Project:

Testing the Distribution of Leading Digits - Lottery Powerball Winning Numbers Dataset

---

There is a statistical phenomenon I wish for you to discover for yourselves.
Given a table of numbers that came from an actual experiment. Note there must be nearly infinitely many different numbers in this data set and not just a few that are repeated very many times.
The question is what is the distribution of the first digit on the very left of each number? Is it true that if you pick a sample of large enough size the probability that any digit from 1 to 9 will occur equally likely or not?
You are required to find data from an actual experiment that had a population.
*** If you use a random number generator you will get the wrong answer. It needs to be a data set from an actual experiment. ***
Your task is to determine whether or not the distribution between the leading digits is equal.
You are required to use the appropriate statistical methods to discern this.

Report Structure

1. **Introduction**
    - Problem description and project objectives
2. **Data Collection and Preparation**
    - Source of data, table of values, and assumptions.
3. **Hypothesis Formulation**
    - Null and alternative hypotheses for both tests.
4. **Statistical Methods**
    - Chi-Square Test and Z-Test methodologies with formulas.
5. **Results**
    - Chi-Square calculations, Z-Test results, and visualizations.
6. **Conclusion and Limitations**
    - Key findings, interpretation, and areas for improvement.

---

1. Problem Description and Objective

a) Problem

- Verify whether the distribution of leading digits (first digits) in a dataset is uniform.

b) Objective:

- Use the **Chi-Square Goodness-of-Fit Test** to test the uniformity of the first-digit distribution.
- Use the **Z-Test** to check if the sample mean matches the population mean.

---

2. Data Collection and Preparation

a) **Table of Collected Values**

- Use data from a **real experiment** (randomly generated numbers are **not allowed**).

  Lottery Powerball Winning Numbers Dataset

  https://www.kaggle.com/datasets/daniilkrasnoproshin/lottery-powerball-winning-numbers-dataset

- Ensure the dataset includes a **variety of numbers** and not repetitive patterns.

b) **Each Value Must Occur At Least 5 Times**

- Verify that all leading digits (1–9) occur frequently enough to ensure statistical reliability.

---

3. Hypothesis Formulation

(Chi-Square Test)

- **Null Hypothesis (H_0)**: The distribution of the leading digits is uniform (equal probability $\frac{1}{9}$ for each digit).
- **Alternative Hypothesis (H_1)**: The distribution of the leading digits is not uniform.

(Z-Test)

- **Null Hypothesis (H_0)**: The sample mean $\bar{X}$ is equal to the population mean $\bar{x}$.
- **Alternative Hypothesis (H_1)**: The sample mean $\bar{X}$ is not equal to the population mean $\bar{x}$.

---

4. Statistical Methods and Analysis

Chi-Square Goodness-of-Fit Test

1. **Prepare Frequency Table**:
   - Calculate **Observed Frequencies (O_i)** for digits 1 through 9.
   - Compute **Expected Frequencies (E_i)**:

$$E_i = N \cdot \frac{1}{9}, \quad \text{where } N \text{ is the total number of observations.}$$

2. **Calculate the Chi-Square Statistic**:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

3. **Degrees of Freedom**:

$$d.f. = k - 1 = 9 - 1 = 8$$

4. **Compare to Critical Value**: Use the **Chi-Square Table** at significance level $\alpha = 0.05$.

Z-Test for Point Estimate

1. **Collect All Needed Information**:
   - Values: $H_0, \bar{X} = \bar{x}, N, \sigma_{\bar{X}}, \sigma$.
2. **Determine Test Type**:
   - Left-tailed, right-tailed, or two-tailed test.
3. **Standardize the Data**: Calculate the Z-Score:

$$Z = \frac{\bar{X} - \bar{x}}{\sigma / \sqrt{N}}$$

4. **Compare to Critical Z-Value**: Use $N(0, 1)$ (standard normal distribution).

---

5. Results and Visualization

- **Chi-Square Test Results**:
  - Report χ^2-statistic, degrees of freedom, and p-value.
  - Compare to the critical value to determine whether H_0 is rejected.
- **Z-Test Results**:
  - Report the Z-Score and compare to the critical value.
- **Visualizations**:
  - **Histogram**: Display observed frequencies of leading digits.
  - **Table**: Compare observed and expected frequencies.
  - **Z-Score Chart** (if applicable).

---

6. Conclusion and Interpretation

- Summarize the results:
  - **Chi-Square Test**: Is the leading-digit distribution uniform?
  - **Z-Test**: Does the sample mean match the population mean?

- Discuss limitations:
    - Data source reliability, sample size, and any assumptions made.
- **Potential Extensions**: Compare results to **Benford's Law** (if applicable).

Final Review

This final structure now includes all critical components:

- Problem statement and goals.
- Data requirements and preparation.
- Hypothesis testing using **Chi-Square** and **Z-Test**.
- Results, visualizations, and interpretation.