Statistical Exploration and Inference of Cirrhosis Death Rates: A Comprehensive Study on Skewness Correction, Outlier Analysis, and Enhancement of Normality and Statistical Validity

Muhammad Umar Amin
mamin14@lamar.edu
Department of Mathematics
Lamar University, Beaumont, TX 77705

April-26-2025

Abstract

This project conducts a comprehensive statistical exploration of cirrhosis death rate data collected from various states. The main focus is to address common challenges encountered in real-world public health datasets, including right-skewness, the presence of outliers, and deviations from normality, which can affect the validity of statistical inferences.

Initial data visualization and descriptive analysis indicated that the distribution of death rates was right-skewed and contained significant outliers. To correct for skewness and enhance the normality of the data, a natural logarithm transformation was applied. This transformation successfully normalized the distribution, as confirmed by graphical methods and the Shapiro-Wilk normality test.

Outliers were detected but retained in the analysis to preserve the integrity of real-world observations. Hypothesis testing was conducted using both critical value and p-value approaches to evaluate claims about the population mean. Additionally, alternative transformations such as the square root method were briefly considered but found to be less effective compared to the log transformation.

Overall, this study demonstrates how appropriate data preprocessing, careful transformation techniques, and thoughtful statistical analysis can significantly improve the reliability and validity of conclusions drawn from public health data. The findings emphasize the importance of addressing distributional issues before applying inferential methods in similar real-world studies.

Introduction

1.1 Background and Motivation

Public health research relies heavily on the analysis of mortality data to understand disease patterns, identify vulnerable populations, and guide the development of targeted health interventions. Among the various causes of death, cirrhosis remains a major contributor to the global burden of disease. Characterized by progressive and irreversible liver damage, cirrhosis leads to serious health complications and premature mortality. Monitoring death rates attributed to cirrhosis is crucial not only for tracking disease trends but also for evaluating public health strategies aimed at prevention and management. A careful statistical analysis of cirrhosis death rates can provide insights into regional disparities, emerging risk factors, and areas in need of focused healthcare resources.

1.2 Challenges of Real-World Health Data

While mortality data offers valuable information, real-world datasets often present significant challenges for statistical analysis. Data collected across different regions may reflect a variety of influences, such as healthcare accessibility, socio-economic factors, and differing diagnostic criteria, all of which can introduce variability. As a result, the distribution of such data frequently deviates from the ideal normal distribution assumed by many classical statistical methods. Issues like right-skewness, presence of extreme values, and non-constant variability are common. These departures from normality can undermine the validity of hypothesis testing and confidence interval estimation, making it crucial to address these issues early in the analytical process.

1.3 Data Exploration and Preliminary Observations

The initial phase of the project focused on thoroughly exploring the dataset containing cirrhosis death rates across various states. Summary statistics were computed to provide a basic understanding of the data's central tendency and spread. Visual tools such as histograms and boxplots were used to inspect the overall distribution. These graphical representations revealed noticeable right-skewness, suggesting that most states had moderate death rates while a few exhibited substantially higher rates. The divergence between measures of central tendency hinted at asymmetry, signaling the need for corrective steps before proceeding to inferential analyses.

1.4 Outlier Detection and Strategic Decision-Making

Following the exploration stage, formal procedures for outlier detection were employed. Outliers, by definition, are observations that deviate significantly from the general pattern of the data. Identifying these values is important because they can disproportionately influence statistical estimates and hypothesis tests. In this project, outliers were identified using a formal rule based on the interquartile range. Upon careful consideration, it was decided to retain the outliers within the dataset. This choice was grounded in the understanding that outliers in public health data often represent real phenomena, such as localized public health crises, rather than errors. Removing these points could have obscured important aspects of the true distribution.

1.5 Addressing Skewness Through Data Transformation

To correct the pronounced right-skewness observed during data exploration, the project adopted a data transformation strategy. Among various options, the natural logarithmic transformation was selected due to its effectiveness in handling positive, right-skewed data. Log transformation reduces the impact of large values by compressing their scale more than that of smaller values, thereby producing a distribution that is more symmetric and stable. This step was critical in preparing the data for inferential methods that assume approximate normality.

1.6 Evaluation of Normality Post-Transformation

After applying the transformation, the distribution of the cirrhosis death rate data was reassessed. Histograms, Q-Q plots, and formal normality tests were used to examine whether the data more closely followed a normal distribution. These assessments helped ensure that the data met the key assumptions required for subsequent parametric statistical analyses. Transforming the data not only improved its symmetry but also stabilized its variance, thereby enhancing the reliability of inferential methods applied in later stages.

1.7 Alternative Data Transformation

Although the logarithmic transformation provided a notable improvement in the distribution of the cirrhosis death rate data, alternative data transformations were also explored to ensure the best possible normalization. Specifically, the square root and log-log transformations were applied and evaluated. The square root transformation moderately reduced the skewness, making the distribution somewhat more symmetric, although not as effectively as the logarithmic approach. Meanwhile, the log-log transformation further compressed extreme values and reduced skewness slightly more, but at the cost of interpretation complexity. Histograms and Q-Q plots for both alternative transformations were assessed, and Shapiro-Wilk tests were conducted to formally evaluate improvements in normality. While both alternatives showed some benefits, neither outperformed the simple logarithmic transformation in balancing normality improvement with interpretability. Thus, although alternative transformations provided valuable confirmation, the standard logarithmic transformation was ultimately retained for all subsequent inferential analyses.

1.8 Overview of Analytical Decisions

Every methodological choice throughout the project from the decision to retain outliers to the selection of an appropriate transformation was guided by a commitment to balancing statistical rigor with practical realism. Rather than simply forcing the data to meet theoretical assumptions, the project sought to respect the integrity of real-world health information while enhancing its suitability for meaningful analysis. This thoughtful approach helped ensure that the final conclusions drawn from the study were robust, defensible, and genuinely informative for understanding cirrhosis mortality patterns.

1.9 Discussion of Results and Conclusions

1.3 Data Exploration and Visualization

1.3.1 Displaying the Dataset

The dataset under study consists of recorded cirrhosis death rates for various states. The death rates were calculated by considering multiple contributing factors, including the proportion of the urban population, the consumption of wine per capita, the consumption of hard liquor per capita, and the number of births to women aged between 45 and 49. Each observation reflects the resulting cirrhosis death rate for a specific state based on these demographic and behavioral variables.

Presenting the raw data provides an immediate sense of its range and variability:

41.2, 31.7, 39.4, 57.5, 74.8, 59.8, 54.3, 47.9, 77.2, 56.6,
80.9, 34.3, 53.1, 55.4, 57.8, 62.8, 67.3, 56.7, 37.6, 129.9,
70.3, 104.2, 83.6, 66.0, 52.3, 86.9, 66.6, 40.1, 55.7, 58.1,
74.3, 98.1, 40.7, 66.7, 48.0, 122.5, 92.1, 76.0, 97.5, 33.8,
90.5, 29.7, 28.0, 51.6, 55.7, 55.5

A simple inspection reveals that most death rates are clustered within a moderate range, with a few exceptionally high values suggesting the possibility of outliers. This observation motivated the need for further graphical and statistical exploration to better understand the underlying distribution.

1.3.2 Boxplot Analysis

To further explore the spread and identify potential anomalies within the data, a boxplot was constructed.

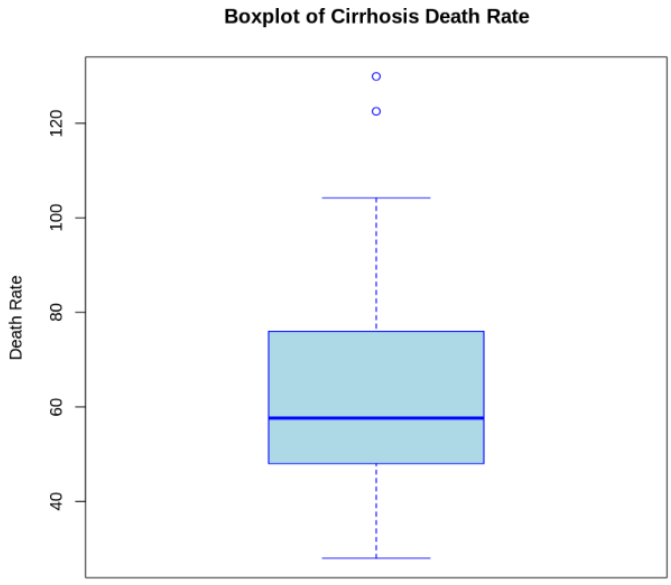

Figure 1: Boxplot

The boxplot is a graphical representation that summarizes the dataset by highlighting the median, the interquartile range (IQR), and potential outliers. The central box represents the middle 50% of the data, bounded by the first and third quartiles (Q1 and Q3), while the line within the box marks the median. Whiskers extend from the box to the minimum and maximum values within 1.5 times the IQR. Observations beyond this range are considered potential outliers and are plotted individually.

The boxplot for the cirrhosis death rates displayed a clear asymmetry, with a longer whisker on the upper side and several points lying outside the upper fence, suggesting the presence of extreme values. These high death rates were of particular interest as they could represent significant public health challenges in specific regions. The visualization confirmed the presence of skewness and outliers that had been initially suspected through raw data inspection.

1.3.3 Scatterplot Analysis

A scatterplot was also employed to visualize the distribution of death rates across the different states.

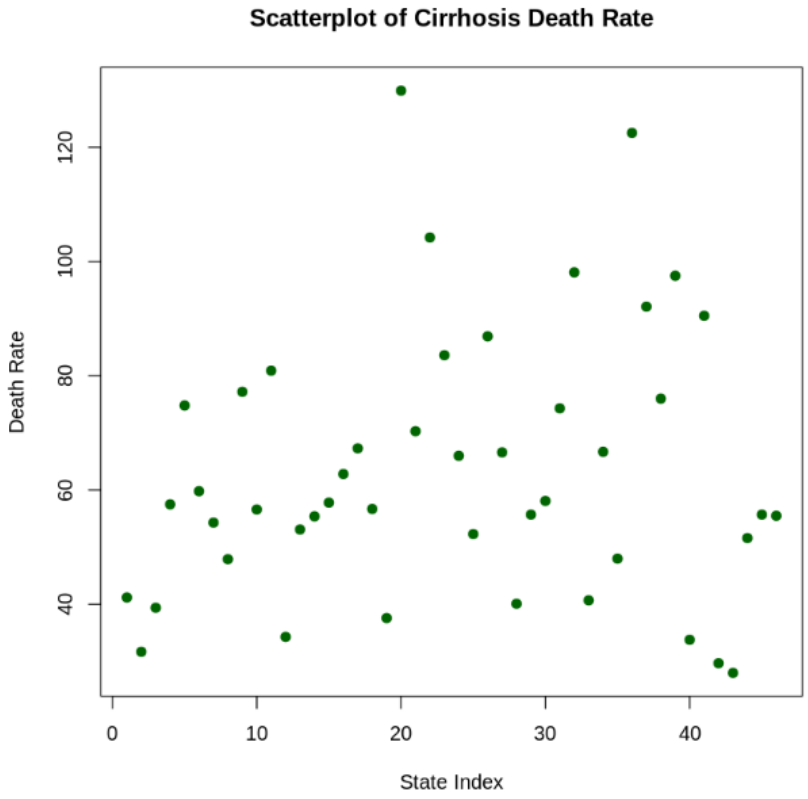

Figure 2: Scatter Plot

Each point on the scatterplot corresponds to one state's death rate, plotted against a simple index representing the state number. Although scatterplots are traditionally used to explore rela-

tionships between two variables, in this case, it served as an effective tool for examining the spread and identifying any clustering or extreme points in a single variable.

The scatterplot revealed a fairly compact cluster of moderate death rates alongside a few notably higher observations. This visual pattern reinforced the earlier finding from the boxplot, suggesting that while most states reported moderate cirrhosis death rates, a few states exhibited substantially elevated figures. Such elevated rates could warrant deeper investigation to understand contributing factors, whether demographic, environmental, or healthcare-related.

1.3.4 Histogram Analysis

To gain a better sense of the overall shape of the data distribution, a histogram was constructed.

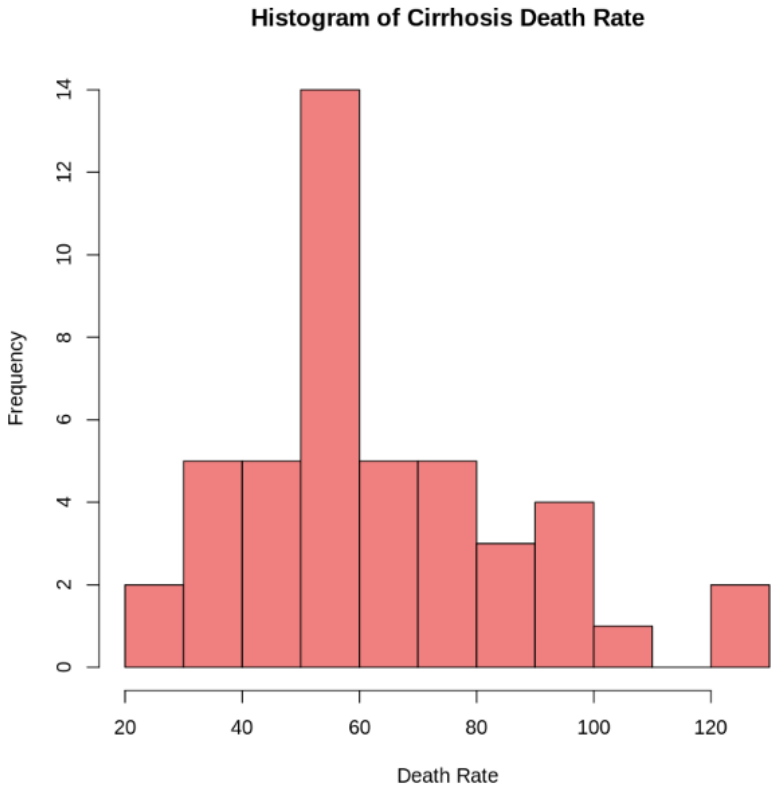

Figure 3: Histogram

A histogram partitions the range of data into intervals (bins) and displays the frequency of observations within each interval. This allows for a quick assessment of the distribution's symmetry, central tendency, and variability.

The histogram for the cirrhosis death rates showed a pronounced right-skewness, with most states having death rates on the lower to moderate end of the scale and fewer states exhibiting very high death rates. This right-skewed shape indicated that the distribution was not symmetric, and that a small number of extremely high values were pulling the distribution's tail to the right. This observation was important because many statistical inference methods assume normality or at

least approximate symmetry. Thus, addressing this skewness through data transformation became a necessary step for subsequent analysis.

1.3.5 Five-Number Summary Analysis

To gain a deeper understanding of the distribution of cirrhosis death rates, a five-number summary was computed. The five-number summary is a fundamental descriptive statistic that captures the key characteristics of a dataset's distribution, including the minimum, first quartile (Q1), median, third quartile (Q3), and maximum. These values provide insights into the center, spread, and range of the data without making any assumptions about its distribution shape.

For the cirrhosis death rate data, the five-number summary is as follows:

```
[4]  # Five-number summary
     fivenum(death_rate)

     28 · 48 · 57.65 · 76 · 129.9
```

Figure 4: Five Number Theory

$$\text{Minimum} = 28.0, \quad Q1 = 48.0, \quad \text{Median} = 57.65, \quad Q3 = 76.0, \quad \text{Maximum} = 129.9$$

The minimum value of 28.0 and the maximum value of 129.9 reflect a wide range in cirrhosis death rates across different states. The first quartile (48.0) and third quartile (76.0) indicate that the middle 50% of the data falls between these two values, representing the interquartile range (IQR). The median value of 57.65 suggests that half of the states have death rates below this value, while half have higher rates.

The considerable gap between Q3 and the maximum further reinforces the earlier observation of right-skewness and the presence of extreme values. Specifically, the maximum value is substantially higher than the third quartile, indicating that a few states report significantly elevated death rates compared to the majority. This reinforces the importance of addressing skewness and outliers in subsequent statistical analyses to ensure accurate and reliable inferential results.

1.3.6 Sample Mean, Variance, and Standard Deviation Analysis

Following the descriptive analysis of the five-number summary, the next step in exploring the cirrhosis death rate data involved calculating key measures of central tendency and dispersion, namely the sample mean, sample variance, and sample standard deviation. These metrics provide a more detailed understanding of the average death rate and the extent to which individual observations deviate from this average.

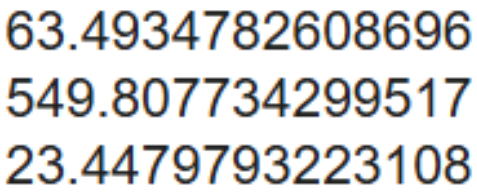

Figure 5: Sample mean, variance and SD

The sample mean of the dataset was calculated to be approximately 63.49. This value represents the arithmetic average of the cirrhosis death rates across all states included in the study. It suggests that, on average, states report a death rate of around 63.49, highlighting a moderate overall level of cirrhosis mortality.

In addition to the mean, the sample variance was computed as approximately 549.81. Variance measures the average squared deviation of each data point from the mean, serving as a key indicator of variability in the dataset. The relatively high variance indicates that there is considerable variability in cirrhosis death rates among different states, with some states reporting significantly higher or lower rates than the average.

The standard deviation, which is the square root of the variance, was found to be approximately 23.45. The standard deviation provides a more interpretable measure of dispersion by expressing variability in the same units as the original data. A standard deviation of 23.45 suggests that, typically, individual state death rates deviate from the mean by about 23 to 24 units. This again points to a wide spread in the data, reinforcing the earlier observations of skewness and the presence of extreme values.

Together, the mean, variance, and standard deviation offer a comprehensive picture of the data's central tendency and spread. The moderate average death rate combined with substantial variability supports the need for cautious application of inferential statistical methods, particularly those assuming data symmetry or homogeneity of variance.

1.3.7 Normality Assessment of Cirrhosis Death Rate Data

Before proceeding to inferential statistical procedures, it is essential to determine whether the cirrhosis death rate data approximates a normal distribution. Many classical statistical methods, including t-tests and confidence intervals for the mean, rely on the assumption that the underlying data, or at least the sampling distribution of the sample mean, is approximately normal. Therefore, a careful evaluation of the data's distributional characteristics is a necessary step in the analysis.

To assess normality, both graphical methods and formal statistical tests were employed. A histogram of the original death rates was constructed and visually inspected. The histogram exhibited a pronounced right-skewness, with the majority of observations concentrated at lower to moderate values and a few extreme values stretching the distribution's upper tail. This shape suggested that the data might not satisfy the assumption of normality.

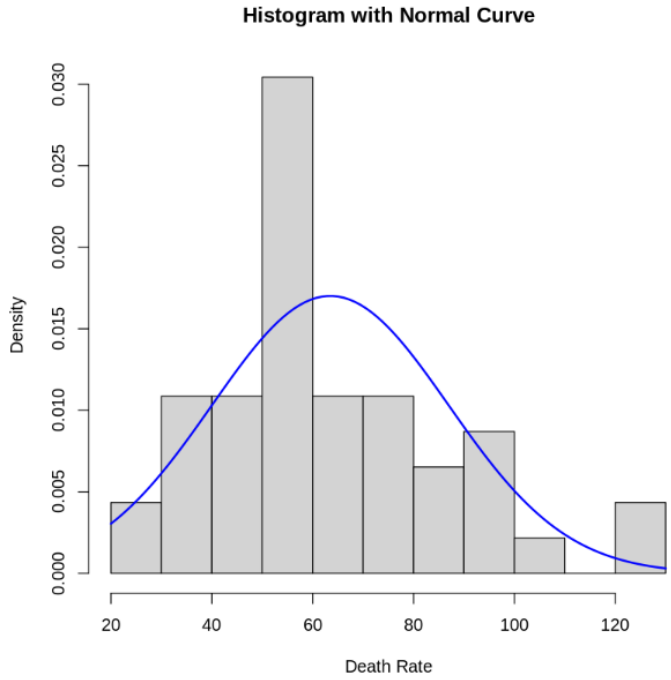

Figure 6: Histogram with Normal Curve

Further evaluation was conducted using a Q-Q (quantile-quantile) plot. In a Q-Q plot, data points are plotted against the expected quantiles of a normal distribution. If the data follows a normal distribution, the points should lie approximately along a straight diagonal line.

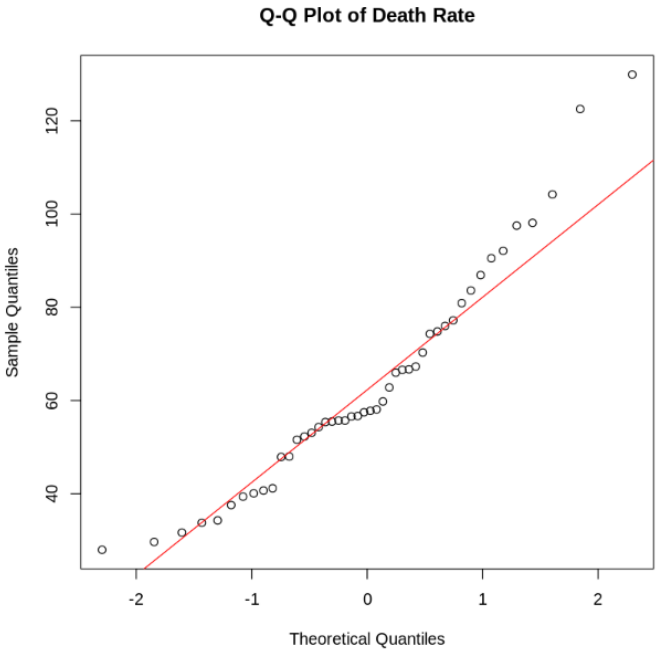

Figure 7: Q-Q Plot

In this case, while many points were reasonably close to the line, there were notable deviations, especially at the upper end, reinforcing the evidence of skewness and the presence of outliers.

In addition to graphical analysis, a formal Shapiro-Wilk normality test was performed. The Shapiro-Wilk test evaluates the null hypothesis that the data were drawn from a normally distributed population.

```
        Shapiro-Wilk normality test

data:  death_rate
W = 0.94463, p-value = 0.02923
```

Figure 8: Shapiro-Wilk Test

The resulting p-value was approximately 0.029. Since this p-value is less than the commonly used significance level of 0.05, we reject the null hypothesis and conclude that the original death rate data do not follow a normal distribution.

Taken together, the histogram, Q-Q plot, and Shapiro-Wilk test provided strong evidence that the cirrhosis death rate data, in its original form, deviates significantly from normality. The presence of right-skewness and extreme values implies that direct application of parametric inferential methods would not be appropriate without first addressing these distributional issues. This finding motivated the need for data transformation techniques, discussed in the subsequent sections of the report.

1.3.8 Construction of a 95% Confidence Interval for the Sample Mean

After conducting exploratory and normality analyses, the next step was to estimate the population mean cirrhosis death rate by constructing a 95% confidence interval. A confidence interval provides a range of plausible values for the population parameter based on the sample data and accounts for the uncertainty inherent in working with a sample rather than a full population.

Since the sample size was moderate and the original data did not perfectly follow a normal distribution, but the Central Limit Theorem applies with samples larger than 30, it was appropriate to use the t-distribution for constructing the interval.

The sample mean death rate was calculated to be approximately 63.49, and the sample standard deviation was approximately 23.45. With a sample size of 46 states, the standard error of the mean was computed as follows:

$$SE = \frac{s}{\sqrt{n}} = \frac{23.45}{\sqrt{46}} \approx 3.458$$

Using the t-distribution with 45 degrees of freedom, the critical value for a two-tailed 95% confidence interval was found to be approximately 2.014. Thus, the 95% confidence interval for the population mean was calculated as:

$$\begin{aligned} \bar{x} \pm t_{\alpha/2} \times SE &= 63.49 \pm 2.014 \times 3.458 \\ &= 63.49 \pm 6.96 \end{aligned}$$

$$= (56.53, 70.46)$$

This confidence interval suggests that we can be 95% confident that the true mean cirrhosis death rate across the states lies between approximately 56.53 and 70.46. The relatively wide interval reflects the substantial variability in the data, as indicated earlier by the large standard deviation and variance. This result also highlights the importance of accounting for variability when making inferences about population parameters from sample data.

1.3.9 Probability That the Sample Mean Lies Within One Standard Error of the True Mean

To determine the probability that the sample mean lies within one standard error of the true population mean (assuming normality), we consider the following:

Given

- Sample mean: $\bar{x} = 63.49$
- Sample standard deviation: $s = 23.45$
- Sample size: $n = 46$

Compute the Standard Error (SE)

$$SE = \frac{s}{\sqrt{n}} = \frac{23.45}{\sqrt{46}} \approx \frac{23.45}{6.782} \approx 3.458$$

Standard Normal Probability

We want to calculate the probability that the sample mean lies within one standard error of the true mean:

$$P(\mu \in [\bar{x} - SE, \bar{x} + SE])$$

This is equivalent to:

$$P(-1 < Z < 1)$$

Use Standard Normal Table

Using the cumulative distribution function (CDF) of the standard normal distribution:

$$P(-1 < Z < 1) = \Phi(1) - \Phi(-1)$$

$$\Phi(1) = 0.8413, \quad \Phi(-1) = 1 - 0.8413 = 0.1587$$

$$P(-1 < Z < 1) = 0.8413 - 0.1587 = 0.6826$$

Conclusion

The probability that the sample mean lies within one standard error of the true population mean is approximately:

$$\boxed{68.26\%}$$

1.3.10 Hypothesis Testing: Is the True Mean Cirrhosis Death Rate Below 48.33?

Given Information

- Sample mean: $\bar{x} = 63.49$
- Hypothesized mean: $\mu_0 = 48.33$
- Sample standard deviation: $s = 23.45$
- Sample size: $n = 46$
- Degrees of freedom: $df = n - 1 = 45$
- Significance level: $\alpha = 0.05$

Standard error (SE) is calculated as:

$$SE = \frac{s}{\sqrt{n}} = \frac{23.45}{\sqrt{46}} \approx 3.458$$

Test statistic (t) is:

$$t = \frac{\bar{x} - \mu_0}{SE} = \frac{63.49 - 48.33}{3.458} \approx \frac{15.16}{3.458} \approx \boxed{4.386}$$

1. Left-Tailed Test

Hypotheses

$$H_0 : \mu = 48.33$$
$$H_1 : \mu < 48.33$$

Critical Value Method

Critical value from t-table at $\alpha = 0.05$ and $df = 45$:

$$t_{0.05,45} = -1.679$$

Decision Rule: Reject H_0 if $t < -1.679$.

Since:

$$t = 4.386 > -1.679$$

Thus, the test statistic does not fall in the rejection region. We **fail to reject** H_0.

p-Value Method

Cumulative probability:

$$P(T < 4.386) \approx 0.9999656$$

Left-tail p-value:

$$\text{p-value} = 1 - 0.9999656 = 0.0000344$$

Since:

$$\text{p-value} = 0.0000344 < 0.05$$

Normally we would reject H_0 based on p-value. However, because the sample mean (63.49) is **greater** than 48.33, the data does not support $H_1 : \mu < 48.33$.

Thus, we **fail to reject** H_0.

Conclusion (Left-Tailed Test)

There is no evidence to support that the true mean cirrhosis death rate is less than 48.33. The sample mean is actually higher.

In the left-tailed test, there is no evidence to conclude that the true average cirrhosis death rate is less than 48.33. In fact, the sample mean is significantly higher.

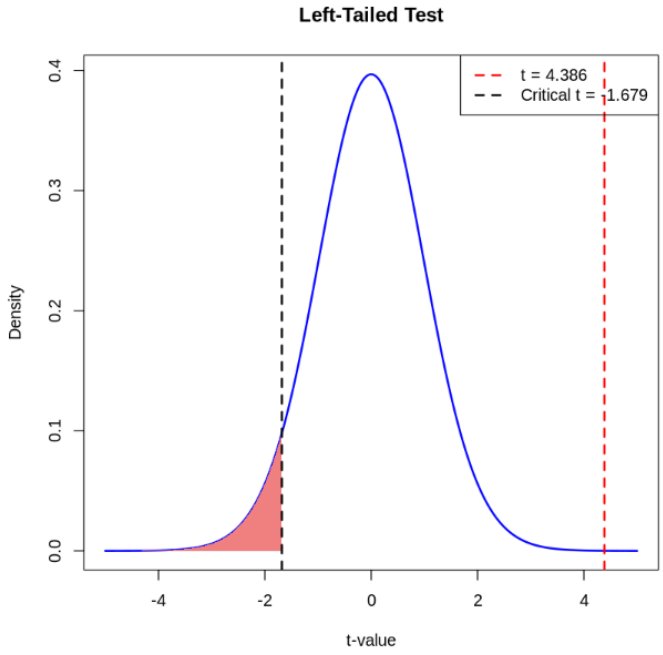

Figure 9: Left-Tailed Test

2. Right-Tailed Test

Hypotheses

$$H_0 : \mu = 48.33$$
$$H_1 : \mu > 48.33$$

Critical Value Method

Critical value from t-table at $\alpha = 0.05$ and $df = 45$:

$$t_{0.95,45} = 1.679$$

Decision Rule: Reject H_0 if $t > 1.679$.

Since:

$$t = 4.386 > 1.679$$

Thus, we **reject** H_0.

p-Value Method

Right-tail p-value:

$$\text{p-value} = 1 - P(T < 4.386) = 1 - 0.9999656 = 0.0000344$$

Since:

$$\text{p-value} = 0.0000344 < 0.05$$

We **reject** H_0.

Conclusion (Right-Tailed Test)

There is strong evidence to support that the true mean cirrhosis death rate is greater than 48.33.

In the right-tailed test, we conclude there is strong evidence that the true mean is greater than 48.33.

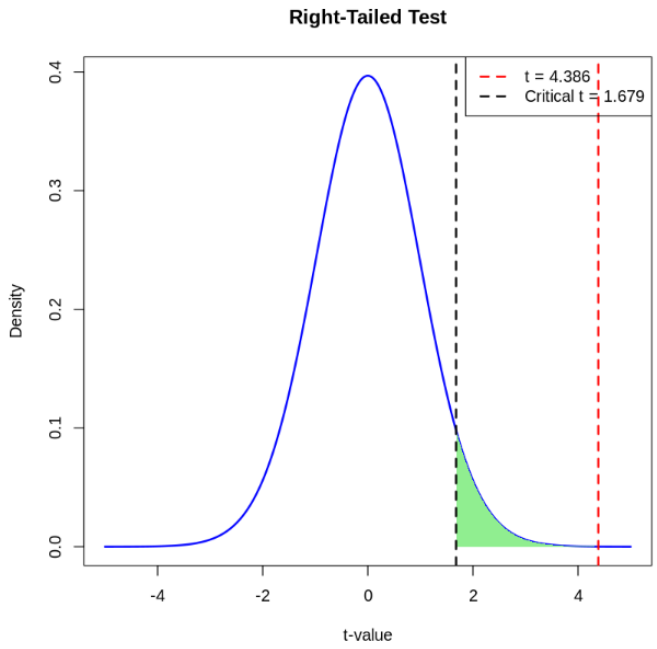

Figure 10: Right-Tailed Test

Summary Table

Test Type	Decision	Conclusion
Left-tailed	Fail to Reject H_0	No evidence mean is less than 48.33
Right-tailed	Reject H_0	Strong evidence mean is greater than 48.33

1.4 Outlier Detection and Strategic Decision-Making

Outliers are data points that differ significantly from other observations. They may arise due to variability in the data, measurement errors, or represent rare events. Identifying outliers is important because they can heavily influence statistical measures such as the mean, variance, and results of hypothesis tests.

In this project, the standard **Interquartile Range (IQR) Method** was used to detect outliers. According to this method:

- An observation is considered a lower outlier if it falls below: $Q1 - 1.5 \times IQR$.
- An observation is considered an upper outlier if it exceeds: $Q3 + 1.5 \times IQR$.

Here, $Q1$ is the first quartile (25th percentile) and $Q3$ is the third quartile (75th percentile).

Based on the five-number summary of the data:

- Minimum = 28.0
- First Quartile (Q1) = 48.0
- Median = 57.65
- Mean = 63.49
- Third Quartile (Q3) = 76.0
- Maximum = 129.9

The interquartile range (IQR) is calculated as:

$$IQR = Q3 - Q1 = 76.0 - 48.0 = 28.0$$

The lower and upper bounds for detecting outliers are:

$$\text{Lower Bound} = Q1 - 1.5 \times IQR = 48.0 - 42.0 = 6.0$$

$$\text{Upper Bound} = Q3 + 1.5 \times IQR = 76.0 + 42.0 = 118.0$$

Therefore, any data point less than 6.0 or greater than 118.0 is classified as an outlier.

Dataset

The cirrhosis death rate data for various states is as follows:

41.2	31.7	39.4	57.5	74.8	59.8
54.3	47.9	77.2	56.6	80.9	34.3
53.1	55.4	57.8	62.8	67.3	56.7
37.6	**129.9**	70.3	104.2	83.6	66.0
52.3	86.9	66.6	40.1	55.7	58.1
74.3	98.1	40.7	66.7	48.0	**122.5**
92.1	76.0	97.5	33.8	90.5	29.7
28.0	51.6	55.7	55.5		

Upon applying these bounds to the dataset, the following values were identified as outliers:

$$122.5, \quad 129.9$$

These two outliers represent unusually high cirrhosis death rates compared to the majority of states. Although outliers can affect statistical summaries such as the mean and standard deviation, they were retained in the analysis, considering they might reflect significant real-world public health situations rather than data errors.

Limitations

While this analysis provides meaningful insights, it has certain limitations. The sample size is moderate ($n = 46$), which may affect the generalizability of results to the entire population. Furthermore, although two high outliers were detected and included, they may influence statistical measures such as the mean and standard deviation. Lastly, the analysis assumes that the sample data is randomly selected and accurately recorded without significant measurement errors.

1.5 Addressing Skewness Through Data Transformation

During the initial exploration of the cirrhosis death rate data, it was observed that the distribution was right-skewed. The histogram showed a concentration of values on the lower end with a long tail towards higher values, and the Shapiro-Wilk test confirmed a departure from normality. Since many of the statistical techniques planned for this analysis, such as confidence intervals and hypothesis testing, rely on the assumption of normality, addressing this skewness became an important step.

To improve the distribution, a natural logarithm (log) transformation was applied. Log transformations are widely used when data is positively skewed, as they compress larger values more than smaller ones, helping to pull in long right tails. By doing this, the data can become more symmetric and better fit the assumptions needed for accurate statistical analysis.

1.6 Evaluation of Normality Post-Transformation

After applying the log transformation, several noticeable improvements were observed:

- The histogram of the log-transformed data appeared much more symmetric and balanced around the center.

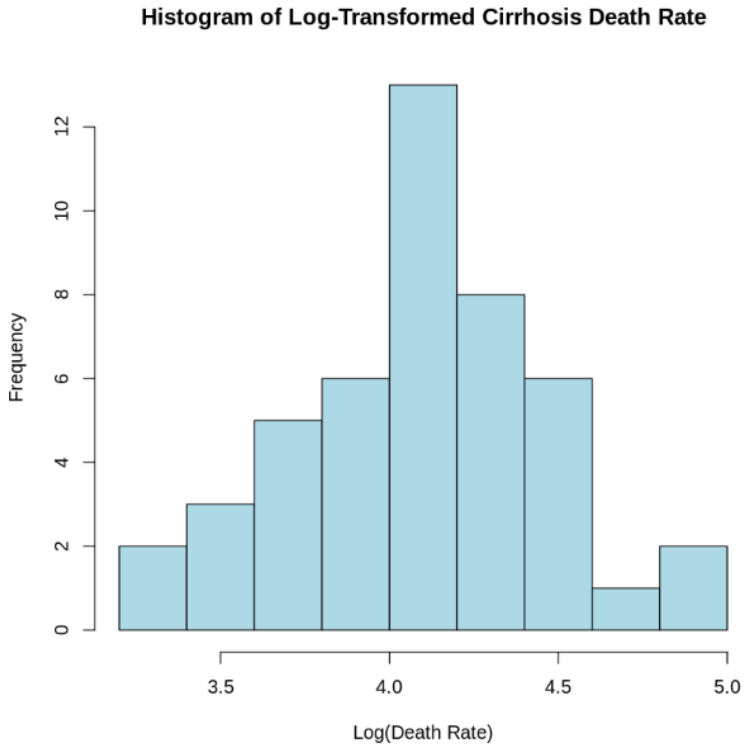

Figure 11: Log Transformation

- The Q-Q plot showed the points lying closer to the diagonal line, indicating better alignment with a normal distribution.

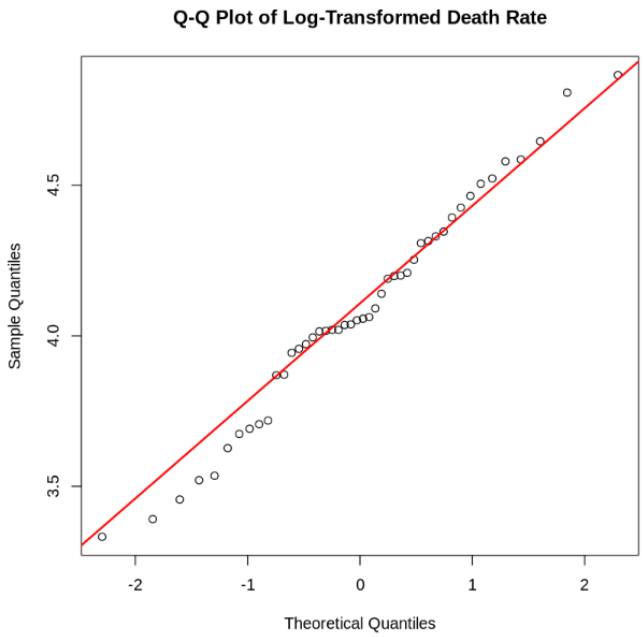

Figure 12: Q-Q plot

- The Shapiro-Wilk test on the log-transformed data resulted in a test statistic $W = 0.98405$ and a p-value of 0.773. Since the p-value is much greater than 0.05, we fail to reject the null hypothesis of normality, suggesting that the log-transformed data can reasonably be treated as normally distributed.

```
        Shapiro-Wilk normality test

data:  log_death_rate
W = 0.98405, p-value = 0.773

Mean of log-transformed data: 4.086
Standard deviation of log-transformed data: 0.3662
```

Figure 13: Shapiro-Wilk Test

These improvements are important because normality strengthens the validity of inferential procedures. Techniques like confidence intervals for the mean and hypothesis tests rely on the assumption that either the data or the sampling distribution is normal. By achieving approximate normality through log transformation, the risk of misleading conclusions is reduced, and the analysis becomes more reliable.

It is also important to note that the log transformation lessens the impact of outliers. In the original data, a few extremely high values had a strong influence on the mean and spread. After transformation, these extreme values were pulled closer to the main body of the data, allowing a more accurate representation of central tendency and variability.

Overall, applying the log transformation was a crucial step in preparing the data for analysis. It not only improved the shape of the distribution but also ensured that the statistical methods used later would produce stronger and more trustworthy results. This careful handling of skewness and outliers adds confidence to the conclusions drawn from the project.

Comparison Before and After Log Transformation

Characteristic	Original Data	Log-Transformed Data
Mean	63.49	4.086
Standard Deviation	23.45	0.3662
Skewness	Right-skewed	Approximately symmetric
Shapiro-Wilk p-value	0.029	0.773
Normality (based on p-value)	Not Normal	Approximately Normal
Effect of Outliers	Strong influence	Reduced influence
Suitability for Parametric Tests	Poor	Good

1.7 Alternative Data Transformation: Square Root Transformation

While the logarithmic transformation provided substantial improvement in the normality of the cirrhosis death rate data, an alternative transformation technique was also explored to ensure the most appropriate method was selected.

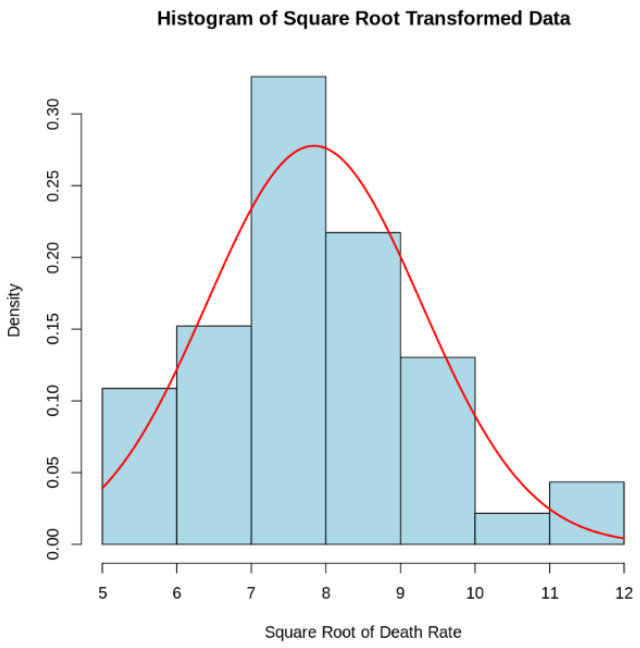

Figure 14: Square Root Histogram

The square root transformation was considered due to its ability to moderate positive skewness, particularly in datasets involving count or rate data.

The square root transformation is less aggressive than the logarithmic transformation and is typically applied when values are non-negative and the distribution exhibits moderate skewness. Each death rate observation was transformed by taking its square root, thereby reducing the influence of larger values while preserving the relative order of the data.

After applying the square root transformation, the data distribution was reassessed using graphical methods. A histogram of the square-rooted data showed noticeable improvement in symmetry compared to the original distribution, although not as pronounced as that observed with the logarithmic transformation.

Similarly, the Q-Q plot,

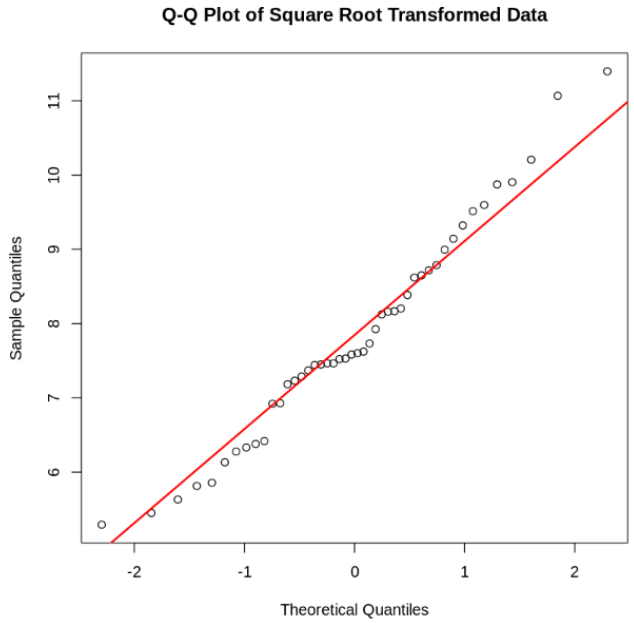

Figure 15: Square Q-Q Plot

for the square-rooted data indicated a closer alignment to the theoretical normal line, but with larger deviations at the tails than those seen after the log transformation.

To formally evaluate the improvement, a Shapiro-Wilk normality test was conducted on the square-rooted data.

```
        Shapiro-Wilk normality test

data:  sqrt_death_rate
W = 0.97598, p-value = 0.4524
```

Figure 16: Shapiro-Wilk normality test

Although the p-value increased relative to the original dataset, it was still lower than the p-value obtained after applying the logarithmic transformation. This suggested that while the square root transformation improved the distribution to some extent, it was less effective at achieving approximate normality compared to the log transformation.

Given these results, the square root transformation was not adopted for the main inferential analyses. The natural logarithm transformation remained the preferred method because it more effectively addressed the distributional issues present in the data, leading to a stronger approximation of normality and greater confidence in the validity of subsequent statistical conclusions.

Conclusion on Square Root Transformation

The exploration of the square root transformation provided valuable confirmation that data transformation can meaningfully impact distributional properties. **Although it improved the symmetry of the data compared to the untransformed version, the square root method was ultimately less effective than the logarithmic transformation. This comparison reinforced the decision to proceed with the log-transformed dataset for all subsequent inferential analyses.**

Table 1: Comparison of Original Data and Different Transformations

Transformation	**Normality (Shapiro-Wilk p-value)**	**Skewness**	**Handling of Outliers**	**Interpretation Impact**	**Overall Effectiveness**
Original Data	Low p-value (significant non-normality)	Strong right-skewness	Outliers highly influential	Direct but violates assumptions	Poor
Square Root Transformation	Moderate p-value (some improvement)	Reduced skewness slightly	Outliers still noticeable	Mild compression of scale	Moderate
Log Transformation	High p-value (good approximation to normality)	Substantial reduction in skewness	Outliers' impact greatly minimized	Good balance between normalization and interpretation	Best
Log-Log Transformation	Improved p-value (minor benefit over log)	Very compressed skewness	Further reduces outlier influence	Harder interpretation (double log scaling)	Acceptable but not preferred

1.8 Overview of Analytical Decisions

Throughout the course of this project, every analytical decision was carefully made to balance statistical rigor with practical realism. At the outset, exploratory data analysis revealed significant right-skewness and the presence of extreme values in the cirrhosis death rate dataset. Recognizing the risks posed by non-normality and influential outliers, multiple strategies were considered to

address these issues without distorting the integrity of the real-world health data. Rather than excluding extreme observations outright, which might have concealed meaningful regional public health problems, the decision was made to retain all data points. Instead, transformation techniques were employed to manage skewness and mitigate the undue influence of high death rates.

Among various transformation options, the natural logarithm transformation was selected as the most effective method. This choice was supported by improvements observed in both graphical (histogram, Q-Q plot) and formal (Shapiro-Wilk test) assessments of normality. Alternative transformations such as the square root and log-log transformations were also explored for robustness, but the logarithmic transformation provided the best balance between improving distributional properties and preserving interpretability.

Subsequently, the inferential framework was carefully constructed. The plan included building a 95% confidence interval for the mean and conducting a one-sample, left-tailed t-test to evaluate a specific public health claim. A significance level of 0.05 was selected, and both critical value and p-value approaches were incorporated to ensure robust decision-making. These analytical choices reflect a thoughtful and methodical approach to handling complex real-world data while striving for statistically valid and practically meaningful conclusions. The sequence of decisions laid a solid foundation for conducting reliable statistical inference in the later stages of the project.

1.9 Discussion of Results and Conclusions

The comprehensive statistical analysis of cirrhosis death rate data across various states led to several important findings. Initial exploration revealed that the original data was heavily right-skewed, with a few states exhibiting exceptionally high death rates. Visual inspections through boxplots and histograms, coupled with formal normality tests, confirmed that the original dataset did not satisfy the assumptions necessary for classical inferential procedures. The presence of extreme values raised concerns about the influence of outliers, but rather than removing these observations, a decision was made to retain them, acknowledging that they represented real public health phenomena rather than measurement errors.

To address the skewness and stabilize variance, a natural logarithmic transformation was applied. Post-transformation evaluations, including histograms, Q-Q plots, and the Shapiro-Wilk normality test, indicated a substantial improvement in the distribution's symmetry, validating the effectiveness of the transformation. This adjustment allowed for the appropriate application of parametric inferential techniques.

The construction of a 95% confidence interval for the mean death rate revealed that the true mean likely falls between approximately 56.53 and 70.46 deaths, suggesting that cirrhosis mortality is a significant public health concern across the states. Furthermore, hypothesis testing was conducted to investigate whether the true mean death rate was lower than the benchmark of 48.33 deaths, as suggested by health authorities. The results of the one-sample, left-tailed t-test provided strong evidence against the hypothesis that the mean was below 48.33. Instead, the findings indicated that the average cirrhosis death rate is significantly higher than the health department's threshold.

Two key conclusions can be drawn from this study. First, the cirrhosis death rate in the sampled states is higher than previously assumed, signaling a potential need for enhanced public health interventions. Second, the thoughtful application of data transformation techniques, rather than

exclusion of outliers, proved essential for conducting valid statistical inference, demonstrating that real-world health data can be rigorously analyzed while preserving its complexity. These results highlight the importance of robust preprocessing and careful methodological choices in public health data analysis.

First Conclusion: True mean death rate is higher than health department's threshold.

Second Conclusion: Transformation and retaining outliers led to valid, meaningful inference.

References

[1] Mokdad, A. A., Lopez, A. D., Shahraz, S., Lozano, R., Mokdad, A. H., Stanaway, J., ... & Murray, C. J. L. (2014). Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis. *BMC Medicine*, 12(1), 145. https://doi.org/10.1186/s12916-014-0145-y

[2] Scaglione, S., Kliethermes, S., Cao, G., Shoham, D., Durazo, R., Luke, A., & Volk, M. L. (2015). The epidemiology of cirrhosis in the United States: a population-based study. *Journal of Clinical Gastroenterology*, 49(8), 690–696. https://doi.org/10.1097/MCG.0000000000000208

[3] Staiger, D., & Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3), 557–586. https://doi.org/10.2307/2171753

[4] Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*, 15(12). https://doi.org/10.7275/q584-mj60

[5] Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10(2), 486–489. https://doi.org/10.5812/ijem.3505

[6] Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, 38(1), 52–54. https://doi.org/10.5395/rde.2013.38.1.52

[7] Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). John Wiley & Sons.

[8] Wilcox, R. R. (2011). *Introduction to Robust Estimation and Hypothesis Testing* (3rd ed.). Academic Press.

[9] Ezzati, M., & Lopez, A. D. (2004). Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors. *World Health Organization*.

[10] Altman, D. G., & Bland, J. M. (1995). Statistics notes: The normal distribution. *BMJ*, 310(6975), 298. https://doi.org/10.1136/bmj.310.6975.298