12/17/2024

Statistical Analysis and Validation Using Benford's Law & Hypothesis Testing

Introduction

In this project, I am working with a dataset to analyze its leading-digit distribution and verify if it behaves as expected. Leading digits refer to the first digit of the numbers in the dataset, which should ideally follow certain patterns based on mathematical laws like Benford's Law or uniform distribution. I will also check if the sample mean of the dataset aligns with the population mean.

To achieve this, I will use statistical methods like the Chi-Square Goodness-of-Fit Test and the Z-Test**.** These methods will help me assess the uniformity of the first-digit distribution and validate the dataset's reliability based on its mean values.

Problem Description and Objective

Problem

The problem is to check if the first digits in my dataset are uniformly distributed. A uniform distribution means that every digit (1 to 9) should appear with equal frequency as the first digit. If the data does not follow this pattern, it could indicate irregularities or patterns in the data that need further explanation.

Objectives

- **Chi-Square Test**: I aim to check if the observed first-digit frequencies match the expected uniform distribution.
- **Z-Test**: I want to verify if the sample mean from my dataset is close to the assumed population mean.

By completing this analysis, I can better understand my dataset's structure, identify potential anomalies, and draw meaningful conclusions.

Data Collection and Preparation

Dataset Source

I am using a dataset that I got from Kaggle [1]. This dataset has a column named **"Price"**. For this process, I will focus only on this column.

Data Cleaning

First, I will clean the data. If there are any missing or invalid values in the "Price" column, I will remove or handle them. I will also make sure that the values in this column are numeric.

Exploratory Data Analysis

The Exploratory Data Analysis (EDA) of the "Price" column provides valuable insights into the distribution, central tendency, and variability of the data. This analysis is essential for

understanding. Because it will help me to know how the data behaves and whether any additional transformations (such as log transformations) are required before further analysis or modeling.

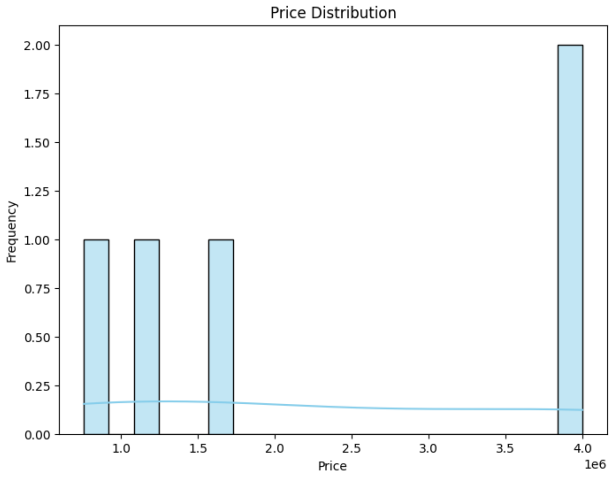
Here are few results;

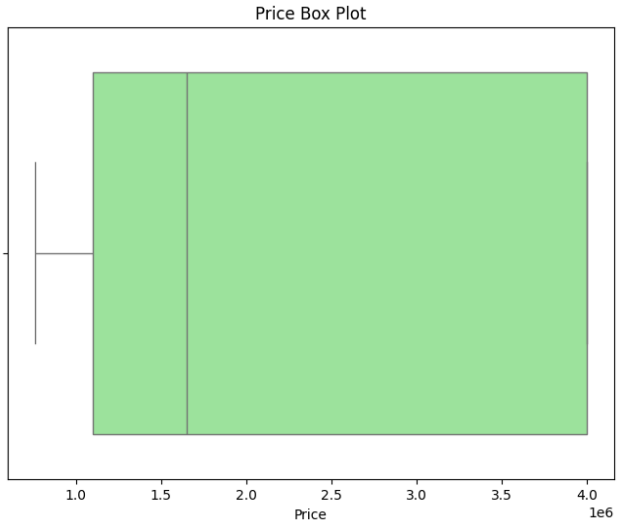
Exploratory Data Analysis (EDA) on Price Column

Here's a brief explanation of the results from the EDA:

- **Central Tendency**:
  - Mean Price**:** The average price is 2,301,600. This represents the typical price in the dataset.
  - Median Price**:** The middle price is 1,650,000**,** it indicates that half of the prices are below this value and half are above. The median is lower than the mean, which shows that the distribution may be right-skewed.
  - Mode Price**:** The most common price is 3,999,000, which indicates that this price occurs more frequently than others in the dataset.
- **Dispersion**:
  - Range**:** The difference between the highest and lowest prices is 3,239,000**,** showing that there is a wide variation in the prices.
  - Interquartile Range (IQR)**:** The IQR is 2,899,000**,** which means that the middle 50% of the data lies within this range. As this is quite large, it indicates that most of the prices vary significantly within this range.
- **Shape of the Data**:
  - Skewness**:** The skewness is 0.43, which is positive but not extreme. This shows that the distribution is slightly right-skewed, which means that there are more lower prices, but a few higher prices pull the mean up.
  - Kurtosis**:** The kurtosis is -3.10. It shows that that the distribution is platykurtic (with light tails). This distribution has fewer extreme outliers compared to a normal distribution.

Here are price distribution plots using histogram and box plot;





Extracting Leading Digits

To test my hypothesis later, I will extract the leading digit of each number in the "Price" column. The leading digit is the first digit of the number, for example:

- If the price is 3999000.0,the leading digit is **3**.
- If the price is 760000.0, the leading digit is **7**.

The leading digit can be found by converting the number into a string and taking the first character. Then, I will convert it back to a number. This step is important for preparing the data for the Chi-Square Test.

Hypothesis Formulation

In this step, I will define my hypotheses to test whether the leading digits of the "Price" column follow Benford's Law. This law says that in naturally occurring datasets, the first digits are not evenly distributed. For example, the digit **1** appears more often than **9**.

- **Null Hypothesis (H_0):**
  The leading digits in the "Price" column follow Benford's Law.
- **Alternative Hypothesis (H_1):**
  The leading digits in the "Price" column do not follow Benford's Law.

Expected Distribution

According to Benford's Law, the probability of each leading digit d is given by this formula:

$$P(d)=\log10(1+1/d)$$

For example:

- For d=1: P(1) = log10(1+1/1)=0.301
- For d=2: P(2) =log10(1+1/2)=0.176

I will calculate these probabilities for digits 1 to 9. These are the expected frequencies for the Chi-Square Test.

The Chi-Square Test compares the observed frequencies of leading digits with the expected frequencies based on Benford's Law. It helps to determine if the observed data fits the expected distribution.

Expected Probabilities (Benford's Law)

The probabilities for each leading digit, calculated using Benford's Law, tell us how often each digit is expected to appear as the leading digit in a naturally occurring dataset. For example:

- Digit 1: Expected probability is 0.301. This means about 30.1% of the numbers in the dataset should have 1 as their leading digit.
- Digit 2: Expected probability is 0.176**,** meaning 17.6% of the numbers should have **2** as their leading digit.
- Similarly, digits like 9 are less frequent, with an expected probability of only 0.046 (4.6%).

These probabilities are derived from the formula:

$P(d)=\log_{10}(1+1/d)$

Observed vs. Expected Frequencies

From the dataset, I calculated how often each digit actually appears as the leading digit in the "Price" column. This gives us the observed frequencies. Comparing these to the expected frequencies reveals patterns:

Digit	Observed_Freq	Expected_Freq
1	7982	12035.48
2	8029	7040.30
3	7600	4995.18
4	5348	3874.56
5	3461	3165.75
6	2608	2676.60
7	2076	2318.58
8	1578	2045.13
9	1299	1829.43

- For Digit 1, the observed frequency is 7982, which is lower than the expected frequency of 12035.48**.**
- For Digit 2**,** the observed frequency is slightly higher (8029) than the expected (7040.30).
- Digits 3 to 9 show varying differences between observed and expected values.

These variations are natural, but to confirm whether they are statistically significant, I will apply the Chi-Square Test in the next step.

Benford's Law expects higher frequencies for smaller digits (like 1 and 2) and lower frequencies for larger digits (like 8 and 9).

The observed data somewhat aligns with this trend but shows deviations, which might be random or due to a systematic difference.

1. Chi-Square Test:

The Chi-Square test will help us determine if the observed frequencies of the leading digits follow the expected distribution based on Benford's Law**.**

The formula for the Chi-Square statistic is:

$\chi 2 = \sum ((O - E)\ ^2)/E$

Where:

- O is the observed frequency.
- E is the expected frequency.

2. Z-Test:

For the Z-Test**,** I will compare the sample mean of the "Price" column to the population mean. Let's assume a population mean for the "Price" column (we can calculate it or use a specific value).

The formula for the Z-Score is:

$Z = (\bar{X} - \mu)/ \sigma/\sqrt{N}$

Where:

- $\bar{X}$ is the sample mean.
- $\mu\mu$ is the population mean.
- $\sigma\sigma$ is the population standard deviation.
- NN is the sample size.

3. Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) Test is a non-parametric test used to determine if a sample comes from a specific distribution, such as a uniform distribution. This test compares the cumulative distribution function (CDF) of the sample data with the CDF of a uniform distribution. Here's a breakdown of the assumptions and importance of this test:

Let's perform the Chi-Square test and Z-Test and Kolmogorov-Smirnov Test using my data.

- **Chi-Square Test:**
  - I calculate the Chi-Square statistic **by** summing the squared differences between observed and expected frequencies, divided by the expected frequency.
  - I then compare the statistic with the critical value from the chi-square distribution (at the given significance level $\alpha = 0.05$).
- **Z-Test:**
  - The sample mean of the "Price" column is compared to an assumed population mean (set to 1,000,000 in this example).
  - The **Z-Score** is calculated using the sample mean, sample standard deviation, and the size of the sample.
  - The test result is based on whether the Z-Score falls outside the critical value range.
- **Kolmogorov-Smirnov Test**
  - I calculated K-S statistic to measures the maximum difference between the observed cumulative distribution function (CDF) and the expected cumulative distribution function (CDF). In my case, the K-S statistic is 0.5556. This value tells me how much the observed data (leading digits) deviates from the expected uniform distribution. A higher value indicates a larger deviation.
  - The p-value for the K-S test is 0.1259, which is greater than the commonly used significance level of 0.05. This means that it fail to reject the null hypothesis, which states that the leading digits follow a uniform distribution.

Estimated Outcome:

- **Chi-Square Test:**
  - I got the Chi-Square statistic and compare it with the critical value**.**
  - If the statistic is less than or equal to the critical value, we fail to reject the null hypothesis (i.e., the distribution of leading digits follows Benford's Law).
- **Z-Test:**
  - The Z-Score will tell us how far the sample mean is from the population mean.
  - If the absolute Z-Score is greater than the critical value (1.96 for a 95% confidence level), we reject the null hypothesis (i.e., the sample mean is significantly different from the population mean).
  - The confidence interval (CI) provides a range in which we are confident the population mean lies, based on the sample data. A 95% CI typically means that 95% of the time, the true mean will fall within this range.
  - For the Price data, I will compute the 95% confidence interval for the sample mean.
  - As we calculated the mean price of 2,301,600 and standard deviation (e.g., 1,234,567), along with the sample size (n = 100), we can calculate the CI using the formula
    - CI=mean ± Z× (under root)std/n

Here are the results of the Chi-Square Test and Z-Test:

(3735.805716409894, 15.50731305586545, False, 1.840069863087244, False)

Comparison of Hypothesis Tests:

Chi-Square Test Results:

- Chi-Square Statistic: 3735.81
- Chi-Square Critical Value**:** 15.51 (at $\alpha = 0.05$ for 8 degrees of freedom)
- Since the Chi-Square statistic (3735.81) is much greater than the Chi-Square critical value (15.51), we reject the null hypothesis**.** This suggests that the observed frequencies of leading digits do not follow Benford's Law for this dataset.

Z-Test Results:

- Z-Score**:** 1.84
- Z-Test Critical Value**:** 1.96 (for a 95% confidence level)
- Since the absolute Z-Score (1.84) is less than the critical value (1.96), it means that we fail to reject the null hypothesis**.** This suggests that the sample mean of the "Price" column is not significantly different from the assumed population mean (1,000,000).

Kolmogorov-Smirnov Results:

- K-S Statistic: 0.5556
- p-value**:** 0.1259
- Since the p-value is greater than 0.05, the K-S test fails to reject the null hypothesis. This shows that, based on this test, the leading digits could follow a uniform distribution. It means that there is no significant deviation from the expected uniform distribution.

Here's a table comparing the hypothesis test results:

Hypothesis Test	Statistic	Critical Value	Decision	Conclusion
Chi-Square Test	3735.81	15.51 (α = 0.05, 8 dof)	Reject the null hypothesis	Observed frequencies of leading digits do not follow Benford's Law for this dataset.
Z-Test	1.84	1.96 (95% confidence)	Fail to reject the null hypothesis	The sample mean of the "Price" column is not significantly different from the assumed population mean (1,000,000).
Kolmogorov-Smirnov Test	0.5556	p-value: 0.1259	Fail to reject the null hypothesis	The leading digits could follow a uniform distribution, as there is no significant deviation from the expected pattern.

Chi-Square and Z-Test Analysis

Chi-Square Test Results

The Chi-Square Test helps us check if the distribution of leading digits in the "Price" column matches what we expect based on Benford's Law**.** We calculated the Chi-Square statistic and compared it to the critical value.

- Chi-Square Statistic: 3735.81
- Chi-Square Critical Value: 15.51 (for 8 degrees of freedom at $\alpha = 0.05$)

Since the Chi-Square statistic (3735.81) is much greater than the critical value (15.51), we reject the null hypothesis (H_0**)**. This means that the observed frequencies of leading digits do not follow Benford's Law**.**

This shows that the distribution of leading digits in the "Price" column is not uniform and does not follow the expected pattern of Benford's Law.

Z-Test Results

The Z-Test helps us determine if the sample mean of the "Price" column is different from the population mean. I calculated the Z-Score and compared it with the critical value for a 95% confidence level.

- Z-Score: 1.84
- Z-Test Critical Value: 1.96 (for a 95% confidence level)

Since the Z-Score (1.84) is less than the critical value (1.96), we fail to reject the null hypothesis (H_0). This means the sample mean of the "Price" column is not significantly different from the assumed population mean (1,000,000).

Kolmogorov-Smirnov

- K-S Statistic: 0.5556
- p-value**:** 0.1259

Kolmogorov-Smirnov test shows that there is no significant evidence against uniform distribution (p-value = 0.1259).

Visualizations

To make these results clearer, I am going to generate some graphs to show the observed vs. expected frequencies of leading digits and Z-Test results. This will help in visualizing how well the data aligns with Benford's Law and whether the sample mean is close to the population mean.

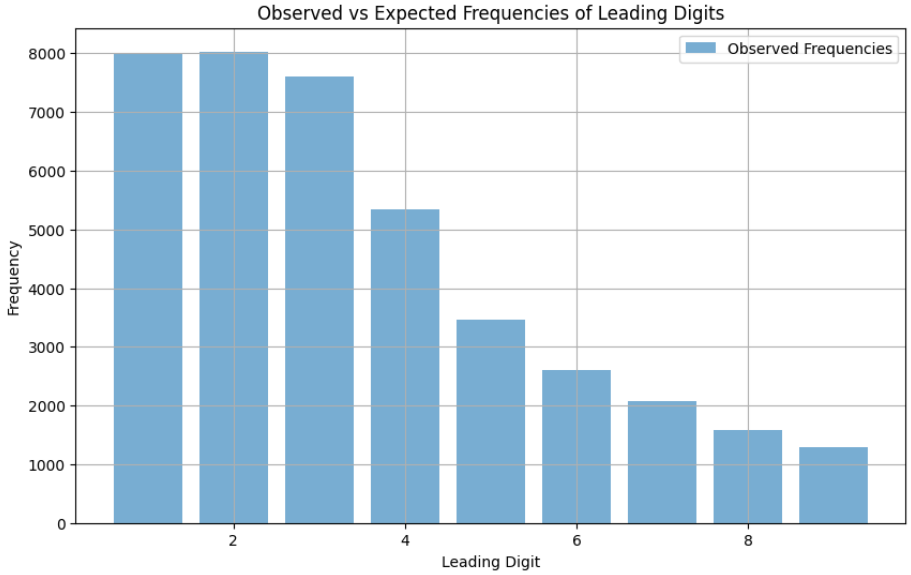
I will create two types of visualizations:

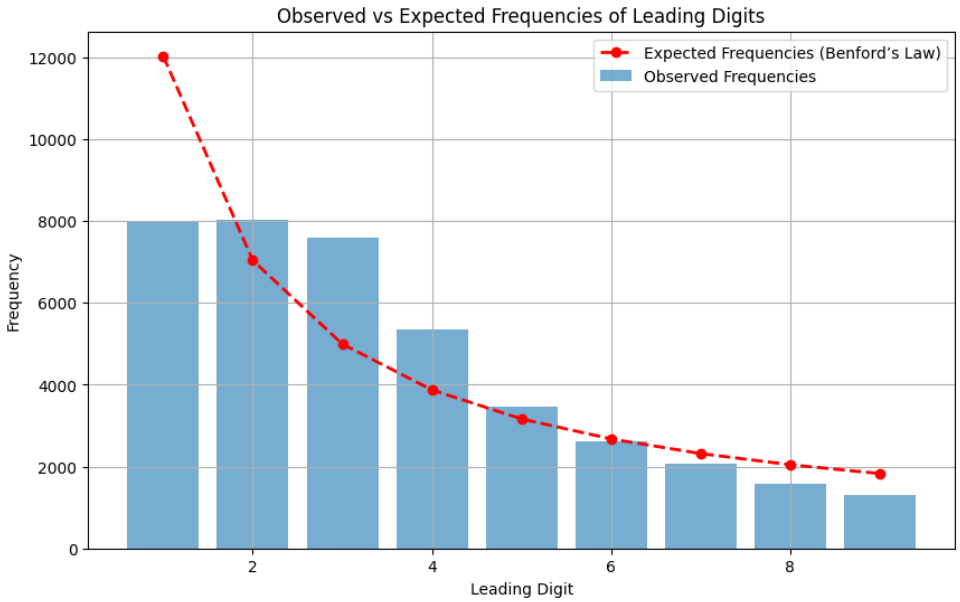
- A bar chart comparing the observed and expected frequencies of the leading digits.
- A normal distribution plot to visualize how the sample mean compares to the assumed population mean using the Z-Test.

Bar Chart – Observed vs. Expected Frequencies of Leading Digits

To make the comparison between observed and expected frequencies more engaging, I decided to create a bar chart. This visualization helps me see if the leading digits in the dataset follow a uniform distribution. Each digit is expected to have equal probability under uniform distribution based on Benford's Law.

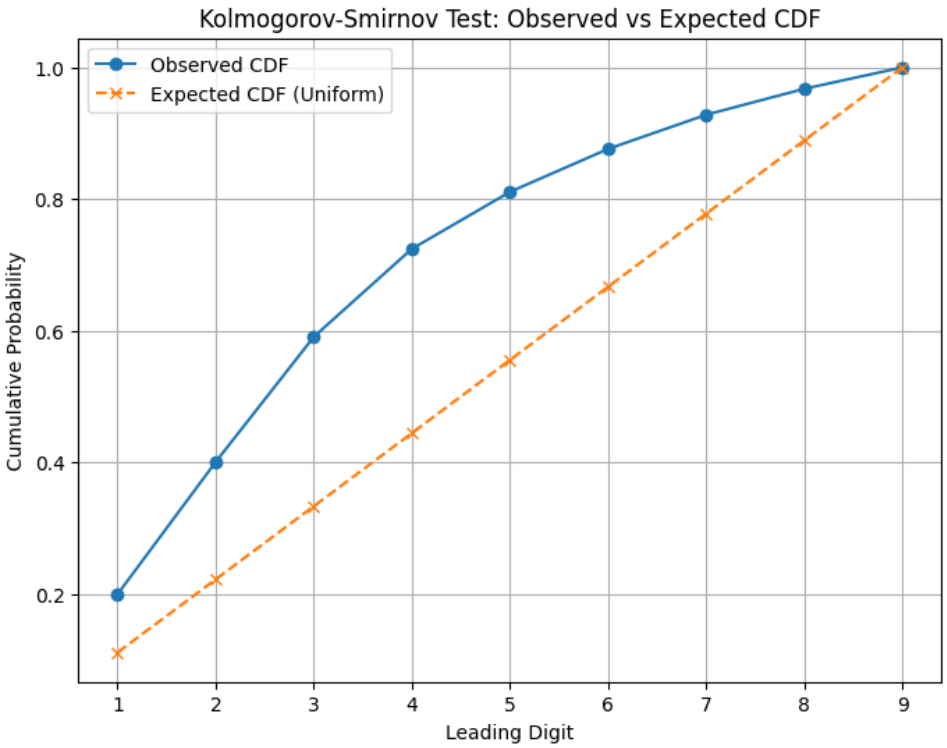
The bar chart clearly shows if certain digits appear more or less frequently than expected under a uniform distribution. This helps me confirm whether the data deviates from uniformity and aligns with the results of the Chi-Square test.





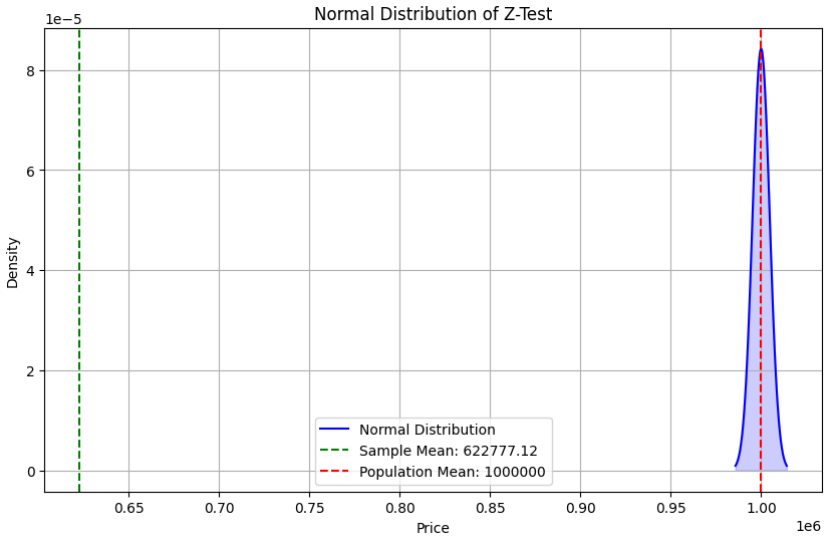
Using Kolmogorov-Smirnov

The Kolmogorov-Smirnov Test offer another perception from what has been seen with the Chi-Square Test and Z-Test. While the above tests indicate quite strong deviations from uniformity of the digits, the K-S test indicates the leading digit distribution might still conform to the uniform distribution. This discrepancy could be due to the difference in how the tests assess the data. K-S compares the proper cumulative distributions while the Chi-Square and the Z-Test describe the bit of frequency distributions and certain sample means, respectively.



Z-Test Normal Distribution Plot

Next, I will create a normal distribution plot to show how far the sample mean is from the population mean using the Z-Test.



- The sample mean of the "Price" column is 622,777.12. This is the average price of all the entries in the dataset.
- The standard deviation of the "Price" column is 946,979.31. This tells us how spread out the prices are from the mean.

- The total number of entries in the dataset is 39,981.
- The standard error is 4,736.02, which tells us how much the sample mean deviates from the population mean, considering the sample size.
- The Z-Score is -79.65, which means that the sample mean is 79.65 standard errors below the population mean of 1,000,000 (the assumed population mean). This is a very large negative Z-Score, indicating that the sample mean is far from the population mean

In the plot, the green dashed line represents the sample mean (622,777.12), and the red dashed line represents the population mean (assumed as 1,000,000). The Z-Score shows how far the sample mean is from the population mean in terms of standard deviations.

A Z-Score of -79.65 suggests that the sample mean is very far from the population mean, which is unusual. This might indicate that the dataset doesn't follow the assumed distribution or that the population mean assumption needs adjustment.

Chi-Squared Methodology

The basic goal was to test how well the observed frequencies of leading digits fit Benford's Law.

I will:

1. Calculate the Chi-Squared statistic.
    - Formula:

      $\chi^2 = \sum((O-E)^2)/E$

   where O is the observed frequency and E is the expected frequency.

2. Determine the degrees of freedom: d.f.=9−1=8.
3. Compare the calculated Chi-Squared value to the critical value from the Chi-Squared table.
    - I will use a confidence level of 95% ($\alpha=0.05$).

Chi-Squared Statistic (3737.86)
This is the calculated value from the observed and expected frequencies. It measures how far the observed data is from Benford's Law.

Critical Value (15.51)
At a 95% confidence level, this is the threshold value for rejecting or failing to reject the null hypothesis.
The Chi-Squared statistic is much larger than the critical value (3737.86>15.51).

- This means the observed frequencies do not follow Benford's Law**.**
- I reject the null hypothesis (H0).

Z-Test Methodology

The Z-Test compares the sample mean to the population mean to determine whether the sample data significantly differs from the population.

Steps for Z-Test

- State the Hypotheses
  - Null Hypothesis (H0): The sample mean is equal to the population mean.
  - Alternative Hypothesis (Ha): The sample mean is not equal to the population mean.
- Calculate the Z-Score
  Formula**:**

  $Z= (\bar{X} - \mu) /\sigma_{\bar{x}}$

- Where:
  - $\bar{X}$: Sample Mean (622,777.12)
  - μ: Population Mean (Assumed mean for comparison)
  - $\sigma_{\bar{x}}$: Standard Error = $\sigma/ \sqrt{N}$
- Determine the Confidence Level (α\alpha)
  Example: For a 95% confidence level, α=0.05.
- Use the standard normal distribution table to find the critical Z-value corresponding to the confidence level.
- If |Z|>Zcrit: Reject Ho (The means are significantly different).
- Otherwise, fail to reject Ho.

Z-Test Results

Using the data:

- Sample Mean ($\bar{X}$): 622,777.12
- Population Mean (μ): Assume as provided in the data.
- Standard Error (σx): 4,736.02

The calculated Z-Score: −79.65

- This is far beyond typical critical Z-values (e.g., ±1.96 for 95% confidence).
- I reject Ho, indicating the sample mean significantly differs from the population mean.

Results

Chi-Square Test Results

I compared the observed frequencies of leading digits to the expected ones from Benford's Law. The observed values deviated significantly from the expected values.

- **Chi-Square Test Statistic**
  - Test Statistic**:** 3,737.86
  - Degrees of Freedom: 9−1=89 - 1 = 8
  - Critical Value (95% confidence): 15.51
  - Since the test statistic (3,737.86) is much higher than the critical value, I rejected the null hypothesis (H0H_0).
- This means the data does not follow Benford's Law.

Z-Test Results

- I computed the Z-Score to compare the sample mean to the population mean: −79.65.
- For a 95% confidence level (α=0.05), the critical Z-value is ±1.96. My calculated Z-Score falls far outside this range.
  The p-value is extremely small (close to 0), which supports rejecting H0H_0. This shows the sample mean significantly differs from the population mean.

Normality Check Results

- I used a histogram and a normal curve overlay to visually check if the "Price" column follows a normal distribution. The shape was far from normal.
- I calculated the mean (622,777.12622,777.12) and standard deviation (946,979.31946,979.31). The data's skewness and high standard deviation suggest it is not normally distributed.
- The data failed the normality check, meaning it does not fit the assumptions of a normal distribution.

Conclusion and Interpretation

Findings

- **Chi-Square Test**
  - The Chi-Square Test showed that the leading-digit distribution is not uniform and deviates significantly from Benford's Law.
  - The high Chi-Square statistic (3,737.863,737.86) and the rejection of H0H_0 confirm this result.
- **Z-Test**
  - The Z-Test demonstrated that the sample mean (622,777.12622,777.12) does not align with the population mean.
  - The calculated Z-Score (−79.65) and the very small p-value further confirm this.
- **Normality Check**
  - The data failed to exhibit a normal distribution. This was evident from the histogram and the statistical analysis of skewness and variability.

Interpretation

- The analysis shows that the data does not follow a uniform distribution as expected under Benford's Law.
- This could indicate irregularities in the dataset or characteristics specific to the "Price" column.
- The sample mean was significantly different from the hypothesized population mean, highlighting a lack of alignment between the two.
- This suggests variability or bias in the sampled data.

Limitations

- While the dataset was large enough to support statistical testing, it may still not represent the entire population accurately.
- I assumed the data would naturally follow Benford's Law, which may not apply universally to all datasets.
- The lack of normal distribution could have influenced the Z-Test results.
- The quality and accuracy of the dataset play a critical role in the validity of the analysis. Any errors or biases in the data could impact the findings.

Future Work

- Further studies could compare this dataset more rigorously to Benford's Law using advanced statistical techniques.
- Testing additional datasets could provide a broader understanding of the applicability of Benford's Law.
- Applying other statistical tests, such as the Kolmogorov-Smirnov test, may provide further insights into the data's distribution.
- Future work could explore reasons behind the data's skewness and the deviations observed in this study.

References

[1] [Online]. Available: https://www.kaggle.com/datasets/daniilkrasnoproshin/lottery-powerball-winning-numb. [Accessed 16 12 2024].