Introduction to Statistics for DS
Individual Assignment 2

Zehao Qian

November 6, 2023

1 Question 1

1. D

Analytics: In statistics, inferential statistics is the process used to make decisions and predictions based on sample data, which usually involves inferring population characteristics from sample data or comparing different populations.

2. D

Analytics: The upper quartile is a location measure that represents values that exceed 25% of the data points in the dataset.

3. C

Analytics: The quality control analysts examined the top 70 items produced within an hour, which is a convenient sampling method because it is based on easy-to-sample criteria rather than random selection.

4. B

Analytics: Because office workers want to test the reliability of all new Christmas lights, there are five strings of 30 bulbs, so the total includes all the bulbs. Although only some of the bulbs were actually tested at random, these tests were designed to infer overall reliability.

5. D

Analytics: The confidence interval itself no longer has a probability of 0.95 after calculation, which is the level of confidence in the interval estimation process, not a single calculated interval.

2 Question 2

2.1 Assessment of Normality Through Q-Q Plot Analysis

```r
# clear the console area
cat("\014")
# clear the environment var area
rm(list = ls())
# set work directory
setwd("C:/Users/QianZ/Documents/Project/ISDS/personal-assignment-2/RCode")

# Import data using base R function
sugar_data <- read.csv("./data/Sugar.csv")

# Generate a Q-Q plot using base R functions
qqnorm(sugar_data$x)
qqline(sugar_data$x, col = "red")
```

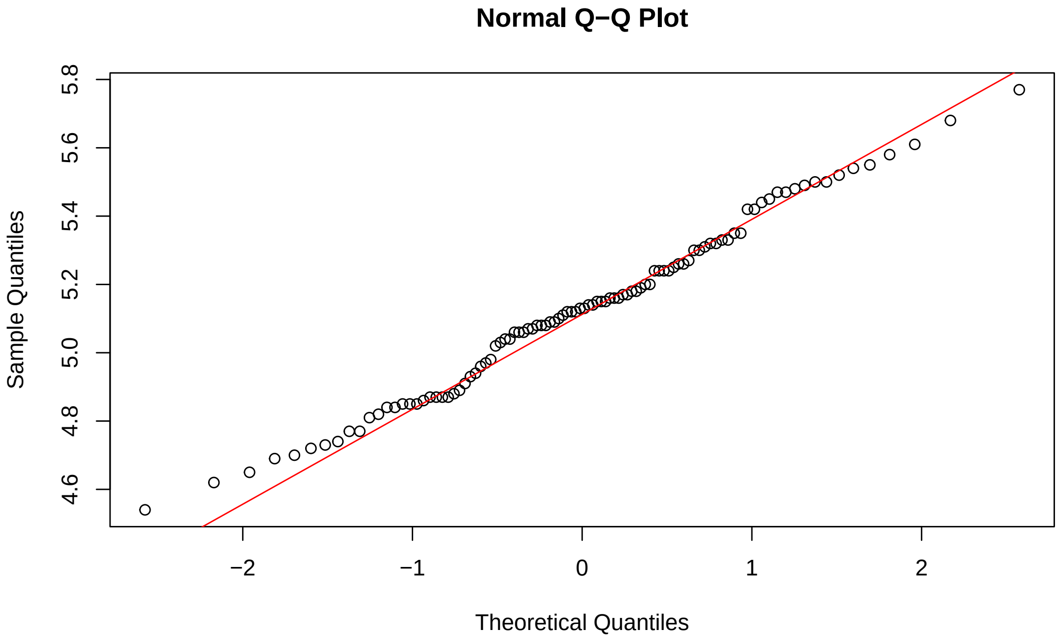

Figure 1: The Q-Q plot of RStudio Output

Comments: From the Q-Q plot (Figure 1), the data points are roughly arranged along the reference line, but there seems to be some deviation at both ends. This means that the data is approximately normally distributed in the central part, but there may be a slight bias in the tail.

2.2 Shapiro-Wilks Test for Normality

```
# Perform the Shapiro-Wilk test on the sugar data
shapiro_test <- shapiro.test(sugar_data$x)

# Output the p-value from the test
shapiro_p_value <- shapiro_test$p.value

# Print the p-value
print(shapiro_p_value)

# Comment on the p-value
if (shapiro_p_value > 0.05) {
    cat("The␣p-value␣is␣greater␣than␣0.05,␣suggesting␣that␣the␣null␣hypothesis␣of␣normality␣cannot␣be␣rejected.")
} else {
    cat("The␣p-value␣is␣less␣than␣or␣equal␣to␣0.05,␣suggesting␣that␣the␣null␣hypothesis␣of␣normality␣can␣be␣rejected.")
}
```

Analytics: The p-value is **0.6553473**, and output ”The p-value is greater than 0.05, suggesting that the null hypothesis of normality cannot be rejected”. This means that we do not have sufficient evidence to reject the null hypothesis of normal distribution, which is consistent with the observation in Q-Q plots that the data are mostly centered around the center, although the tail is slightly off.

2.3 Two-Tailed Hypothesis Test for Sugar Amount

- Zero hypothesis (h_0): the average amount of sugar add equals 5 ml.
- Alternative hypothesis (h_1): the average amount of sugar add is not equal to 5 ml.

The test statistic is calculated and the corresponding critical value is found. Since the sample size is less than 30, a t-distribution is used to determine the cut-off value.

```
# Define the significance level
alpha <- 0.05

```

```
# The null hypothesis (H0): The mean amount of sugar added is 5 ml
# The alternative hypothesis (H1): The mean amount of sugar added is not 5 ml

# Calculate the test statistic (t-value)
t_test <- t.test(sugar_data$x, mu = 5, alternative = "two.sided")

# Output the test statistic
t_value <- t_test$statistic

# Print the value of the test statistic
print(t_value)

# Find the critical t-value for a two-tailed test at the 5% significance level
critical_t_value <- qt(alpha/2, df=length(sugar_data$x)-1, lower.tail=FALSE)

# Print the critical t-value
print(critical_t_value)

# Conclusion based on the t-test
if (abs(t_value) > critical_t_value) {
    cat("Conclusion:␣Reject␣the␣null␣hypothesis.␣There␣is␣a␣significant␣difference␣between␣the␣amount␣of␣sugar␣I␣am␣adding␣and␣the␣5␣ml␣I␣am␣supposed␣to␣be␣adding.")
} else {
    cat("Conclusion:␣Fail␣to␣reject␣the␣null␣hypothesis.␣There␣is␣no␣significant␣difference␣between␣the␣amount␣of␣sugar␣I␣am␣adding␣and␣the␣5␣ml␣I␣am␣supposed␣to␣be␣adding.")
}
```

Conclusion:

- In the two-tailed hypothesis test, The test statistic (t-value) is **4.8981**;
- on a t-distribution with 99 degrees of freedom, the critical t-value of the 5% significance level is **1.984217** ⇒ Because the absolute value of the test statistic is greater than the critical value, we can reject the null hypothesis.
- P-value is much smaller than the significance level α (0.05), which further supports the rejection of the null hypothesis.

3 Question 3

3.1 Hypothesis Test for Seasonal Differences in Beaver Body Temperature

- The null hypothesis (H_0): There is no difference in mean internal body temperature of beavers between November and December.

- The alternative hypothesis (H_1): Beavers have a lower mean internal body temperature in December than in November. (H_1 is complementary to H_0).

I apply a one-tailed t-test since we are specifically interested in whether the mean temperature in December is lower.

Because we are interested in whether the body temperature in December is lower than in November, this is a directional testing question.

```r
# clear the console area
cat("\014")
# clear the environment var area
rm(list = ls())

# Load the datasets
beaver1
beaver2

# Define the significance level
alpha <- 0.05

# Perform the t-test for independent samples
t_test <- t.test(beaver2$temp, beaver1$temp, alternative = "less", var.equal = TRUE)

# The value of the test statistic
t_value <- t_test$statistic

# A bound on the p-value for the test
p_value_bound <- t_test$p.value

# Print the test statistic and the bound on the p-value
print(t_value)
print(p_value_bound)

```

```
# Conclusion
if (p_value_bound < alpha) {
    cat("Conclusion:␣Reject␣the␣null␣hypothesis.␣There␣is␣
        evidence␣to␣suggest␣that␣beavers␣have␣a␣lower␣mean␣
        internal␣body␣temperature␣in␣December␣than␣in␣November."
        )
} else {
    cat("Conclusion:␣Fail␣to␣reject␣the␣null␣hypothesis.␣There␣
        is␣not␣sufficient␣evidence␣to␣suggest␣that␣beavers␣have␣
        a␣lower␣mean␣internal␣body␣temperature␣in␣December␣than␣
        in␣November.")
}
```

Output:

- The test statistic is **-15.235** .
- A bound on the p-value for the test is **4.224052e-31**.
- **Conclusion:** Reject the null hypothesis. There is evidence to suggest that beavers have a lower mean internal body temperature in December than in November.