Statistical Interference - Project, part 2

Introduction

In this report the ToothGrowth data in the R datasets package is analyzed. First, some basic exploratory data analyses are performed and a basic summary of the data is provided. Then, confidence intervals and hypothesis tests are used to compare tooth growth by supp and dose.
The dataset's variable `len` contains the length of odontoblasts (teeth) in each of 10 guinea pigs at each of three dose levels of Vitamin C (0.5, 1, and 2 mg, values are in variable named `dose`) with each of two delivery methods (orange juice or ascorbic acid, values OJ and VC in variable `supp`).

The used R code is included in this report so that every part of it is reproducible.

Exploratory Data Analysis and Data Summary

After loading the data we print out the structure of the data frame, i.e. the number of observations and variables and their data types. The combined counts of the `supp` and `dose` variables are printed using the *table* function, and an overview of the main statistics using the *summary* function:

```
set.seed(1)
library(ggplot2)

data(ToothGrowth)
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 ...
```

```
table(ToothGrowth$supp, ToothGrowth$dose)
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

```
summary(ToothGrowth)
```

```
##       len        supp         dose      
##  Min.   : 4.2   OJ:30   Min.   :0.50  
##  1st Qu.:13.1   VC:30   1st Qu.:0.50  
##  Median :19.2           Median :1.00  
##  Mean   :18.8           Mean   :1.17  
##  3rd Qu.:25.3           3rd Qu.:2.00  
##  Max.   :33.9           Max.   :2.00
```

Data Transformations

We convert the `dose` variable to a factor variable as its numeric interpretation are not of interest.

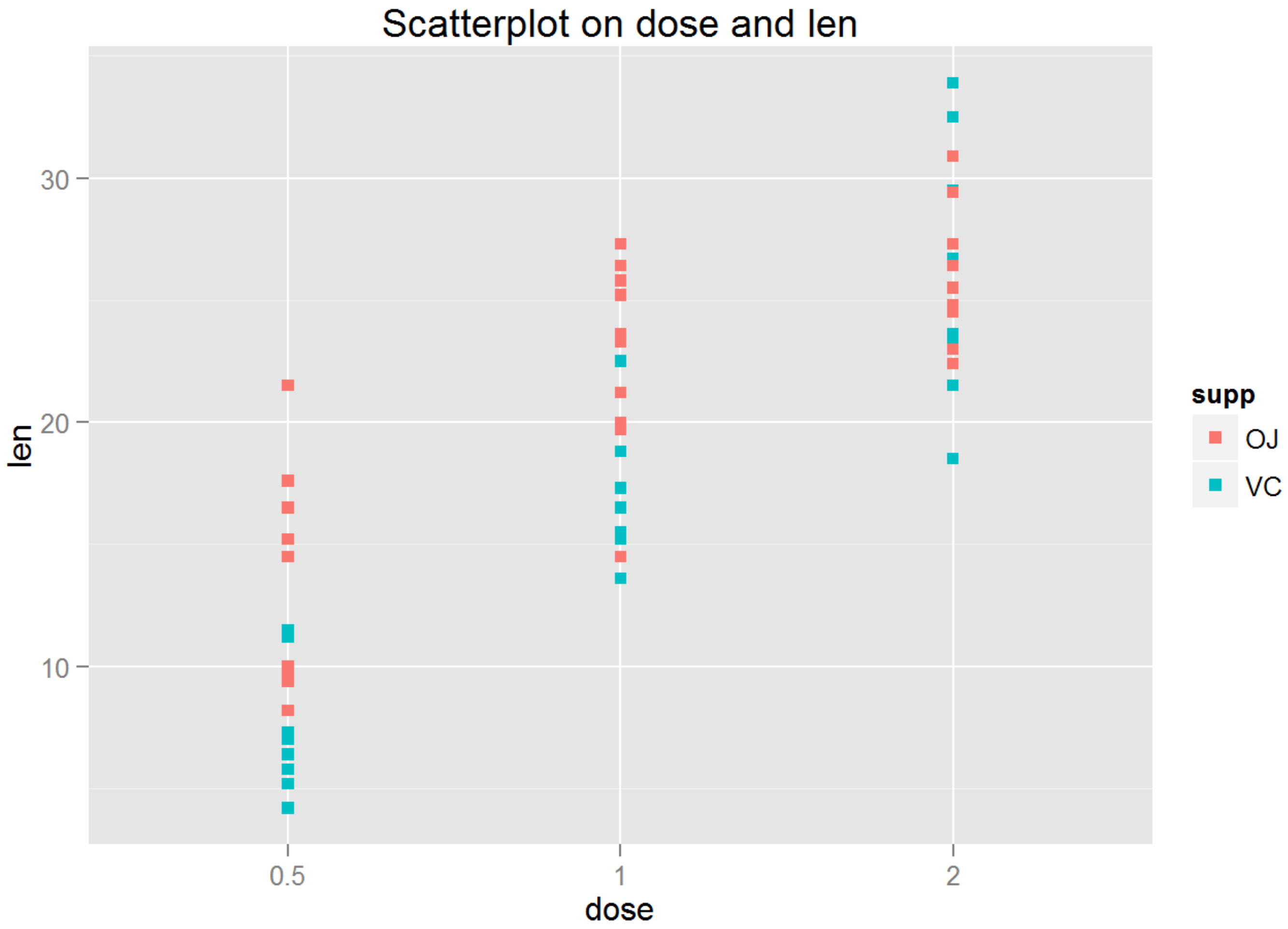
```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
summary(ToothGrowth)
```

```
##       len        supp     dose   
##  Min.   : 4.2   OJ:30   0.5:20  
##  1st Qu.:13.1   VC:30   1  :20  
##  Median :19.2           2  :20  
##  Mean   :18.8                   
##  3rd Qu.:25.3                   
##  Max.   :33.9
```

Data Plots

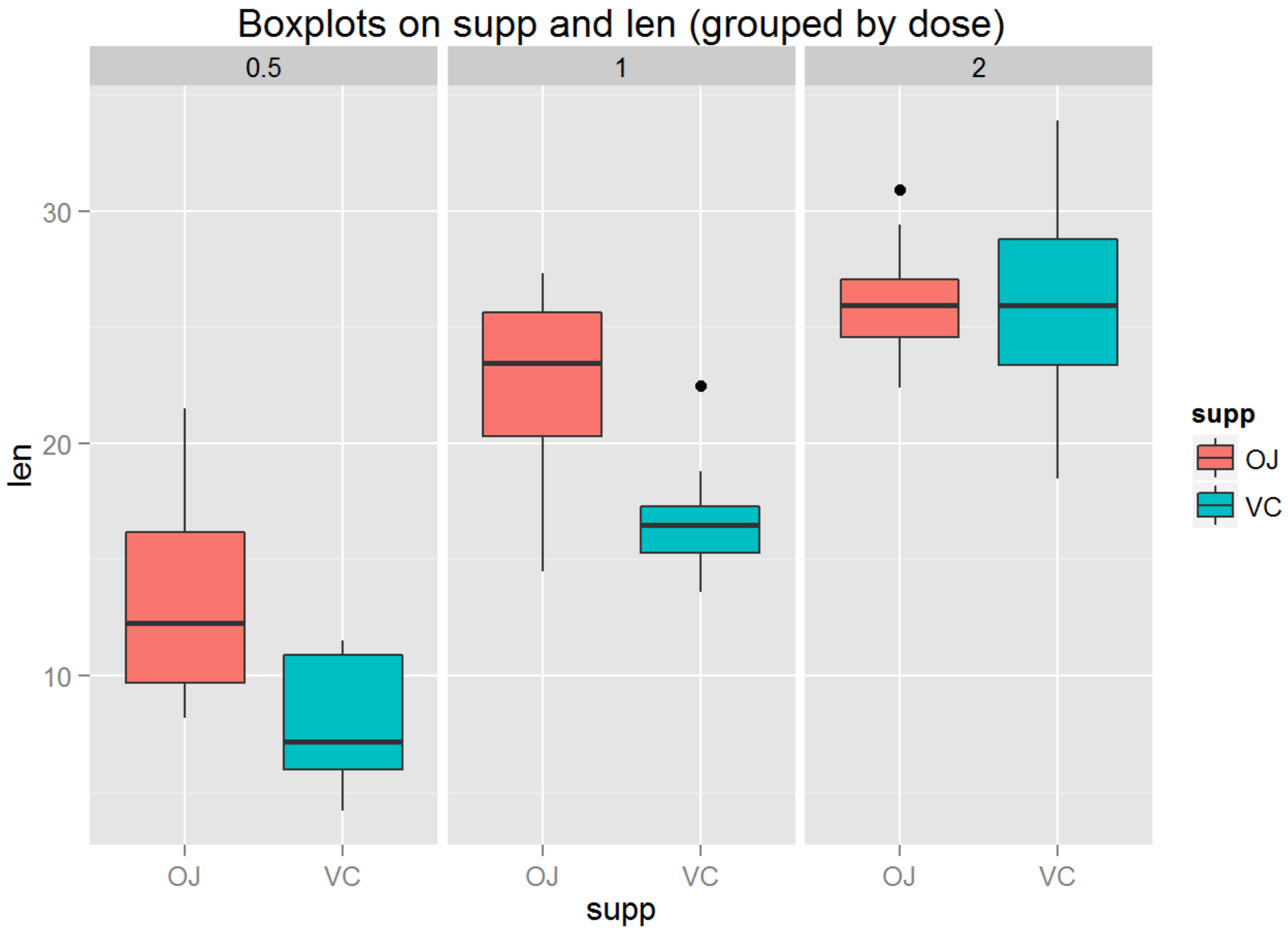
The following scatterplot on the variables `dose` and `len` shows both types of delivery methods using different colors:

```
g <- ggplot(data=ToothGrowth, aes(x=dose, y=len, color=supp)) + labs(title="Scatterplot on dose and l
en")
g <- g + geom_point(shape=15)
print(g)
```

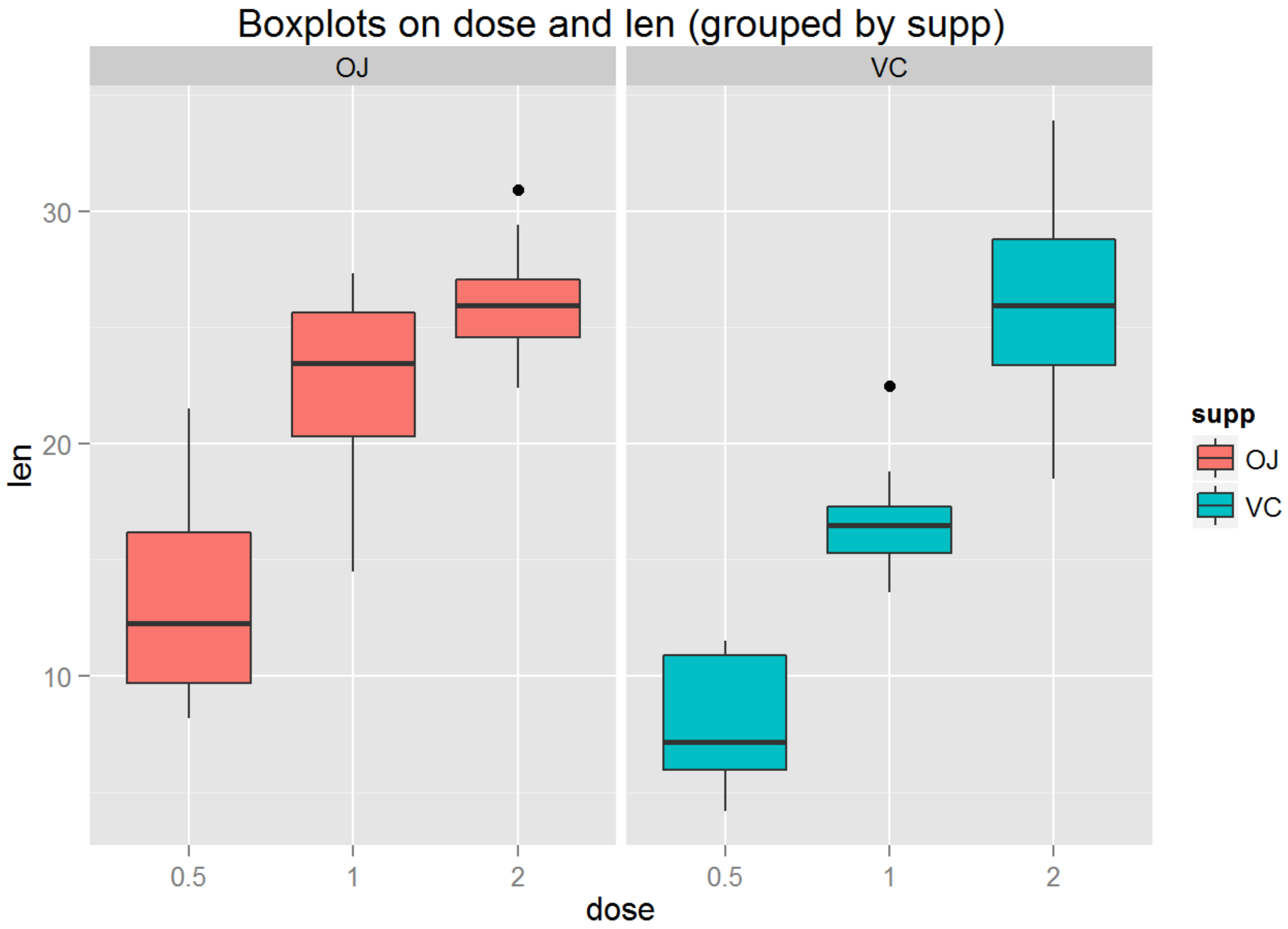


As there are now two categorical variables (`supp` and `dose`) there are two ways of visualizing side-by-side boxplots: with the numerical variable `len` on the y-axis, one categorical variable is used as the variable on the x-axis and the other variable is used as the grouping (or “panel”) variable:

```
g <- ggplot(data=ToothGrowth, aes(x=supp, y=len, fill=supp))
g <- g + labs(title="Boxplots on supp and len (grouped by dose)")
g <- g + geom_boxplot() + facet_wrap(~ dose)
print(g)
```



```
g <- ggplot(data=ToothGrowth, aes(x=dose, y=len, fill=supp))
g <- g + labs(title="Boxplots on dose and len (grouped by supp)")
g <- g + geom_boxplot() + facet_wrap(~ supp)
print(g)
```



Confidence Intervals and Hypothesis Testing

95% *t* Confidence Intervals

In order to compare the two types of delivery methods, represented in the variable `supp` with values OJ and VC, we calculate an **independent group *t* confidence interval** for the tooth growth difference (as we assume there's no link between the pigs in the two groups). As there are three different dose levels (variable `dose`) and we want the comparision to be made for the same dose level, we will calculate one confidence interval for each of the three dose levels 0.5, 1.0, and 2.0.

If we assume that the **variances from the two samples are equal**, we can calculate the 95% *t* confidence intervals using the pooled variance S_p^2 to estimate the variance and $n_{OJ} + n_{VC} - 2 (= 18)$ as the number for the degrees of freedom:

```
v_len.05.OJ <- ToothGrowth[ToothGrowth$dose == 0.5 & ToothGrowth$supp == 'OJ', 1]
v_len.05.VC <- ToothGrowth[ToothGrowth$dose == 0.5 & ToothGrowth$supp == 'VC', 1]
v_len.10.OJ <- ToothGrowth[ToothGrowth$dose == 1.0 & ToothGrowth$supp == 'OJ', 1]
v_len.10.VC <- ToothGrowth[ToothGrowth$dose == 1.0 & ToothGrowth$supp == 'VC', 1]
v_len.20.OJ <- ToothGrowth[ToothGrowth$dose == 2.0 & ToothGrowth$supp == 'OJ', 1]
v_len.20.VC <- ToothGrowth[ToothGrowth$dose == 2.0 & ToothGrowth$supp == 'VC', 1]

n_OJ <- length(v_len.05.OJ); n_VC <- length(v_len.05.VC)  ## sample size is the same for both types
df <- n_OJ + n_VC - 2  ## degrees of freedom is n+m-2
mean_05_OJ <- mean(v_len.05.OJ); mean_05_VC <- mean(v_len.05.VC)
sd_05_OJ <- sd(v_len.05.OJ); sd_05_VC <- sd(v_len.05.VC)
mean_10_OJ <- mean(v_len.10.OJ); mean_10_VC <- mean(v_len.10.VC)
sd_10_OJ <- sd(v_len.10.OJ); sd_10_VC <- sd(v_len.10.VC)
mean_20_OJ <- mean(v_len.20.OJ); mean_20_VC <- mean(v_len.20.VC)
sd_20_OJ <- sd(v_len.20.OJ); sd_20_VC <- sd(v_len.20.VC)

# Pooled sd estimator, sp, is the square root of the pooled variance estimator.
sp_05 <- sqrt( ( (n_OJ-1)*sd_05_OJ^2 + (n_VC-1)*sd_05_VC^2 ) / df )
sp_10 <- sqrt( ( (n_OJ-1)*sd_10_OJ^2 + (n_VC-1)*sd_10_VC^2 ) / df )
sp_20 <- sqrt( ( (n_OJ-1)*sd_20_OJ^2 + (n_VC-1)*sd_20_VC^2 ) / df )

# Calculate the 95% t confidence interval using pooled varaince estimate:
mean_05_OJ - mean_05_VC + c(-1, 1) * qt(.975, df) * sp_05 * sqrt(1/n_OJ + 1/n_VC)
```

```
## [1] 1.77 8.73
```

```
mean_10_OJ - mean_10_VC + c(-1, 1) * qt(.975, df) * sp_10 * sqrt(1/n_OJ + 1/n_VC)
```

```
## [1] 2.841 9.019
```

```
mean_20_OJ - mean_20_VC + c(-1, 1) * qt(.975, df) * sp_20 * sqrt(1/n_OJ + 1/n_VC)
```

```
## [1] -3.723  3.563
```

Note: We can get the same confidence intervals using the *t.test()* function with parameter `var.equal = TRUE`. If we assume that the **variances from the two samples are not equal**, we can calculate the 95% *t* confidence intervals using the *t.test* function with parameter `var.equal = FALSE` (the Welch or Satterthwaite approximation to the degrees of freedom is used):

```
# Using the t.test() function we can set var.equal = FALSE
t.test(v_len.05.OJ, v_len.05.VC, paired = FALSE, var.equal = FALSE)$conf
```

```
## [1] 1.719 8.781
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(v_len.10.OJ, v_len.10.VC, paired = FALSE, var.equal = FALSE)$conf
```

```
## [1] 2.802 9.058
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(v_len.20.OJ, v_len.20.VC, paired = FALSE, var.equal = FALSE)$conf
```

```
## [1] -3.798  3.638
## attr(,"conf.level")
## [1] 0.95
```

From the three 95% confidence intervals only the one for **dose level 2.0 contains 0**. This means that if one were to repeatedly get samples of size 10, about **95% of the intervals obtained would contain 0**. Therefore we cannot say that there's a statistical significant difference between the two treatment methods for a dose of 2.0 mg.

Hypothesis Testing with Type I Error Rate α = 5%

Now, let's see if we come to the same conclusion using hypothesis testing. We define our null hypothesis to represent the status quo, i.e. that there's no difference in tooth growth between the two treatment methods. We perform a two-sided *t*-test:

$H_0 : \mu = \mu_0$ vs. $H_A : \mu \neq$ μ_0 (where μ_0 is the population mean of the difference under the assumption that H_0 is true, so $\mu_0 = \mu_{OJ} - \mu_{VC} = 0$)

If we assume that the **variances from the two samples are equal**, we can calculate the 95% *t* statistics using the pooled variance S_p^2 to estimate the variance and $n_{OJ} + n_{VC} - 2 (= 18)$ as the number for the degrees of freedom:

We will reject H_0 when $| \frac{\bar{X}-\mu_0}{S_p * \sqrt{\frac{1}{n_{OJ}}+\frac{1}{n_{VC}}}} | \geq t_{1-\alpha/2, n_{OJ}+n_{VC}-2}$, i.e. when $| \frac{(\bar{X}_{OJ}-\bar{X}_{VC})-0}{S_p * \sqrt{\frac{1}{10}+\frac{1}{10}}} | \geq t_{0.975,18} (= 2.1009)$.

For the three dose levels we get the following *t*-test statistics (after putting the already calculated values in the left side of the formula above):

- dose level 0.5: $| \frac{(13.23-7.98)-0}{3.7036 \times 0.4472} | = |3.1697|$
- dose level 1.0: $| \frac{(22.7-16.77)-0}{3.288 \times 0.4472} | = |4.0328|$
- dose level 2.0: $| \frac{(26.06-26.14)-0}{3.8773 \times 0.4472} | = |-0.0461|$

Note: We can get the same *t*-test statistics using the *t.test* function with parameter `var.equal = TRUE`.
If we assume that the **variances from the two samples are not equal**, we can calculate the 95% *t*-test statistics using the *t.test* function with parameter `var.equal = FALSE` (the Welch or Satterthwaite approximation to the degrees of freedom is used):

```
# Using the t.test() function we can set var.equal = FALSE
t.test(v_len.05.OJ, v_len.05.VC, alternative="two.sided", paired = FALSE, var.equal = FALSE)$statisti
c
```

```
##    t
## 3.17
```

```
t.test(v_len.10.OJ, v_len.10.VC, alternative="two.sided", paired = FALSE, var.equal = FALSE)$statisti
c
```

```
##     t
## 4.033
```

```
t.test(v_len.20.OJ, v_len.20.VC, alternative="two.sided", paired = FALSE, var.equal = FALSE)$statisti
c
```

```
##        t
## -0.04614
```

From the three *t*-test statistics for $\alpha = 0.5$ only the one for **dose level 2.0 is not big enough to reject the null hypothesis**. Therefore we cannot say that there's a statistical significant difference between the two treatment methods for a dose of 2.0 mg.

Results Interpretation

We can state what the sidy-by-side boxplot on `supp` and `len` grouped by `dose` already suggested, i.e. that there is a statistical significant difference on the tooth growth between the two treatment methods for a dose of 0.5 and 1.0 mg, but not for a dose of 2.0 mg.
We have also seen that the connection between hypothesis testing and confidence intervals holds, i.e. that if a $(1 - \alpha) \times 100\%$ interval contains μ_0 then we also fail to reject H_0.