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.

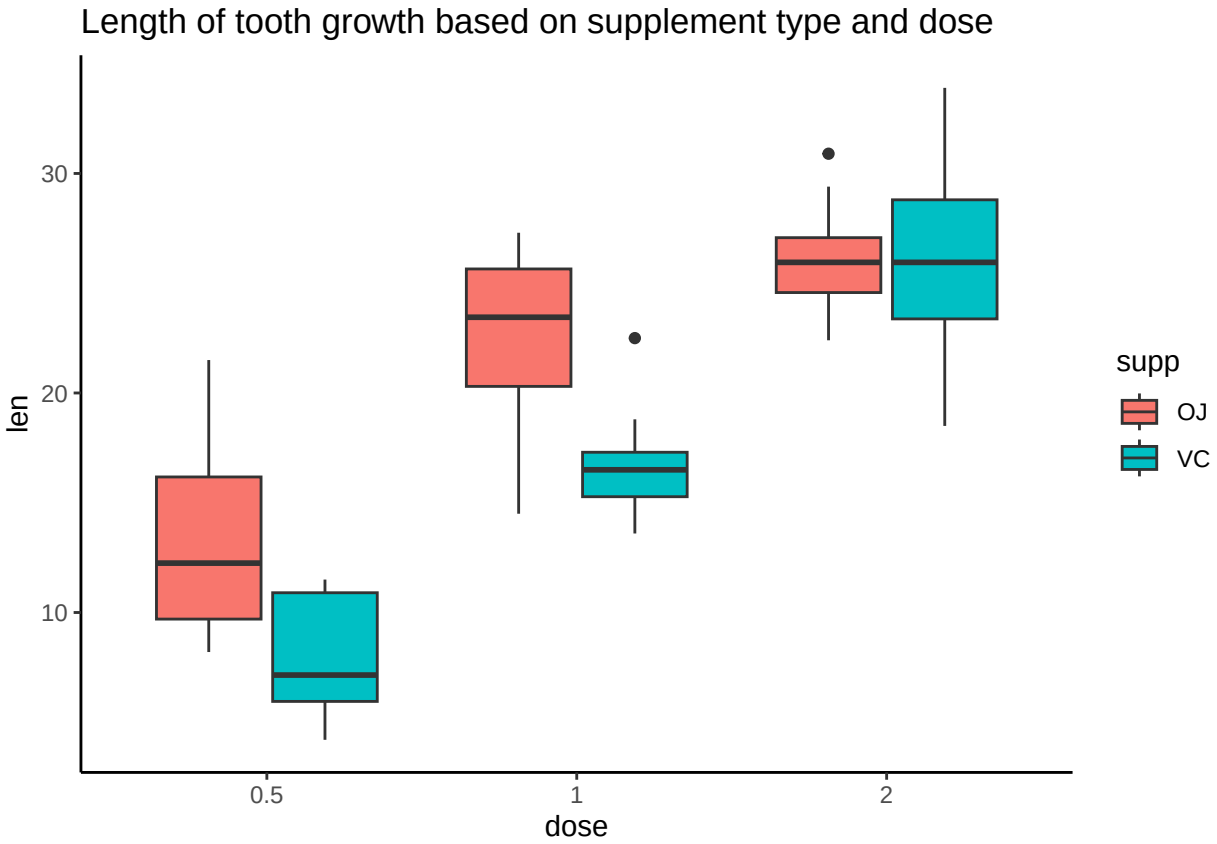
```
summary(ToothGrowth)
```

```
##       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

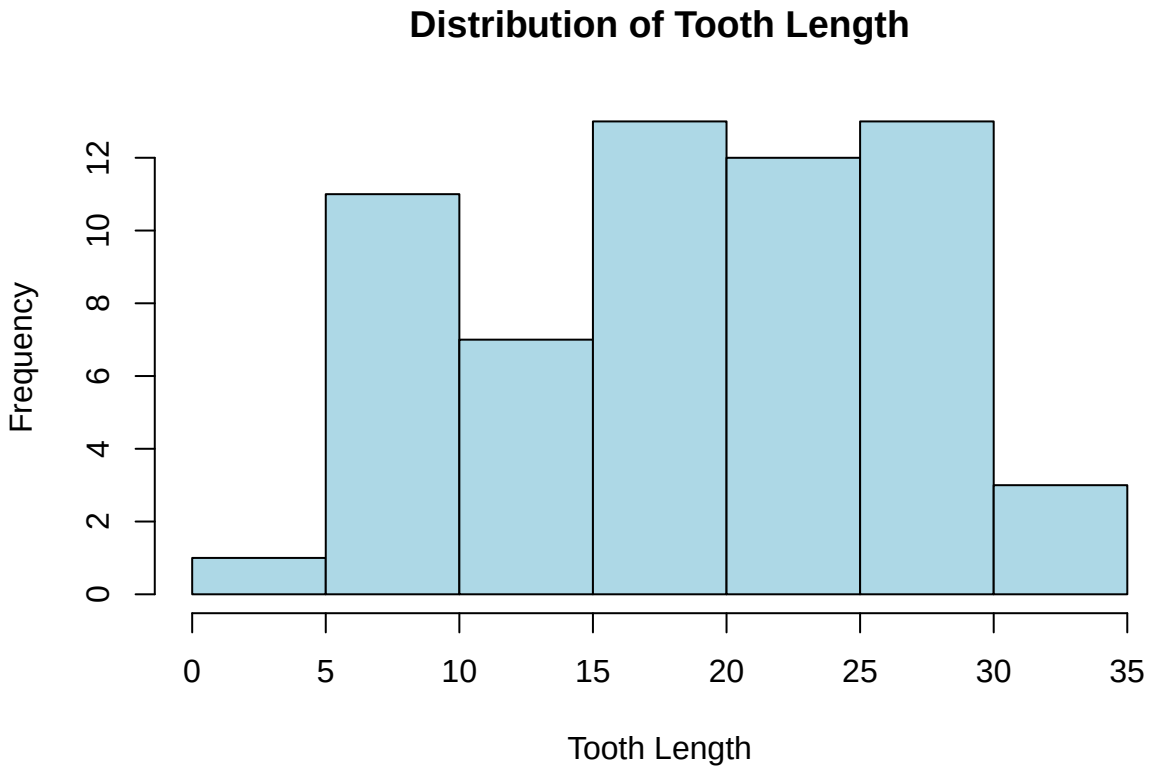
```
ToothGrowth$dose <- factor(ToothGrowth$dose)
library(ggplot2)

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)

- Assumptions:

1. Normality: The t-test assumes that the distribution of the dependent variable (in this case, "len") within each group (OJ and VC) follows a normal distribution.
2. Equal Variances: In a standard t-test (not Welch's t-test), it is assumed that the variances in the two groups are equal (homogeneity of variances). However, in this case, you used Welch's t-test, which does not assume equal variances and is more robust when the variances are unequal.
3. Independence: The observations within each group should be independent of one another. • Validity:
4. The normality assumption can be checked using Q-Q plots or formal normality tests (e.g., the Shapiro-Wilk test). If the data significantly deviates from normality, the validity of the t-test may be compromised.
5. Welch's t-test is robust to unequal variances, so we don't need to worry about the equal variances assumption in this case.
6. Independence is assumed based on the experimental design, which should ensure that each guinea pig is independent of others in terms of their tooth growth. • Remedial Measures:
7. If the normality assumption is violated, you could either transform the data (e.g., log or square-root transformation) to make it more normally distributed or use a non-parametric test (like the Wilcoxon Rank-Sum test, as performed).
8. Since Welch's t-test handles unequal variances, no further action is needed regarding this assumption.

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

- Assumptions:

1. Independence: The Wilcoxon Rank-Sum test assumes that the observations in each group are independent.
2. Ordinality or Continuous Data: The test assumes that the dependent variable (tooth length) can be meaningfully ranked.
3. No Assumption of Normality: Unlike the t-test, the Wilcoxon test does not require the data to follow a normal distribution, making it more robust in situations where normality is questionable. • Validity:
4. The independence assumption holds if the guinea pigs were randomly assigned to the two supplement groups.
5. Ties (equal values of tooth length between groups) can lead to the warning that you encountered. This does not invalidate the test but indicates that an exact p-value cannot be computed due to the presence of tied ranks, and an approximate method is used instead.
6. Since the Wilcoxon test does not assume normality, it is valid even if the normality assumption for the t-test is violated. • Remedial Measures:
7. The ties warning is a result of identical tooth length values in both groups (e.g., both OJ and VC having a length of 16.5). This can be ignored in most cases because the approximation used is usually sufficient, especially with larger sample sizes.
8. If the number of ties is large and you are concerned about the approximation, you can either use a different non-parametric test or report the limitations in the interpretation of results.

Summary:

- The t-test is valid if the data is normally distributed (or approximately so), and since you used W
- The Wilcoxon Rank-Sum test is more robust when normality is violated and is a good alternative if t
- In both cases, independence of observations is key, and the warning in the Wilcoxon test regarding