GR5291 Advanced Data Analysis Problem Set 2

Francis Zhang

September 20, 2024

Question

Consider the ToothGrowth data in R, concerning the Effect of Vitamin C on Tooth Growth in Guinea Pigs.

1.Ignore the data for Dose=2.0, and determine whether there is a significant difference in the mean "len" between the two groups (i.e., OJ vs VC), combining the data for Doses =0.5 and 1.0:

a.Using a parametric procedure

b.Using a non-parametric procedure

c.Discuss the assumption underlying each of the analyses, their validity, and any remedial measures to be taken.

Overview

I will be taking the ToothGrowth data and do a basic EDA on the data, then I will use t.test as a parametric procedure and wilcox.test as a non-parametric procedure to perform hypothesis testing for the effectiveness of the supplement types on tooth growth length under the respective dose levels. And discuss the assumption underlying each of the analyses, their validity, and any remedial measures to be taken.

Solution

Basic EDA

```
# Load the ToothGrowth data
data("ToothGrowth")
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

The data has 60 rows and 3 columns

```
unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

There are three unique values for dose.

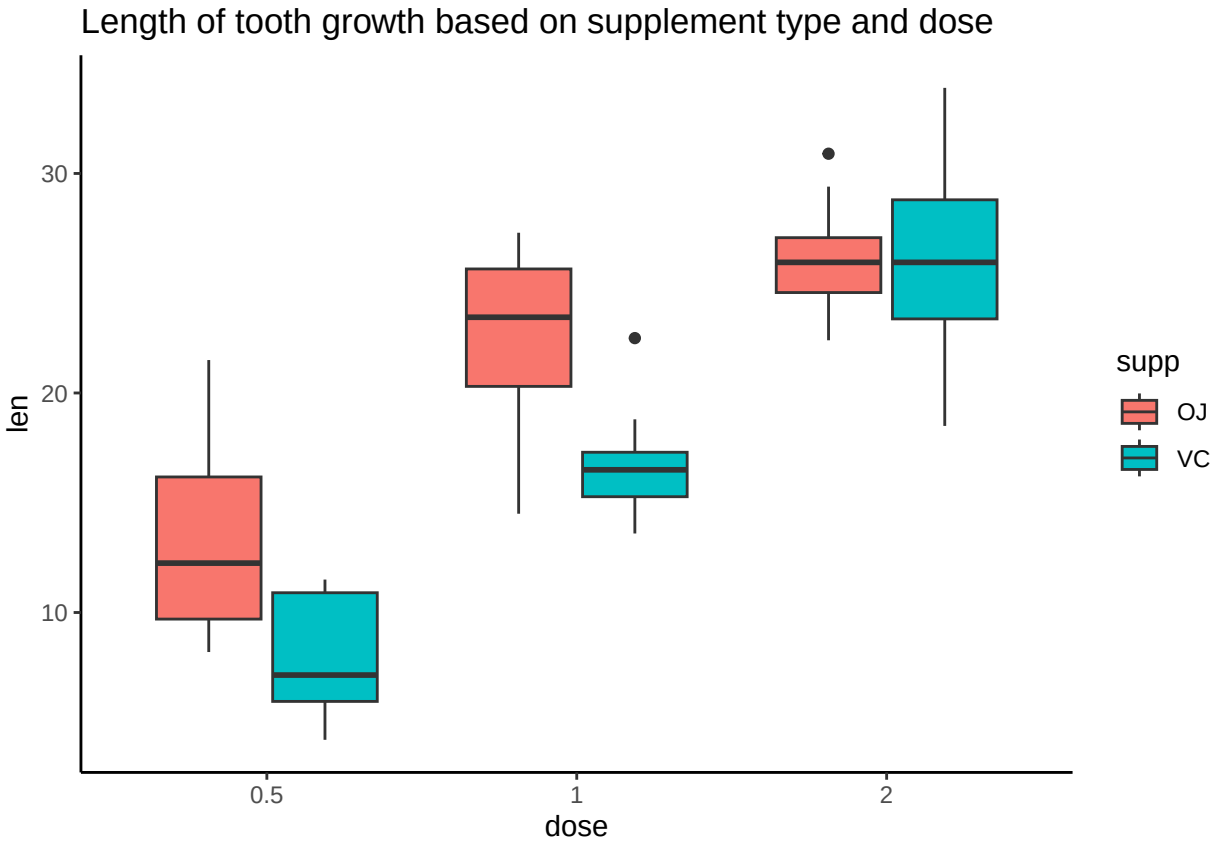
```
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

Plotting box plot based on supplement type and dose

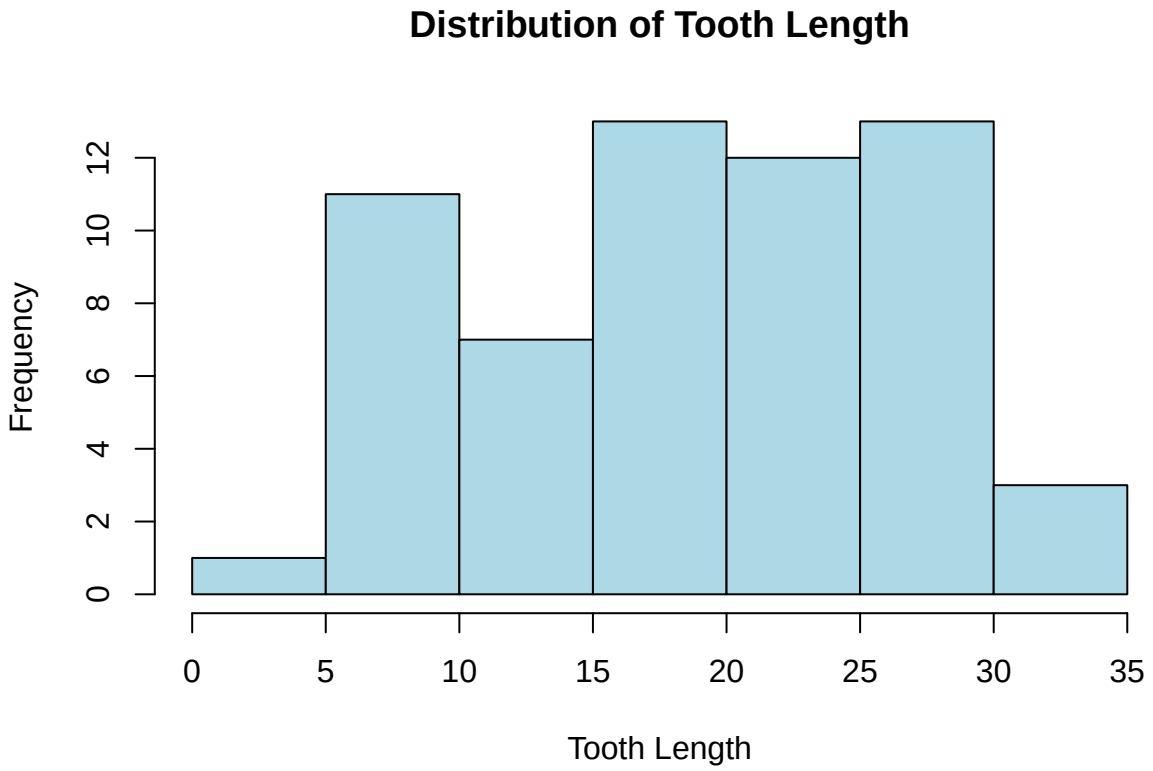
```
ToothGrowth$dose <- factor(ToothGrowth$dose)
library(ggplot2)

ggplot(ToothGrowth, aes(x = dose, y = len, fill = supp)) +
  geom_boxplot() +
  ggtitle("Length of tooth growth based on supplement type and dose") +
  theme_classic()
```




```
hist(ToothGrowth$len,
     main = "Distribution of Tooth Length",
     xlab = "Tooth Length",
```

```
     ylab = "Frequency",
     col = "lightblue",
     border = "black")
```



1. Filter Data for Doses 0.5 and 1.0, Exclude Dose 2.0

```
# Filter the data for doses 0.5 and 1.0, excluding dose 2.0
filtered_data <- subset(ToothGrowth, dose != 2.0)
filtered_data
```

```
##     len supp dose
## 1   4.2   VC  0.5
## 2  11.5   VC  0.5
## 3   7.3   VC  0.5
## 4   5.8   VC  0.5
## 5   6.4   VC  0.5
## 6  10.0   VC  0.5
## 7  11.2   VC  0.5
## 8  11.2   VC  0.5
## 9   5.2   VC  0.5
## 10  7.0   VC  0.5
## 11 16.5   VC    1
## 12 16.5   VC    1
## 13 15.2   VC    1
## 14 17.3   VC    1
## 15 22.5   VC    1
## 16 17.3   VC    1
## 17 13.6   VC    1
## 18 14.5   VC    1
```

```
## 19 18.8   VC    1
## 20 15.5   VC    1
## 31 15.2   OJ  0.5
## 32 21.5   OJ  0.5
## 33 17.6   OJ  0.5
## 34  9.7   OJ  0.5
## 35 14.5   OJ  0.5
## 36 10.0   OJ  0.5
## 37  8.2   OJ  0.5
## 38  9.4   OJ  0.5
## 39 16.5   OJ  0.5
## 40  9.7   OJ  0.5
## 41 19.7   OJ    1
## 42 23.3   OJ    1
## 43 23.6   OJ    1
## 44 26.4   OJ    1
## 45 20.0   OJ    1
## 46 25.2   OJ    1
## 47 25.8   OJ    1
## 48 21.2   OJ    1
## 49 14.5   OJ    1
## 50 27.3   OJ    1
```

```r
# Perform a t-test for the two supplement groups (OJ vs VC)
t_test_result <- t.test(len ~ supp, data = filtered_data)

# View the result
t_test_result
```

a. Using a parametric procedure

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.0503, df = 36.553, p-value = 0.004239
## alternative hypothesis: true difference in means between group OJ and group VC is not equal to 0
## 95 percent confidence interval:
##  1.875234 9.304766
## sample estimates:
## mean in group OJ mean in group VC
##           17.965           12.375
```

```r
# Perform a Wilcoxon Rank-Sum test
wilcox_test_result <- wilcox.test(len ~ supp, data = filtered_data)
```

b. Using a non-parametric procedure

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties
```

```r
# View the result
wilcox_test_result
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  len by supp
## W = 295, p-value = 0.01053
## alternative hypothesis: true location shift is not equal to 0
```

c. Discussion of Assumptions for Each Analysis

a. Parametric Procedure (t-test)

• Hypothesis: The null hypothesis for the t-test is that there is no difference in the mean tooth length ("len") between the two groups (OJ and VC). The alternative hypothesis is that the true difference in means is not equal to zero.

• Result: Based on the t-test output, we obtained a t-statistic of 3.0503 and a p-value of 0.004239. Since the p-value is below the common significance threshold of 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in mean tooth length between the OJ and VC groups.

• Validity:

• The validity of this test depends on the normality of the data within each group. If the data deviate substantially from a normal distribution, the results of the t-test may not be reliable.

• The small p-value (0.004239) suggests that there is a strong statistical difference between the two groups, assuming normality holds.

• Remedial Measures:

• If normality is questionable, transformations (like log or square-root) can be applied to the data to stabilize variances and better meet the normality assumption.

• Alternatively, if normality is violated, a non-parametric test, such as the Wilcoxon Rank-Sum test, could be used to confirm the results.

b. Non-Parametric Procedure (Wilcoxon Rank-Sum Test)

• Hypothesis: The null hypothesis for the Wilcoxon Rank-Sum test is that the distribution of tooth length is the same in both the OJ and VC groups. The alternative hypothesis is that the distribution of tooth lengths differs between the two groups.

• Result: The test produced a W statistic of 295 and a p-value of 0.01053. Since the p-value is below 0.05, we reject the null hypothesis, indicating that there is a statistically significant difference in the distribution of tooth lengths between the two groups.

• Validity:

• The Wilcoxon Rank-Sum test is more robust to non-normal data, making it a valid choice even if the t-test's assumptions are violated.

• The presence of ties (same values in both groups) prompted a warning. While this can slightly affect the test's power, the p-value of 0.01053 still suggests a significant difference between the groups.

• Remedial Measures:

• The ties in the data prevent the computation of an exact p-value. However, the approximation used here is generally reliable, especially given the moderate sample size.

• If ties are frequent and significantly impact the result, alternative non-parametric methods, such as the Kruskal-Wallis test, could be considered, although in this case, the Wilcoxon test result seems valid.

c. Conclusion:

- Both the t-test and Wilcoxon Rank-Sum test provide evidence of a significant difference in tooth length between the OJ and VC groups. If the normality assumption of the t-test is violated, the Wilcoxon test serves as a reliable alternative, especially given its robustness to ties and non-normal distributions. Based on both tests, we can confidently conclude that the two groups differ in tooth length, with strong evidence provided by the low p-values in both analyses.