An engineer is monitoring the pressure inside an oil pipeline. Due to varying flow rates and environmental conditions, the pressure in the pipeline fluctuates slightly with time. The true average pressure of the pipeline is unknown. Pressure measurements, $X_1, X_2, \ldots, X_n$ satisfy the following model:

$$X_i = \mu + \epsilon_i$$

where $\mu$is the unknown true average pressure, and ϵ_i represents random error. The errors are i.i.d. with mean 0 and unknown standard deviation σ. The pipeline's pressure is measured 100 times. If we construct an approximate 95% confidence interval for μ, this interval was constructed for one of the following purposes. Indicate which is correct and explain why:

1. To estimate the average of the 100 pressure measurements and give ourselves some room for error in the estimate.
2. To estimate the true average pressure of the pipeline and give ourselves some room for error in the estimate.
3. To provide a range in which 95 of the 100 pressure measurements are likely to have fallen.
4. To provide a range in which 95% of all possible pressure measurements are likely to fall.

Notes

- Upload your work as a zip file in this format on the website: DS_CA0_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- We will run your code during the project delivery, so make sure your results are reproducible.