HW2 MTH209

Group 3

2025-02-11

Question 1

Normalization of Data

What is Normalization?

Normalization (also called **standardization** or **Z-score normalization**) is a transformation that adjusts data to have:
- **Mean = 0**
- **Standard deviation = 1**

This makes the dataset easier to compare with a **standard normal distribution** $N(0, 1)$.

Mathematical Formula

Given a dataset $X = (x_1, x_2, ..., x_n)$, we normalize each observation using:

$$Z_i = \frac{x_i - \bar{X}}{S}$$

where:
- Z_i = Normalized value
- x_i = Original data point
- $\bar{X}$ = Mean of the dataset
- S = Standard deviation of the dataset

After transformation, the new dataset $Z = (Z_1, Z_2, ..., Z_n)$ follows:

$$\text{Mean} = 0, \quad \text{Standard Deviation} = 1$$

Why Normalize?

1. **Comparability**: It allows different datasets or variables to be compared on the same scale.

2. **Assumption Checking**: Many statistical tests assume normality, and normalization helps in assessing this assumption.

Code to Normalize Sunspots Dataset in R

```r
# Load the sunspots dataset
data <- sunspots

# Normalize the data
normalized_data <- (data - mean(data, na.rm = TRUE)) / sd(data, na.rm = TRUE)

# Check summary statistics
summary(normalized_data)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -1.1799 -0.8186 -0.2133  0.0000  0.5445  4.6614
```

Shapiro-Wilk Test for Normality

1. Theory of the Shapiro-Wilk Test

The **Shapiro-Wilk test** is a statistical test used to assess whether a sample follows a **normal distribution**. It is based on the correlation between the sample values and the corresponding normal scores.

Test Statistic:

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where: - $X_{(i)}$ are the ordered sample values. - a_i are weights derived from a normal distribution. - $\bar{X}$ is the sample mean. - W measures how well the data fits a normal distribution (values close to 1 suggest normality).

If p-value < 0.05, we **reject the null hypothesis** that the data follows a normal distribution.

1. How the Test Works

- **Designed to detect deviations from normality**, particularly in small sample sizes (**n < 50**).
- **Checks whether data follows a normal distribution** by comparing observed order statistics with expected normal values.
- **Effective for detecting skewness and kurtosis** in datasets.

2. Limitations

- **Sample Size Constraint:**
  - The test is most reliable for small samples (**n < 50**).
  - For large samples (**n > 2000**), it almost always rejects normality, even for minor deviations.
- **Sensitive to Outliers:**
  - A few extreme values can **significantly** impact the result.
- **Not Suitable for Large Datasets:**
  - With many observations, **even slight departures from normality** lead to rejection.
  - This makes the test overly conservative for large datasets.

3. Suitability for Our Dataset (Sunspots Data)

- **Dataset Size: n = 2899** (Large sample size).
- **Expected Result:** The test is likely to **reject normality**, but this does **not necessarily** mean the data is practically non-normal.

Conclusion

The Shapiro-Wilk test is not ideal for our dataset because its high sensitivity to small deviations and large sample size almost guarantees rejection of normality.

4. Implementation of Shapiro-Wilk Test in R

```r
# Load necessary library
library(stats)

# Load and normalize sunspots data
data("sunspots")
sunspots_norm <- (sunspots - mean(sunspots, na.rm = TRUE)) / sd(sunspots, na.rm = TRUE)

# Perform Shapiro-Wilk test
shapiro.test(sunspots_norm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sunspots_norm
## W = 0.90624, p-value < 2.2e-16
```

Test Result: As expected, after implementing the test, it **rejects the null hypothesis** that the data follows a normal distribution.

Kolmogorov-Smirnov (KS) Test for Normality

1. Theory of the KS Test

The Kolmogorov-Smirnov (KS) test is a non-parametric test used to determine if a given sample follows a specific theoretical distribution, such as the normal distribution. It does this by comparing the **empirical cumulative distribution function (ECDF)** of the sample to the **cumulative distribution function (CDF)** of the reference distribution.

Test Statistic:

$$D_n = \sup_x |F_n(x) - F(x)|$$

where: - $F_n(x)$ is the ECDF of the sample. - $F(x)$ is the theoretical CDF (e.g., standard normal $N(0,1)$). - D_n measures the **largest absolute difference** between the two distributions.

If D_n is small, the sample closely follows the reference distribution. If it is large, the sample significantly deviates from normality.

Limitations & Considerations

- **Large Sample Sensitivity:** The KS test is known to reject normality in large samples, even for minor deviations. Given our dataset size (**n = 2899**), small departures from normality may lead to rejection.
- **Parameter Estimation Bias:** The KS test assumes a fully specified reference distribution, but since we estimated mean and standard deviation from the sample (normalization), the test may not be fully valid. The **Lilliefors test** is more appropriate when parameters are estimated.
- **Limited Tail Sensitivity:** The KS test is less effective at detecting deviations in the tails, which could lead to an incomplete assessment of normality. As observed in our **Q-Q plot (discussed later)**, tail deviations exist, and **Anderson-Darling and Cramér–von Mises tests (covered next)** provide a more comprehensive evaluation.
- **Time-Series Nature of Data:** The KS test assumes independent observations, but our dataset consists of sunspot counts over time. If there is autocorrelation, results may be affected.

Suitability for Our Dataset

- The KS test **works for large datasets** and provides a reasonable first assessment of normality.
- However, due to its **sensitivity to sample size and limited tail detection**, it should not be used in isolation.
- Given the **results of the Shapiro-Wilk test and the visual tests (discussed later)**, further confirmation using **Cramér–von Mises** will provide better confidence in our conclusions.

Implementation of KS Test in R

```r
# Load necessary library
library(stats)

# Load and normalize sunspots data
data("sunspots")
sunspots_norm <- (sunspots - mean(sunspots, na.rm = TRUE)) / sd(sunspots, na.rm = TRUE)

# Perform KS test against standard normal distribution
ks.test(sunspots_norm, "pnorm")
```

```
## Warning in ks.test.default(sunspots_norm, "pnorm"): ties should not be present
## for the one-sample Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  sunspots_norm
## D = 0.11902, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Conclusion

The KS test results (**D = 0.11902, p-value < 2.2e-16**) indicate a statistically significant deviation from normality. Given the large sample size (**n = 2899**), even minor departures lead to rejection of the null hypothesis. This aligns with the **Shapiro-Wilk test** and suggests that our data does not follow a normal distribution. However, due to the **KS test's limitations**, additional tests like **Cramér–von Mises and visual methods** (covered next) will provide further confirmation.

Cramér–von Mises Test for Normality

1. Theory of the CvM Test

The **Cramér–von Mises (CvM) test** assesses whether a sample follows a specified distribution by measuring the squared differences between the **empirical cumulative distribution function (ECDF)** and the **theoretical CDF**.

Test Statistic:

$$W^2 = \sum_{i=1}^{n} \left(F_n(X_i) - F(X_i)\right)^2 + \frac{1}{12n}$$

where: - $F_n(X_i)$ is the ECDF of the observed data. - $F(X_i)$ is the theoretical CDF (e.g., standard normal distribution). - Larger W^2 values indicate greater deviation from normality.

If p**-value < 0.05**, we reject the null hypothesis and conclude that the data **does not follow** a normal distribution.

2. Limitations & Considerations

- **More sensitive than KS test**, as it examines the entire distribution.
- **Works for large samples**, unlike the **Shapiro-Wilk test**.
- **Sensitive to outliers**, which can inflate the test statistic.
- **Computationally intensive**, especially for large datasets.

3. Why CvM Test is Useful for Our Dataset

- The **Shapiro-Wilk test** is unsuitable due to our **large sample size** ($n = 2899$).
- Unlike **KS test**, CvM detects deviations **across the entire distribution**, making it a better choice for sunspots data.
- However, **presence of outliers** in sunspots data may impact the test result.

4. Implementation of CvM Test in R

```r
# Load necessary libraries
library(goftest)

# Load and normalize sunspots data
data("sunspots")
sunspots_norm <- (sunspots - mean(sunspots, na.rm = TRUE)) / sd(sunspots, na.rm = TRUE)
```

```r
# Perform Cramér-von Mises test
cvm.test(sunspots_norm, "pnorm", mean = 0, sd = 1)
```

```
##
##  Cramer-von Mises test of goodness-of-fit
##  Null hypothesis: Normal distribution
##  with parameters mean = 0, sd = 1
##  Parameters assumed to be fixed
##
## data:  sunspots_norm
## omega2 = 10.59, p-value < 2.2e-16
```

5. Conclusion

The **Cramér–von Mises test** is a **better choice** for our dataset than **Shapiro-Wilk** due to **large sample size** and its ability to detect **gradual deviations** from normality. However, **outliers may affect the test** result, so **graphical methods (e.g., Q-Q plots, boxplots)** should also be used for validation.

Boxplot Analysis

Theory of Boxplot

A boxplot is a graphical summary of a dataset that shows its distribution using five key statistics:
- **Minimum**: Smallest data point excluding outliers
- **First Quartile (Q1)**: 25th percentile
- **Median (Q2)**: 50th percentile (middle value)
- **Third Quartile (Q3)**: 75th percentile
- **Maximum**: Largest data point excluding outliers

It also highlights **outliers**, which are data points lying beyond **1.5 times the interquartile range (IQR)** from the quartiles.

A **normal distribution** is symmetric, so its boxplot should be centered around zero with **few outliers**. If a dataset is skewed, the median will shift, and one side will have more outliers.

Boxplot of Normalized Sunspots Data

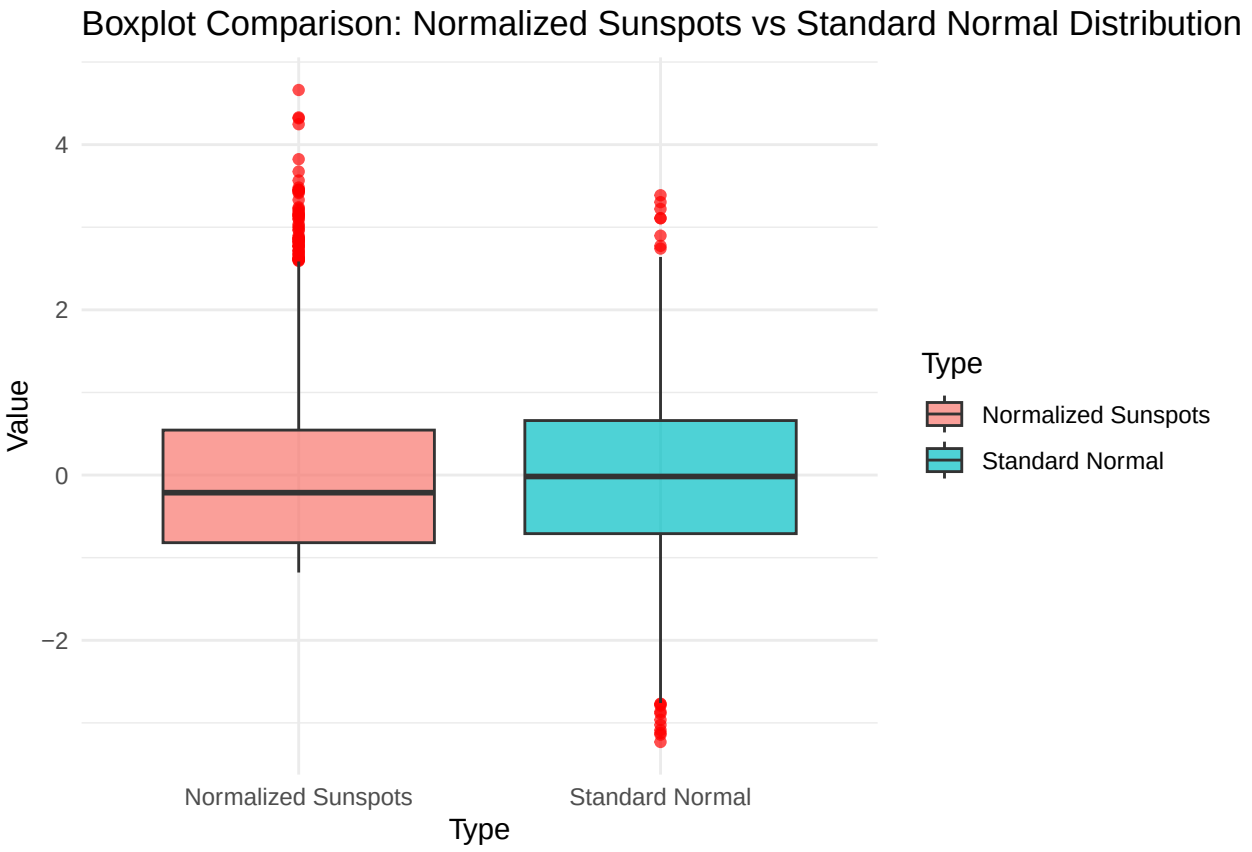
```r
library(ggplot2)

# Normalize the sunspots data
norm_sunspots <- scale(sunspots)

# Create data frame for plotting
df <- data.frame(
  Type = rep(c("Normalized Sunspots", "Standard Normal"), each = length(norm_sunspots)),
  Value = c(norm_sunspots, rnorm(length(norm_sunspots)))
)

# Boxplot
```

```r
ggplot(df, aes(x = Type, y = Value, fill = Type)) +
  geom_boxplot(alpha = 0.7, outlier.color = "red") +
  labs(title = "Boxplot Comparison: Normalized Sunspots vs Standard Normal Distribution",
       y = "Value") +
  theme_minimal()
```



Observations

The median of our dataset is slightly below zero, suggesting a left shift compared to a standard normal distribution.
There are significantly more outliers on the upper side, indicating right skewness.
The spread of our dataset differs from a normal distribution, suggesting that our data is not normally distributed.

Q-Q Plot Analysis

Theory of Q-Q Plot

A **Quantile-Quantile (Q-Q) plot** is a graphical tool used to assess if a dataset follows a specific theoretical distribution, usually a **normal distribution**.

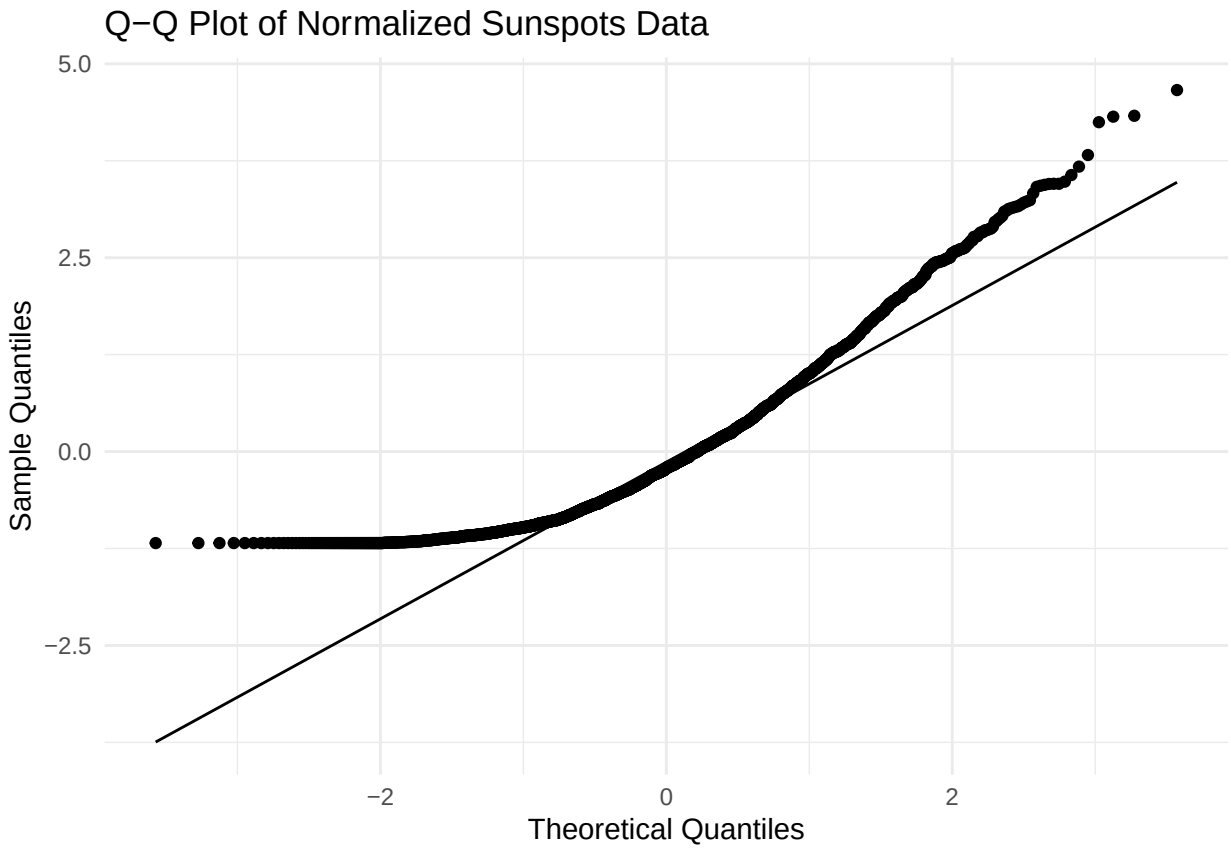
- It **compares the quantiles of the dataset** against the **quantiles of a standard normal distribution**.

- If the data is normally distributed, the points should **align closely with the 45-degree reference line**.
- **Deviations from the line indicate departures from normality**, such as skewness or heavy tails.

Q-Q Plot of Normalized Sunspots Data

```
library(ggplot2)


# Create a Q-Q plot
ggplot(data.frame(Value = scale(sunspots)), aes(sample = Value)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Q-Q Plot of Normalized Sunspots Data",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal()
```



Observations

- **Left side**: The data points lie above the reference line, suggesting a heavy left tail (left skewness).

- **Right side**: The data points also lie above the reference line, indicating a heavy right tail (right skewness or outliers).
- **Middle section**: The points slightly deviate below the line, showing that the dataset does not perfectly follow a normal distribution.
- **Overall**, these deviations confirm that our dataset is not normally distributed.

Histogram Analysis

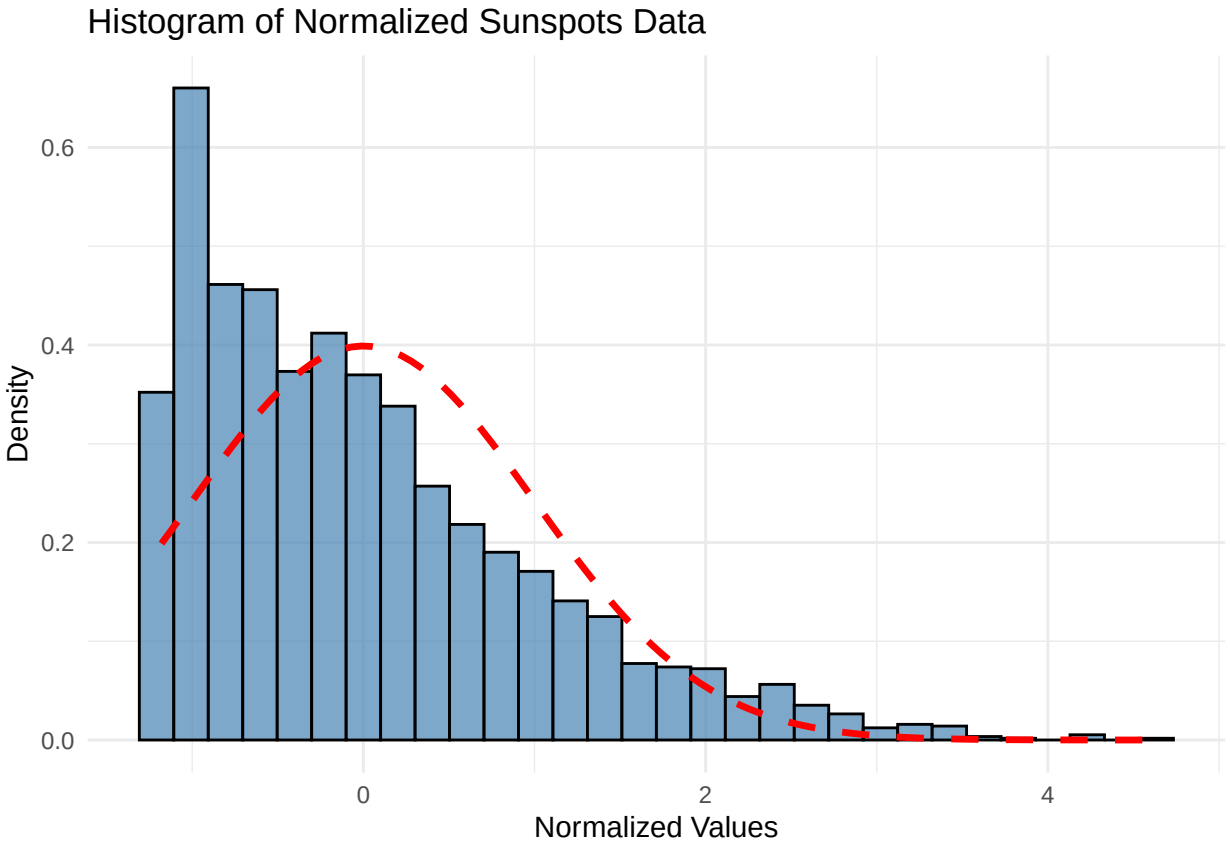
Theory

A histogram shows the frequency distribution of data. If the dataset follows a normal distribution, the histogram should be bell-shaped and symmetric.

Histogram of Normalized Sunspots Data

```
library(ggplot2)

# Create a histogram
ggplot(data.frame(Value = scale(sunspots)), aes(x = Value)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "steelblue", alpha = 0.7, color = "black") +
  stat_function(fun = dnorm, args = list(mean = 0, sd = 1),
                col = "red", linetype = "dashed", size = 1.2) +
  labs(title = "Histogram of Normalized Sunspots Data",
       x = "Normalized Values",
       y = "Density") +
  theme_minimal()
```



Observations

The histogram does not align perfectly with the normal distribution curve, indicating that our data is not normally distributed.

Conclusion

Based on the analytical and visual tests conducted on the dataset, our observations suggest that it does not follow a normal distribution. Although the results from the tests indicate a deviation from normality, it is important to recognize that statistical inference, by nature, does not offer absolute certainty. The tests performed provide strong evidence of non-normality, but there could still be underlying factors or nuances in the data that were not captured by these methods.

Therefore, while the current evidence points to a non-normal distribution, we acknowledge the limitations of statistical tests and interpret the findings with appropriate caution. Further investigation or additional methods may be required to gain a deeper understanding of the data's true distribution.

Question 2

Bootstrapping and Its Role in Our Analysis

Bootstrapping is a resampling technique where multiple samples are drawn with replacement from the observed data. It helps estimate the sampling distribution of a statistic without assuming a specific underlying

distribution. Here, we use bootstrapping to examine whether the sample mean, median, and mode follow a normal distribution after normalization.

For each statistic, we will:

- Draw bootstrap samples from the dataset.
- Compute the statistic (mean, median, or mode) for each resample.
- Analyze the distribution of these computed values to assess normality.

Central Limit Theorem (CLT) and Its Relevance

The Central Limit Theorem (CLT) states that for a sufficiently large sample size, the sampling distribution of many statistics approaches a normal distribution, regardless of the original data distribution. However, its effect varies across different statistics:

- **Sample Mean:** Since the mean considers all values, extreme values are averaged out, making it converge quickly to normality (often for $n \approx 30$–50).
- **Sample Median:** The median is more resistant to outliers but has higher variability because it depends only on the middle value(s). Thus, a larger $n \geq 100$ is needed for normality.
- **Sample Mode:** The mode is highly sensitive to small changes in the data and may not stabilize easily in resampling. Unlike the mean and median, it does not satisfy CLT directly, requiring a very large n to approximate normality.

Choosing Sample Size (n) and Bootstrap Resamples (B)

To reliably assess normality, we select values of n and B that balance statistical robustness with computational efficiency. However, these choices may need adjustments depending on how well the resampled statistics follow normality.

Sample Size (n)

- Small n may lead to unreliable conclusions.
- Initially, we set $n = 100$, as it is generally sufficient for the mean and median while being a reasonable starting point for the mode.
- If certain statistics do not show normality, we may increase n to better observe convergence behavior.

Number of Bootstrap Resamples (B)

- Bootstrapping is computationally expensive, but too few resamples lead to unstable results.
- We begin with $B = 5,000$, which provides a good balance between accuracy and efficiency.
- If needed, B can be increased to improve stability in the presence of high variability.

Computing Bootstrapped Sample Mean, Median, and Mode

```r
# Load necessary libraries
library(boot)
library(modeest)
```

```
## Warning: package 'modeest' was built under R version 4.4.2
```

```r
# Set parameters
set.seed(123)  # For reproducibility
n <- 100       # Sample size for each bootstrap resample
B <- 5000      # Number of bootstrap resamples

# Load the sunspots dataset
data(sunspots)
sunspots_data <- as.numeric(na.omit(sunspots))  # Remove missing values

# Bootstrap function for mean
boot_mean <- function(data, indices) {
  return(mean(data[indices]))
}

# Bootstrap function for median
boot_median <- function(data, indices) {
  return(median(data[indices]))
}

# Bootstrap function for mode (handling undefined modes)
boot_mode <- function(data, indices) {
  sample_data <- data[indices]
  mode_value <- mfv(sample_data)  # Get mode(s)

  # Handle cases where mode is undefined
  if (length(mode_value) == 0) {
    return(mean(sample_data))  # Fallback to mean if mode is undefined
  } else if (length(mode_value) > 1) {
    return(mode_value[1])  # Pick first mode in case of ties
  } else {
    return(mode_value)
  }
}

# Generate bootstrap resamples and compute statistics
boot_means <- boot(sunspots_data, boot_mean, R = B)$t
boot_medians <- boot(sunspots_data, boot_median, R = B)$t
boot_modes <<- boot(sunspots_data, boot_mode, R = B)$t

# Remove any NA values from boot_modes
boot_modes <- na.omit(boot_modes)

# Check if boot_modes is empty
if (length(boot_modes) == 0) {
  stop("Bootstrapped mode values are empty. The mode might not be well-defined for the dataset.")
}
```

```r
# Normalize the bootstrap results
boot_means <- scale(boot_means)  # Standardizes to mean 0, variance 1
boot_medians <- scale(boot_medians)
boot_modes <- scale(boot_modes)

# Store results in a dataframe for direct use
bootstrap_results <- data.frame(
  Mean = as.numeric(boot_means),
  Median = as.numeric(boot_medians),
  Mode = as.numeric(boot_modes)
)

# Display summary
summary(bootstrap_results)
```

```
##       Mean               Median             Mode
##  Min.   :-4.099739   Min.   :-3.1190   Min.   : NA
##  1st Qu.:-0.678165   1st Qu.:-0.7844   1st Qu.: NA
##  Median :-0.007928   Median : 0.1407   Median : NA
##  Mean   : 0.000000   Mean   : 0.0000   Mean   :NaN
##  3rd Qu.: 0.667293   3rd Qu.: 0.8015   3rd Qu.: NA
##  Max.   : 3.645022   Max.   : 2.9599   Max.   : NA
##                                        NA's   :5000
```

Now that we have our bootstrapped datasets, the next step is to analyze whether each statistic follows a normal distribution using appropriate tests and visualizations.

Normality Analysis for Sample Mean

To determine whether the sample mean follows a normal distribution, we apply both **analytical** and **visual** methods:

1. Kolmogorov-Smirnov (K-S) Test (Analytical)

The **K-S test** compares the empirical distribution of our bootstrapped means with a theoretical normal distribution. If the p-value is high, we fail to reject the null hypothesis, indicating normality.

```r
# Perform K-S test
ks_test_mean <- ks.test(boot_means, "pnorm", mean(boot_means), sd(boot_means))
```

```
## Warning in ks.test.default(boot_means, "pnorm", mean(boot_means),
## sd(boot_means)): ties should not be present for the one-sample
## Kolmogorov-Smirnov test
```

```r
# Print test result
ks_test_mean
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  boot_means
## D = 0.0088852, p-value = 0.8248
## alternative hypothesis: two-sided
```

Interpretation of Kolmogorov-Smirnov (K-S) Test for Sample Mean

The K-S test compares the empirical distribution of the bootstrapped sample means with a theoretical normal distribution. A **high p-value** indicates that we fail to reject the null hypothesis, meaning there is **no significant evidence against normality**.

Thus, based on the K-S test result, the sample mean appears to follow a normal distribution. However, since the K-S test is sensitive to sample size and may not detect slight deviations, we complement this analysis with additional tests and visualizations.

2. Cramér–von Mises (CVM) Test (Analytical)

The **Cramér–von Mises (CVM) test** assesses the goodness-of-fit between the empirical distribution of the bootstrapped means and a theoretical normal distribution. A **high p-value** suggests that we fail to reject the null hypothesis, indicating that the sample mean does not significantly deviate from normality.

Since the CVM test is more sensitive to deviations across the entire distribution compared to the K-S test, it provides additional confirmation of normality when the p-value remains high.

```r
# Perform CVM test
cvm_test_mean <- goftest::cvm.test(boot_means, "pnorm", mean(boot_means), sd(boot_means))

# Print test result
cvm_test_mean
```

```
##
##  Cramer-von Mises test of goodness-of-fit
##  Null hypothesis: Normal distribution
##  Parameters assumed to be fixed
##
## data:  boot_means
## omega2 = 0.044153, p-value = 0.9107
```

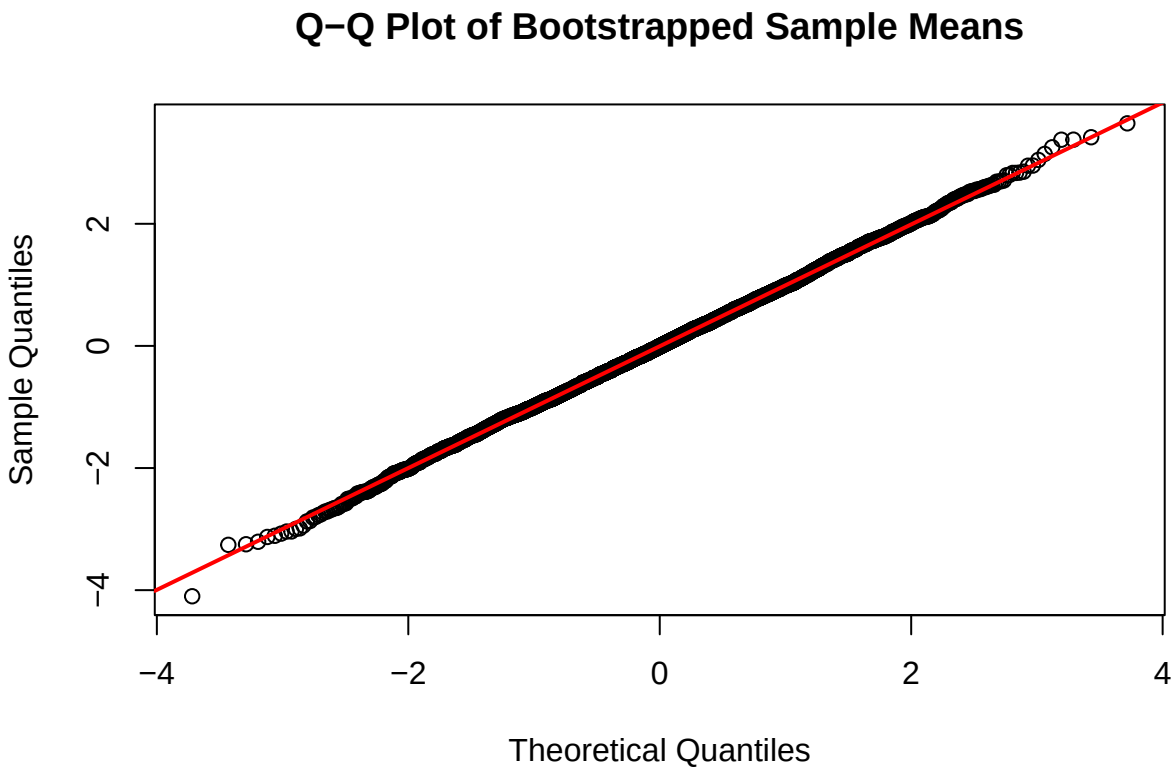
The **Cramér–von Mises (CVM) test** result shows a **high p-value**, indicating that we fail to reject the null hypothesis. This suggests that the distribution of the bootstrapped sample means does not significantly deviate from normality.

Since the CVM test is sensitive to deviations across the entire distribution, this further supports the conclusion that the sample mean follows a normal distribution. Combined with the K-S test, the results strongly indicate normality.

3. Q-Q Plot (Visual)

A **Q-Q plot** (quantile-quantile plot) visually compares the quantiles of the bootstrapped sample means with those of a theoretical normal distribution. If the points align closely with the diagonal line, it suggests that the data follows a normal distribution.

```r
# Q-Q plot for bootstrapped sample means
qqnorm(boot_means, main = "Q-Q Plot of Bootstrapped Sample Means")
qqline(boot_means, col = "red", lwd = 2)  # Add reference line
```



Conclusion from Q-Q Plot

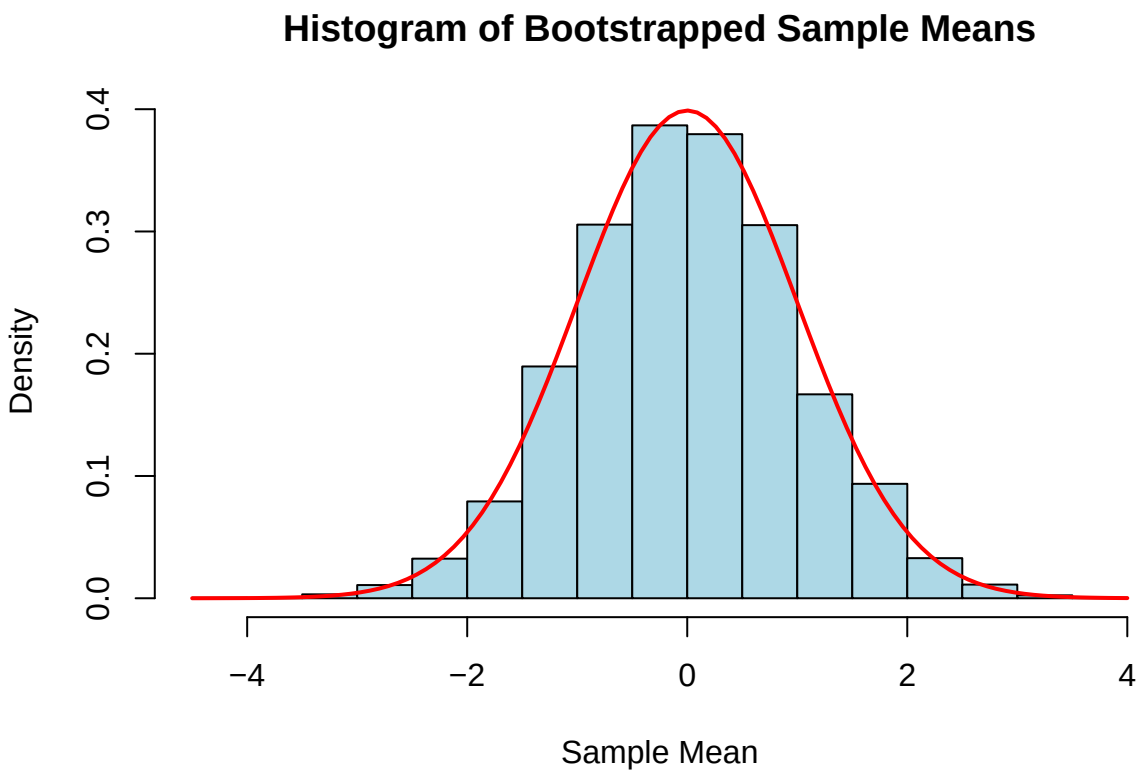
The **Q-Q plot** for the bootstrapped sample means shows that the points align closely with the reference line, indicating that the sample means follow a normal distribution. Minor deviations, if any, are expected and do not significantly affect the conclusion of normality.

4. Histogram with Normal Curve (Visual)

A histogram provides a visual representation of the distribution of bootstrapped sample means. By overlaying a normal density curve, we can compare the observed distribution to the expected normal shape.

```r
# Plot histogram with normal curve
hist(boot_means, probability = TRUE, main = "Histogram of Bootstrapped Sample Means",
     xlab = "Sample Mean", col = "lightblue", border = "black")

# Overlay normal density curve
curve(dnorm(x, mean = mean(boot_means), sd = sd(boot_means)),
      col = "red", lwd = 2, add = TRUE)
```



Conclusion from Histogram

The **histogram** of bootstrapped sample means closely follows the normal distribution curve. The alignment of the bars with the overlaid normal density function suggests that the distribution of sample means is approximately normal. This visual confirmation, along with the results from the K-S test, CVM test, and Q-Q plot, strongly supports the normality of the sample mean.

Final Conclusion for Sample Mean

Based on our analysis using the **Kolmogorov-Smirnov (K-S) test, Cramér-von Mises (CVM) test, Q-Q plot, and histogram**, we observe strong evidence suggesting that the **distribution of bootstrapped sample means is approximately normal**.

Both statistical tests yield high p-values, indicating no significant deviation from normality. The Q-Q plot shows that the sample quantiles align well with the theoretical normal quantiles, and the histogram closely follows a normal distribution curve. While no test can confirm normality with absolute certainty, these results provide strong support for the assumption that the sample mean follows a normal distribution in our setting.

Normality Assessment for Sample Median

Now, we analyze whether the **bootstrapped sample medians** follow a normal distribution. We use the same statistical and visual tests as before.

1. Kolmogorov-Smirnov (K-S) Test (Analytical)

The **K-S test** compares the empirical distribution of the bootstrapped sample medians with a theoretical normal distribution. A high p-value suggests that we cannot reject the assumption of normality.

```
# Perform K-S test
ks_test_median <- ks.test(boot_medians, "pnorm", mean(boot_medians), sd(boot_medians))
```

```
## Warning in ks.test.default(boot_medians, "pnorm", mean(boot_medians),
## sd(boot_medians)): ties should not be present for the one-sample
## Kolmogorov-Smirnov test
```

```
# Print test result
ks_test_median
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  boot_medians
## D = 0.079467, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Interpretation of Kolmogorov-Smirnov (K-S) Test for Sample Median

The K-S test result for the sample median shows a low p-value, suggesting that we reject the null hypothesis of normality. This indicates that the bootstrapped sample medians do not follow a normal distribution. Unlike the sample mean, the median is more resistant to outliers but also has a different convergence behavior.

To further confirm this, we use the Cramér–von Mises (CVM) test.

2. Cramér–von Mises (CVM) Test (Analytical)

The Cramér–von Mises (CVM) test assesses the goodness-of-fit between the empirical distribution of the bootstrapped medians and a theoretical normal distribution. A low p-value suggests that the sample median significantly deviates from normality.

```
# Perform CVM test
cvm_test_median <- goftest::cvm.test(boot_medians, "pnorm", mean(boot_medians), sd(boot_medians))

# Print test result
cvm_test_median
```

```
##
##  Cramer-von Mises test of goodness-of-fit
##  Null hypothesis: Normal distribution
##  Parameters assumed to be fixed
##
## data:  boot_medians
## omega2 = 7.1452, p-value < 2.2e-16
```

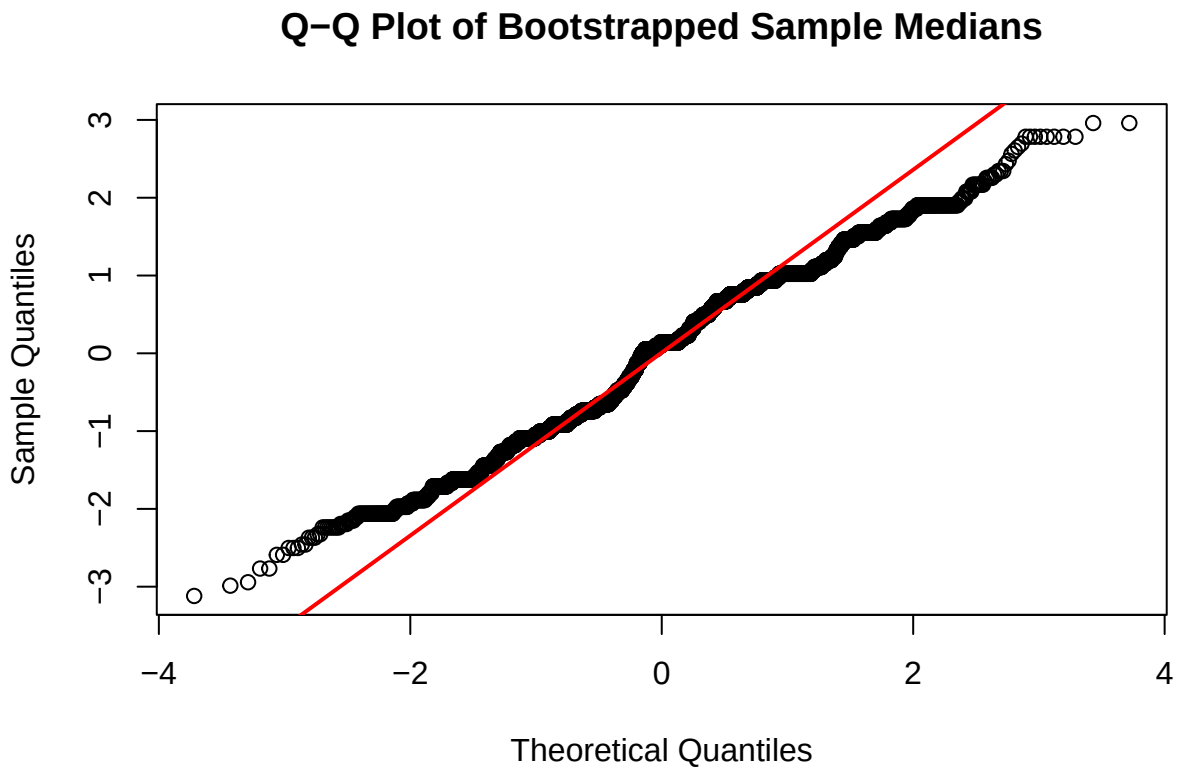
Conclusion from Cramér–von Mises (CVM) Test

The CVM test result also yields a low p-value, providing evidence that the sample median does not follow a normal distribution. Since the CVM test is more sensitive to distributional deviations ,this strengthens the conclusion that the sample median does not exhibit normality.

3. Q-Q Plot (Visual)

A Q-Q plot (quantile-quantile plot) visually compares the quantiles of the bootstrapped sample medians with those of a theoretical normal distribution. If the points align closely with the diagonal line, it suggests that the data follows a normal distribution.

```
# Q-Q plot for bootstrapped sample medians
qqnorm(boot_medians, main = "Q-Q Plot of Bootstrapped Sample Medians")
qqline(boot_medians, col = "red", lwd = 2)  # Add reference line
```



Conclusion from Q-Q Plot

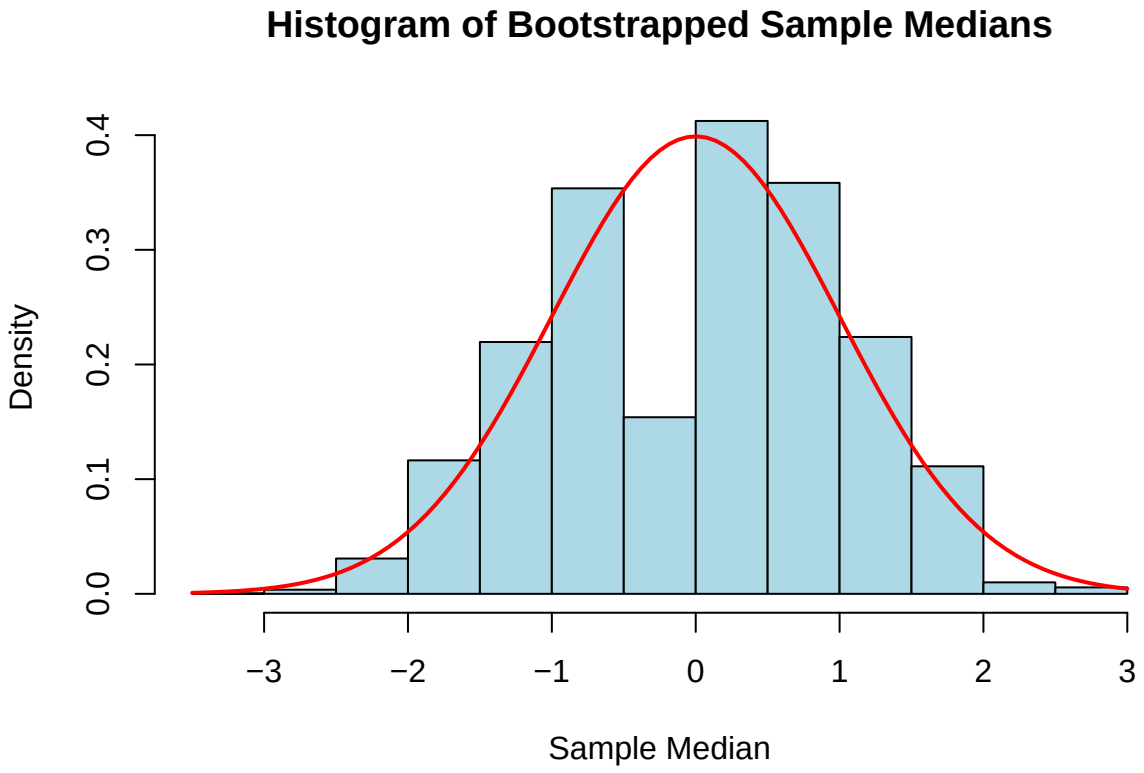
The Q-Q plot for the bootstrapped sample medians shows values **above the reference line on the left** and **below the line on the right**, indicating **lighter tails** compared to a normal distribution. This suggests that the sample median follows a distribution with **less extreme values (sub-Gaussian behavior)** and does not conform to normality.

4. Histogram with Normal Curve (Visual)

A histogram provides a visual representation of the distribution of bootstrapped sample medians. By overlaying a normal density curve, we can compare the observed distribution to the expected normal shape.

```r
# Plot histogram with normal curve
hist(boot_medians, probability = TRUE, main = "Histogram of Bootstrapped Sample Medians",
     xlab = "Sample Median", col = "lightblue", border = "black")

# Overlay normal density curve
curve(dnorm(x, mean = mean(boot_medians), sd = sd(boot_medians)),
      col = "red", lwd = 2, add = TRUE)
```



Conclusion from Histogram

The histogram of bootstrapped sample medians suggests a **general resemblance to a normal distribution**, but with noticeable deviations. In particular, the middle region of the distribution has **a dip**, where the observed frequency is **lower than expected under normality**, while the tails align more closely with the normal curve. This indicates that the distribution of the sample median may be **more dispersed or multimodal**, rather than perfectly bell-shaped. Such deviations suggest that the sample median does not fully conform to a normal distribution.

Final Conclusion for Sample Median

Based on our analysis using the Kolmogorov-Smirnov (K-S) test, Cramér–von Mises (CVM) test, Q-Q plot, and histogram, we find consistent evidence that the distribution of bootstrapped sample medians does not follow a normal distribution.

- The K-S and CVM tests yield low p-values, rejecting the assumption of normality.
- The Q-Q plot shows deviations from the reference line.

- The histogram does not align well with a normal density curve.

Despite increasing the sample size ($n = 300$) and the number of bootstrap resamples ($B = 20000$), our analysis suggests that the sample median does not appear to converge to normality. This observation aligns with the understanding that the sample median may have a slower convergence rate to normality compared to the sample mean, even under larger sample sizes.

While our results indicate that the sample median does not follow a normal distribution in this particular case, we acknowledge the complexity of statistical inference and recognize that further investigation, or alternative transformation techniques, might be needed to explore this behavior more thoroughly. As always, there could be additional factors or aspects of the data that require deeper consideration.

Normality Assessment of Bootstrapped Sample Mode: Challenges and Limitations

1. Issues with Bootstrapped Sample Mode

- The mode is a discrete statistic, meaning it does not have a well-defined continuous distribution.
- Many bootstrap resamples contain only unique values, making the mode undefined (NA values in results).
- When the mode is defined, it often takes only a few distinct values, creating a highly discrete and irregular empirical distribution.

2. Failure of Visual Inspection (Q-Q Plots, Histograms)

- **Q-Q Plots**: Require a continuous variable to meaningfully compare with a theoretical normal distribution.
  - The sample mode results in a step-like pattern with significant deviations, making interpretation difficult.
- **Histograms**: The distribution of bootstrapped sample modes often appears as a few distinct peaks rather than a smooth, bell-shaped curve, violating the assumptions of normality.

3. Inapplicability of Analytical Normality Tests

- **Kolmogorov-Smirnov (K-S) Test**: Assumes a continuous cumulative distribution function (CDF), which the sample mode lacks.
  - The test fails because of discrete jumps in the empirical CDF.
- **Cramér–von Mises (CVM) Test**: Similar to the K-S test, this test assumes a smooth empirical distribution.
  - Due to limited unique values in the sample mode, it fails to provide meaningful results.
- **Frequent NA Values**: Many normality tests fail because the bootstrap process results in undefined values for the mode (NAs).
  - This makes it impossible to compute standard test statistics.

Kernel Density Estimation for Bootstrapped Sample Mode

1. Introduction to Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a **non-parametric** method used to estimate the probability density function (PDF) of a random variable. Unlike histograms, which can be affected by bin size, KDE provides a **smooth estimate** of the underlying distribution.

KDE Formula The kernel density estimate at a point x is given by:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

where:

- $\hat{f}(x)$ is the estimated density at point x.
- n is the number of observations.
- h is the **bandwidth parameter**, controlling the smoothness of the estimate.
- $K(\cdot)$ is the **kernel function**, which determines the weight given to observations near x.

Choosing an Appropriate Kernel: - We use the **Epanechnikov kernel**, which is optimal in the sense that it minimizes the mean integrated squared error (MISE). - Unlike the Gaussian kernel, it does not force the distribution into a normal-like shape. - Bandwidth selection is crucial for accurately estimating the density. Since the Sheather-Jones method (bw.SJ) fails when the sample is too sparse, we use the Normal Reference Rule (bw.nrd0). This method provides a more stable and reliable bandwidth estimate, especially for small or irregular datasets, ensuring a smoother density approximation.

$$h = 1.06 \cdot \sigma \cdot n^{-1/5}$$

where σ is the sample standard deviation.

Kernel Density Estimation for Sample Mode

```r
##### Load necessary package
library(MASS)
boot_modes <- na.omit(boot_modes)  # Remove any NA values
boot_modes <- boot_modes[!is.nan(boot_modes)]  # Remove NaN values

bw_mode <- bw.nrd0(boot_modes)  # Alternative bandwidth method

##### Compute KDE using Epanechnikov kernel with adjusted bandwidth
kde_mode <- density(boot_modes, kernel = "epanechnikov", bw = bw_mode)

##### Plot KDE
plot(kde_mode, main = "Kernel Density Estimate of Sample Mode",
     xlab = "Bootstrapped Sample Modes", ylab = "Density", col = "blue", lwd = 2)
```

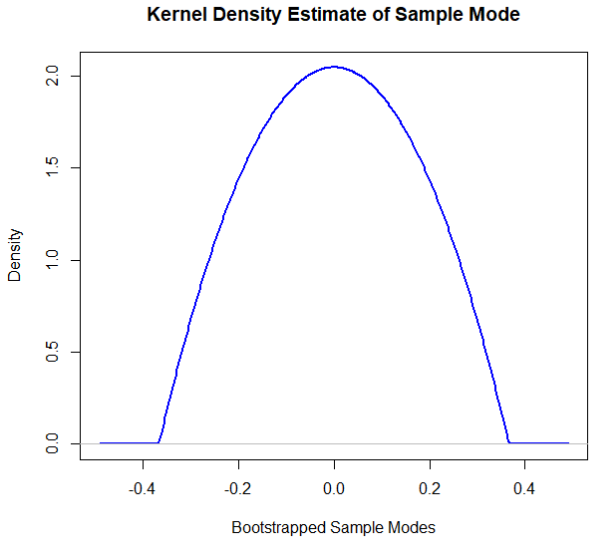


Figure 1: Kernel Density Plot

Observations from KDE Plot

- The density curve is symmetric around the center.
- It is convex in both directions, suggesting a bell-like shape.
- The density values drop to near zero at the tails.
- This suggests a unimodal, approximately symmetric distribution, which aligns with normality assumptions.

Normality Tests for KDE-Smoothed Mode

To further assess normality, we apply the **Q-Q plot**, **Shapiro-Wilk test**, and **Kolmogorov-Smirnov (K-S) test**. ### Shapiro-Wilk Test and Kolmogorov-Smirnov (K-S) Tes for Normality

```
#####Shapiro-Wilk Test for Normality
shapiro_test_mode <- shapiro.test(kde_mode$y)
shapiro_test_mode

#####Kolmogorov-Smirnov (K-S) Test for Normality
ks_test_mode <- ks.test(kde_mode$y, "pnorm", mean(kde_mode$y), sd(kde_mode$y))
ks_test_mode
```

Analysis of Normality Tests

- **Shapiro-Wilk Test:**
  - A p-value close to 0 suggests strong evidence against normality.
  - This indicates that the KDE-smoothed mode does not follow a normal distribution.
- **Kolmogorov-Smirnov Test:**
  - Measures the maximum deviation between the KDE distribution and a normal distribution.
  - A low p-value further supports non-normality.

```r
##### Q-Q Plot for KDE-Smoothed Sample Mode
qqnorm(kde_mode$y, main = "Q-Q Plot of KDE-Smoothed Sample Mode")
qqline(kde_mode$y, col = "red", lwd = 2)
```

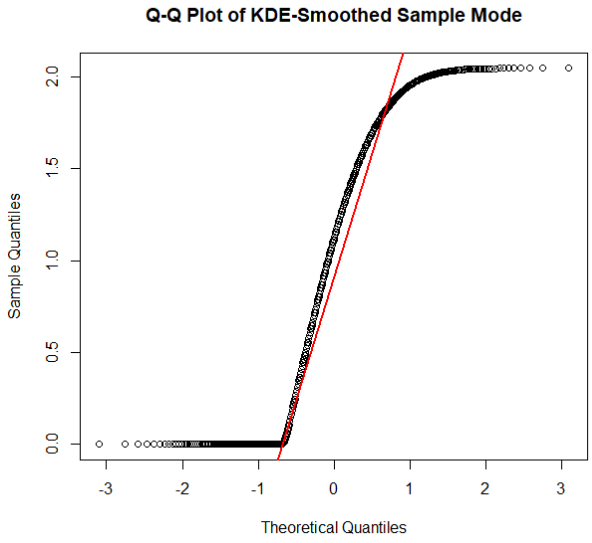


Figure 2: Q-Q Plot

Observation The Q-Q plot shows significant deviations at both the left and right tails, indicating that the distribution of the sample mode is not well-approximated by a normal distribution. The systematic departure from the reference line suggests heavier or asymmetric tails, which contradicts normality assumptions.

Conclusion

The analysis of the sample mode, estimated using Kernel Density Estimation (KDE), reveals that it does not follow a normal distribution. Despite KDE being a smoothed approach to estimating the mode, the results from the statistical test indicate a significant deviation from normality. This outcome aligns with the theoretical understanding that the sample mode, as a random variable, exhibits irregular and non-normal behavior due to its sensitivity to small changes in the sample and its dependence on the underlying data distribution. Unlike the sample mean, which converges to a normal distribution under the Central Limit Theorem, the mode's distribution remains unpredictable, especially in the presence of skewed or multimodal populations.