Assignment-4

Ashish Verma

2024-02-05

PART 1

Section 7.2 Problem #12

Solution(a) Yes, because the sample size is large enough and greater than 40 which is 43 in this case and also sample show reasonable normal pattern.We can use the CLT to approximate sampling distribution go normal.

Solution(b)

```r
# Given values
mean_x <- 1191.6
sd <- 506.6
n <- 43
alpha <- 0.01
t_alpha_over_2 <- qt(1 - alpha / 2, df = n - 1)

# Standard Error
se <- sd / sqrt(n)

# Confidence Interval Calculation
confidence_interval_lower <- mean_x - t_alpha_over_2 * se
confidence_interval_upper <- mean_x + t_alpha_over_2 * se
# Displaying the results
cat("Lower Bound:", confidence_interval_lower, "\n")
```

```
## Lower Bound: 983.1588
```

```r
cat("Upper Bound:", confidence_interval_upper, "\n")
```

```
## Upper Bound: 1400.041
```

Section 7.2 Problem #20

Solution (a)

```r
# Given values
z_value_99_percent = 2.576
p = 0.53
n = 2343
```

```
# Confidence Interval calculation
lower_bound = p - z_value_99_percent * sqrt(p * (1 - p) / n)
upper_bound = p + z_value_99_percent * sqrt(p * (1 - p) / n)

# Displaying the result
cat("The 99% confidence interval is (", round(lower_bound, 4), ",", round(upper_bound, 4), ")\n")
```

```
## The 99% confidence interval is ( 0.5034 , 0.5566 )
```

Solution(b)

```
# Given values
ME = 0.05
p_b = 0.5
z_value_99_percent_b = 2.576

# Sample size calculation
n_b = (z_value_99_percent_b^2) * (p_b * (1 - p_b)) / (ME^2)

# Displaying the result
cat("The required sample size for a margin of error of 0.05 is:", ceiling(n_b), "\n")
```

```
## The required sample size for a margin of error of 0.05 is: 664
```

Section 7.3 Problem #33

Solution(a)

```
# Install and load necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: ’car’
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
# Assuming you have a dataset named 'data'
data <- c(418, 421, 421, 422, 425, 427, 431, 434, 437, 439, 446, 447, 448, 453, 454, 463, 465)
## Descriptive Statistics

# Compute descriptive statistics
summary_stats <- summary(data)
mean_value <- mean(data)
median_value <- median(data)
sd_value <- sd(data)

# Print the results
cat("Summary Statistics:\n", summary_stats, "\n\n")
```

```
## Summary Statistics:
##  418 425 437 438.2941 448 465
```

```r
cat("Mean:", mean_value, "\n")
```

```
## Mean: 438.2941
```

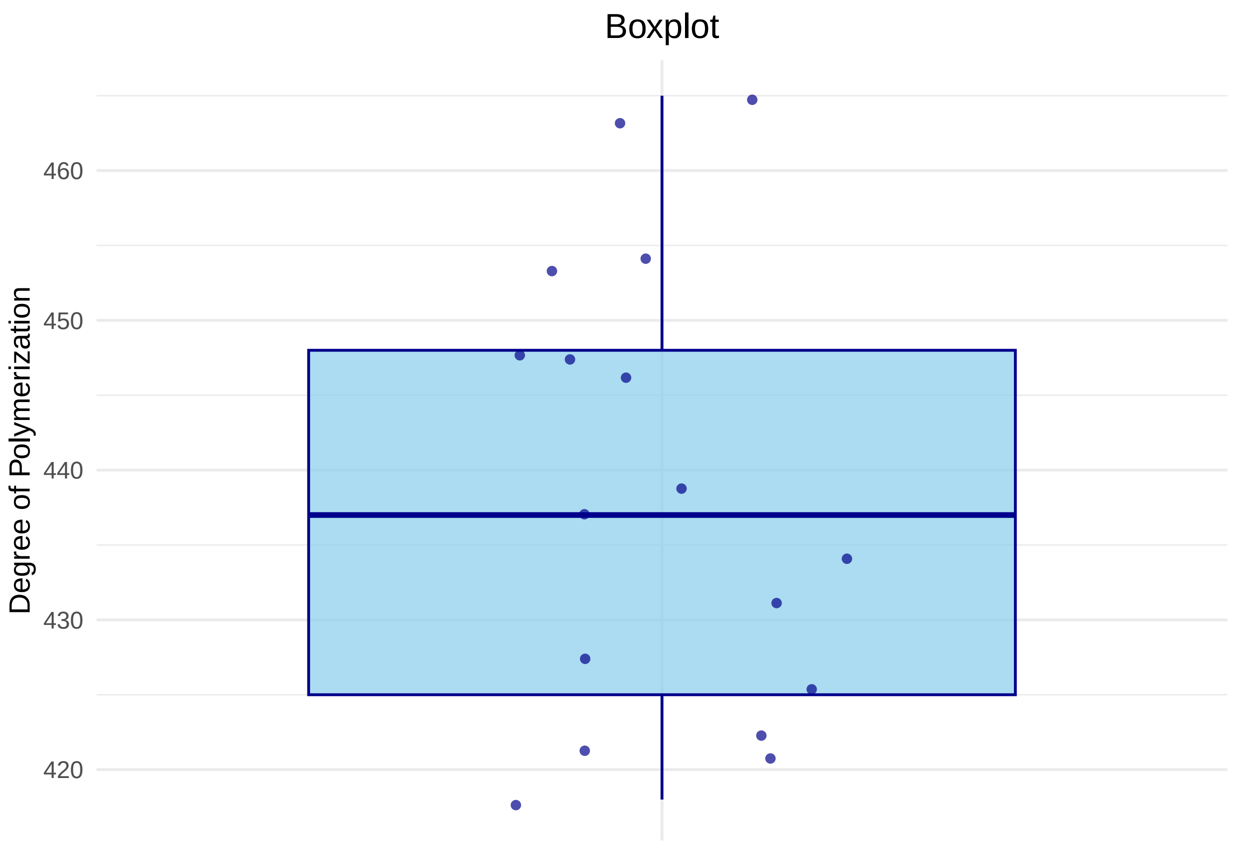
```r
cat("Median:", median_value, "\n")
```

```
## Median: 437
```

```r
cat("Standard Deviation:", sd_value, "\n")
```

```
## Standard Deviation: 15.14416
```

```r
# Create a boxplot
boxplot <- ggplot(data.frame(DegreeOfPolymerization = data), aes(x = "", y = DegreeOfPolymerization)) +
  geom_boxplot(fill = "skyblue", color = "darkblue", alpha = 0.7) +
  geom_jitter(shape = 16, position = position_jitter(0.2), color = "darkblue", alpha = 0.7) +
  theme_minimal() +
  labs(title = "Boxplot",
       x = NULL,
       y = "Degree of Polymerization") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        legend.position = "none")
print(boxplot)
```



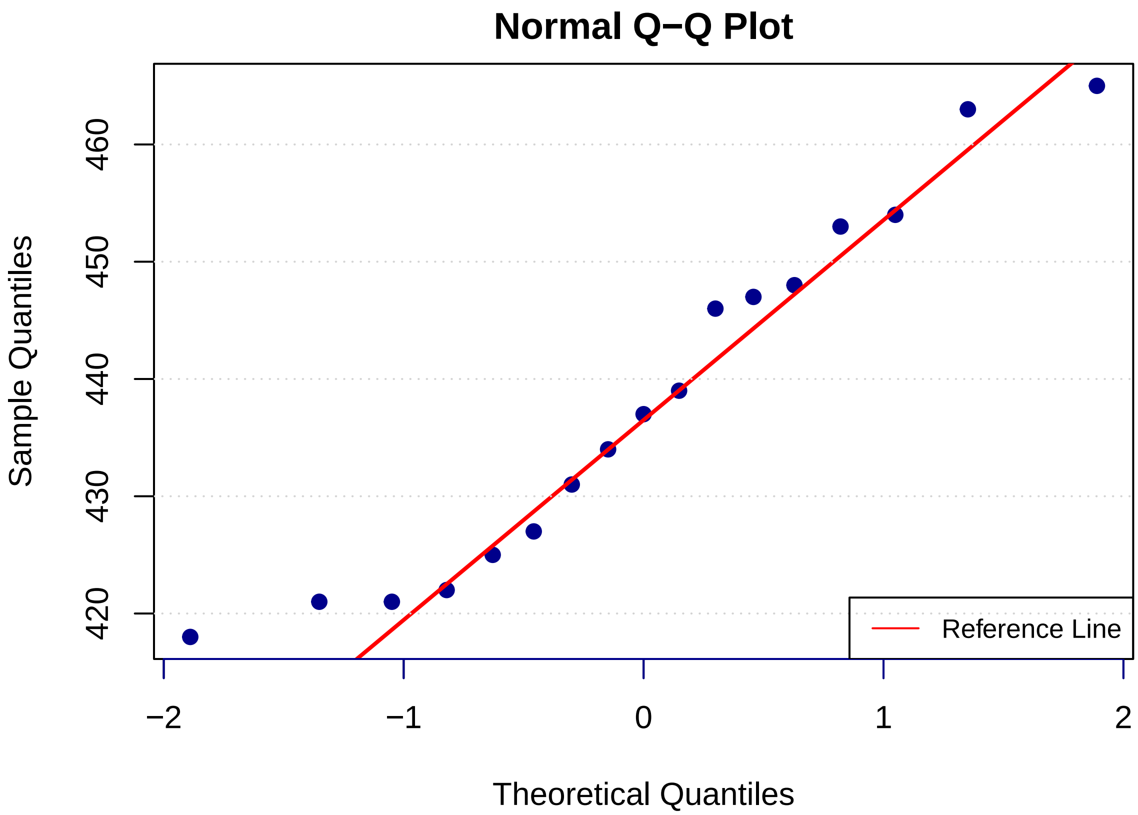
Solution(b)

```
data <- c(418, 421, 421, 422, 425, 427, 431, 434, 437, 439, 446, 447, 448, 453, 454, 463, 465)

# Create a nice probability plot
par(mar = c(5, 5, 2, 2))  # Adjust plot margins
qqnorm(data, col = "darkblue", pch = 20, cex = 1.5)
qqline(data, col = "red", lwd = 2)  # Add a reference line

# Add labels and grid
axis(1, col = "darkblue", col.axis = "darkblue", cex.axis = 1.2, labels = FALSE)
#axis(2, col = "darkblue", col.axis = "darkblue", cex.axis = 1.2)
abline(h = seq(400, 500, by = 10), col = "lightgray", lty = 3)

# Add legend
legend("bottomright", legend = "Reference Line", col = "red", lty = 1, cex = 0.8)
```



Solution(c) # Confidence Interval Calculation

The mean ($\bar{x}$) and standard deviation (s) for the given data are calculated as follows:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{418 + 421 + \ldots + 465}{17} = \frac{7451}{17} = 438.29$$

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(418 - 438.29)^2 + (421 - 438.29)^2 + \ldots + (465 - 438.29)^2}{16}} = \sqrt{229.3456} = 15.14$$

The degrees of freedom is $df = n - 1 = 17 - 1 = 16$.

Confidence Interval Calculation

Using the formula for a 95% confidence interval:

$$95\% CI = \bar{x} \pm t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

From the t-distribution table, for the two-tail test, the critical value corresponding to 16 degrees of freedom at a 0.05 level of significance is $t_{0.025,16} = 2.12$.

The 95% confidence interval for the true average degree of polymerization is calculated as follows:

$$95\% CI = 438.29 \pm 2.12 \left(\frac{15.14}{\sqrt{17}} \right) = (430.5, 446.1)$$

Interpretation

- Yes, the interval suggests that 440 is a plausible value for the true average degree of polymerization since it lies within the interval.
- No, the interval does not suggest that 450 is a plausible value for the true average degree of polymerization since it does not lie within the interval.

PART 2

Solution(1)

Data

```r
total_births <- 2009
female_births <- 1026
male_births <- 983

# Sample proportions
p_female <- female_births / total_births
p_male <- male_births / total_births

# Confidence level
confidence_level <- 0.99

# Z-score for a two-tailed test
z_score <- qnorm((1 + confidence_level) / 2)
# Confidence interval calculation for females
lower_ci_female <- p_female - z_score * sqrt((p_female * (1 - p_female)) / total_births)
upper_ci_female <- p_female + z_score * sqrt((p_female * (1 - p_female)) / total_births)

# Print results
cat("99% Confidence Interval for Female Births: [", lower_ci_female, ", ", upper_ci_female, "]\n")
```

```
## 99% Confidence Interval for Female Births: [ 0.4819744 ,  0.5394293 ]
```

```r
# Confidence interval calculation for males
lower_ci_male <- p_male - z_score * sqrt((p_male * (1 - p_male)) / total_births)
upper_ci_male <- p_male + z_score * sqrt((p_male * (1 - p_male)) / total_births)

# Print results
cat("99% Confidence Interval for Male Births: [", lower_ci_male, ", ", upper_ci_male, "]\n")
```

```
## 99% Confidence Interval for Male Births: [ 0.4605707 ,  0.5180256 ]
```

```r
# Check if the intervals overlap
intervals_overlap <- upper_ci_female >= lower_ci_male && upper_ci_male >= lower_ci_female

# Print results
cat("Do the confidence intervals overlap? ", ifelse(intervals_overlap, "Yes", "No"))
```

```
## Do the confidence intervals overlap?  Yes
```

Solution(2)

```r
# Load the dataset
# Replace 'your_dataset.csv' with the actual file path or URL

url <- "https://archive.ics.uci.edu/static/public/109/wine.zip"

# Specify the destination folder for the downloaded ZIP file
zip_file <- "wine.zip"

# Download the ZIP file
download.file(url, zip_file)

# Unzip the file
unzip(zip_file)

column_names <- c("Region", "Alcohol", "Malic_Acid", "Ash", "Alcalinity_of_Ash", "Magnesium",
                  "Total_Phenols", "Flavanoids", "Nonflavanoid_Phenols", "Proanthocyanins",
                  "Color_Intensity", "Hue", "OD280_OD315_of_Diluted_Wines", "Proline")

# Read CSV with schema
data <- read.csv("wine.data", header = FALSE, col.names = column_names)

# Assuming the dataset has columns 'region' and 'flavanoids'
# Subset the data for region 1 and region 3
data_region1 <- filter(data, Region == 1)
data_region3 <- filter(data, Region == 3)

# Confidence level
confidence_level <- 0.99

# Calculate 99% confidence intervals for the mean value of flavanoids
ci_region1 <- t.test(data_region1$Flavanoids, conf.level = confidence_level)$conf.int
ci_region3 <- t.test(data_region3$Flavanoids, conf.level = confidence_level)$conf.int

# Print results
cat("99% Confidence Interval for Region 1 Flavanoids: [", ci_region1[1], ", ", ci_region1[2], "]\n")
```

```
## 99% Confidence Interval for Region 1 Flavanoids: [ 2.84455 ,  3.120196 ]
```

```r
cat("99% Confidence Interval for Region 3 Flavanoids: [", ci_region3[1], ", ", ci_region3[2], "]\n")
```

```
## 99% Confidence Interval for Region 3 Flavanoids: [ 0.6677307 ,  0.8951859 ]
```

```r
# Check if the intervals overlap
intervals_overlap <- ci_region1[2] >= ci_region3[1] && ci_region3[2] >= ci_region1[1]

# Print results
cat("Do the confidence intervals overlap? ", ifelse(intervals_overlap, "Yes", "No"))
```

```
## Do the confidence intervals overlap?  No
```

Interpretation:

If the confidence intervals overlap, it suggests that there is no statistically significant difference between the mean values of the Flavanoids variable in regions 1 and 3.

If the confidence intervals do not overlap, it may indicate a significant difference between the mean values of the flavanoids variable in the two regions.