Inferential Data Analysis - Toothrowth Dataset

Willem Abrie

2022-11-07

Loading Initial Packages

We will mostly rely on the "tidyverse" family of packages for data manipulation.

```
# Load Packages and get the Data

# Install pacakages
check_and_install_package <- function(package_name){
    if(!package_name %in% installed.packages()){
        install.packages(package_name)
    }
}

check_and_load_package <- function(package_name){
    if(!package_name %in% (.packages())){
        library(package_name, character.only = TRUE)
    }
}

check_and_install_package("tidyverse")
check_and_load_package("tidyverse")

check_and_install_package("GGally")
check_and_load_package("GGally")

check_and_install_package("rstatix")
check_and_load_package("rstatix")

check_and_install_package("ggpubr")
check_and_load_package("ggpubr")
(.packages())
```

Load the data - ToothGrowth from the R Datasets package.

Standrard R description of the data:

The Effect of Vitamin C on Tooth Growth in Guinea Pigs

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

```
tooth_raw <- ToothGrowth

summary(tooth_raw)
```

```
##       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  
```

```
str(tooth_raw)
```

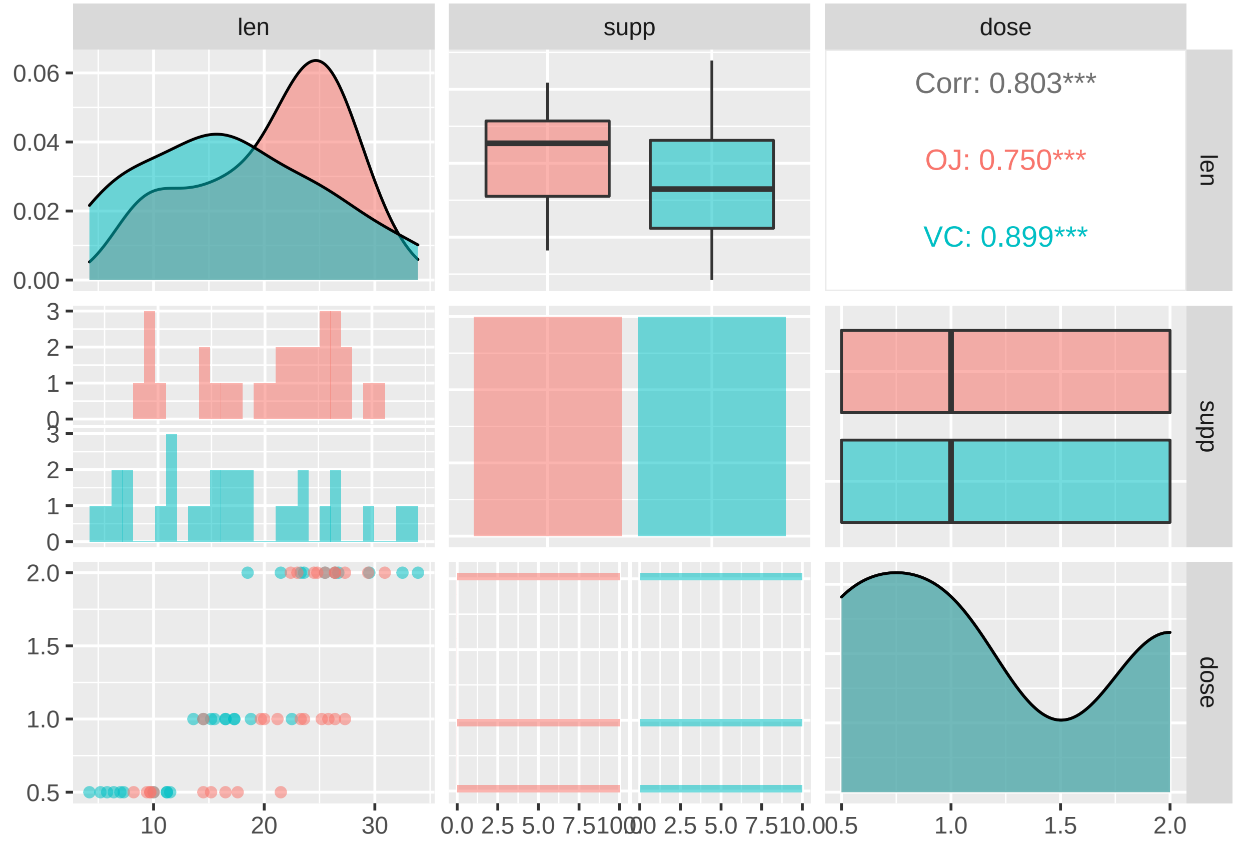
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

Data Exploration

We note we only have 60 observations balanced accross both groups. That means only 10 observations per category (i.e. supplement and dose combination). No missing values. For a quick exploration we run the ggpairs() function from the GGally package.

```
ggpairs(tooth_raw
        ,ggplot2::aes(colour=supp, alpha = 0.8)#,
        #upper = list(continuous = "density", combo = "box_no_facet")
        )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



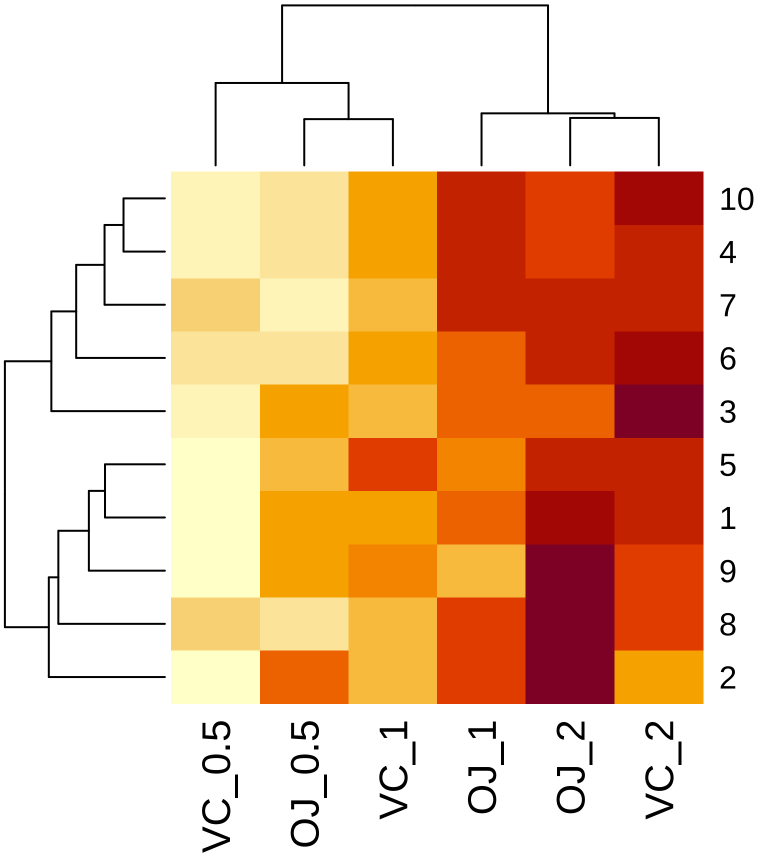
We note the following from the plots above: 1. The OJ supplements seemed to have resulted in a distribution with higher quantities of longer cells, 2. The median length is also a bit larger than that produced by the VC, but since there is a fair amount of overlap of the interquartile range, we'll have to rely on our confidence intervals and hypothesis testing to determine if the difference is meaningful. 3. It seems that with both supplements there was a bit of diminishing returns in the sense that doubling the dose from 1.0 to 2.0mg didn't have as profound an effect as going from 0.5 to 1. This trend was more intense for the OJ, hence it has a slightly lower linear correlation of length to dose.

Data Summary

We draw a heatmap of the data to get a bit of a feel for how they compare, and also look at the means broken down by supplement and dose

```
data_heat <- tooth_raw %>%
        mutate(item = rep(1:10, 6)) %>%
        pivot_wider(names_from = c(supp, dose), values_from = len) %>%
        select(-item)%>%
        sapply(as.numeric)


heatmap(as.matrix(data_heat))
```



```r
data_means <- tooth_raw %>%
        pivot_wider(names_from = c(supp, dose), values_from = len, values_fn = mean)
data_var <- tooth_raw %>%
        pivot_wider(names_from = c(supp, dose), values_from = len, values_fn = var)

data_plus <- rbind(data_means, data_var) %>%
        mutate(Stat = c("Means", "Variance"))

data_plus
```

```
## # A tibble: 2 x 7
##   VC_0.5  VC_1  VC_2 OJ_0.5  OJ_1  OJ_2 Stat
##    <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <chr>
## 1   7.98 16.8   26.1   13.2  22.7 26.1  Means
## 2   7.54  6.33  23.0   19.9  15.3  7.05 Variance
```

We see that OJ seems significantly better than VC at the 0.5 and 1.0mg doses but at a much larger variance. At 2.0mg the two types of supplements perform very similarly on average, but the VC this time has a much larger variance.

Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose

Null hypothesis: There is no difference in in mean cell length as you vary the dose of OJ. I.e. mu_1 - mu_0.5 = 0 Alternate Hypothesis: The cells are longer for a higher dose. I.e. mu_1 - mu_0.5 > 0

```r
data <- tooth_raw %>%
        mutate(item = rep(1:10, 6)) %>%
        filter(supp == "OJ") %>%
        pivot_wider(names_from = dose, values_from = len) %>%
        select(-item)

t.test(data$`1`-data$`0.5`, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  data$‘1‘ - data$‘0.5‘
## t = 4.1635, df = 9, p-value = 0.001218
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  5.300497      Inf
## sample estimates:
## mean of x
##      9.47
```

This shows that we will reject the null hypothesis. Similar for the rest of the dose comparisons for both OJ and VC.

Let’s also check for OJ comparing 1 and 2mg doses:

```r
t.test(data$`2`-data$`1`, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  data$‘2‘ - data$‘1‘
## t = 1.9435, df = 9, p-value = 0.04192
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.1908169       Inf
## sample estimates:
## mean of x
##      3.36
```

Notice the p-value is almost 5%, but we’ll still reject the null hypothesis.

Now we compare supplements:

Null hypothesis: There is no difference in in mean cell length outcome depending on whether OJ or VC is given for a specific dose. I.e. mu_OJ_1.0 - mu_VC_1.0 = 0 Alternate Hypothesis: The cells are longer for a higher dose. I.e. mu_OJ_1.0 - mu_VC_1.0 > 0

```r
data2 <- tooth_raw %>%
        mutate(item = rep(1:10, 6)) %>%
        filter(dose == 1.0) %>%
        pivot_wider(names_from = supp, values_from = len) %>%
        select(-item)

t.test(data2$OJ-data2$VC, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  data2$OJ - data2$VC
## t = 3.3721, df = 9, p-value = 0.004115
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  2.706401      Inf
## sample estimates:
## mean of x
##      5.93
```

We see that there is a significant difference between supplements at 1.0 mg dose.

Let's look at 2 mg:

Null hypothesis: There is no difference in in mean cell length outcome depending on whether OJ or VC is given for a specific dose. I.e. mu_OJ_2.0 - mu_VC_2.0 = 0 Alternate Hypothesis: The cells are longer for a higher dose. I.e. mu_OJ_2.0 - mu_VC_2.0 != 0

```
data3 <- tooth_raw %>%
        mutate(item = rep(1:10, 6)) %>%
        filter(dose == 2.0) %>%
        pivot_wider(names_from = supp, values_from = len) %>%
        select(-item)

t.test(data3$OJ-data3$VC, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  data3$OJ - data3$VC
## t = -0.042592, df = 9, p-value = 0.967
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.328976  4.168976
## sample estimates:
## mean of x
##     -0.08
```

The p=value is very large in this instance, so we cannot reject the null hypothesis, ie. that the type of supplement makes no difference at this dose.

Conclusions and Discussion of Assumptions

We conclude that dose of supplement does have a positive outcome on tooth cell length. We also conclude that type of supplement does make difference (OJ is superior) at 0.5 and 1mg doses, but not at 2mg.

Assumptions were as follows: I leaned on this (datanovia page)[https://www.datanovia.com/en/lessons/t-test-assumptions/independent-t-test-assumptions/].

1. Independence of the observations. Each subject should belong to only one group. There is no relationship between the observations in each group. - O.K.

2. No significant outliers in the two groups - O.K. no extreme outliers

```
outliers <- tooth_raw %>%
        group_by(supp, dose) %>%
        identify_outliers(len)

outliers
```

```
## # A tibble: 2 x 5
##   supp   dose   len is.outlier is.extreme
##   <fct> <dbl> <dbl> <lgl>      <lgl>
## 1 OJ        2  30.9 TRUE       FALSE
## 2 VC        1  22.5 TRUE       FALSE
```

4. Normality. the data for each group should be approximately normally distributed. - O.K. The normality assumption can be checked by computing the Shapiro-Wilk test for each group. If the data is normally distributed, the p-value should be greater than 0.05.

```
normality <- tooth_raw %>%
        group_by(supp, dose) %>%
        shapiro_test(len)

normality
```

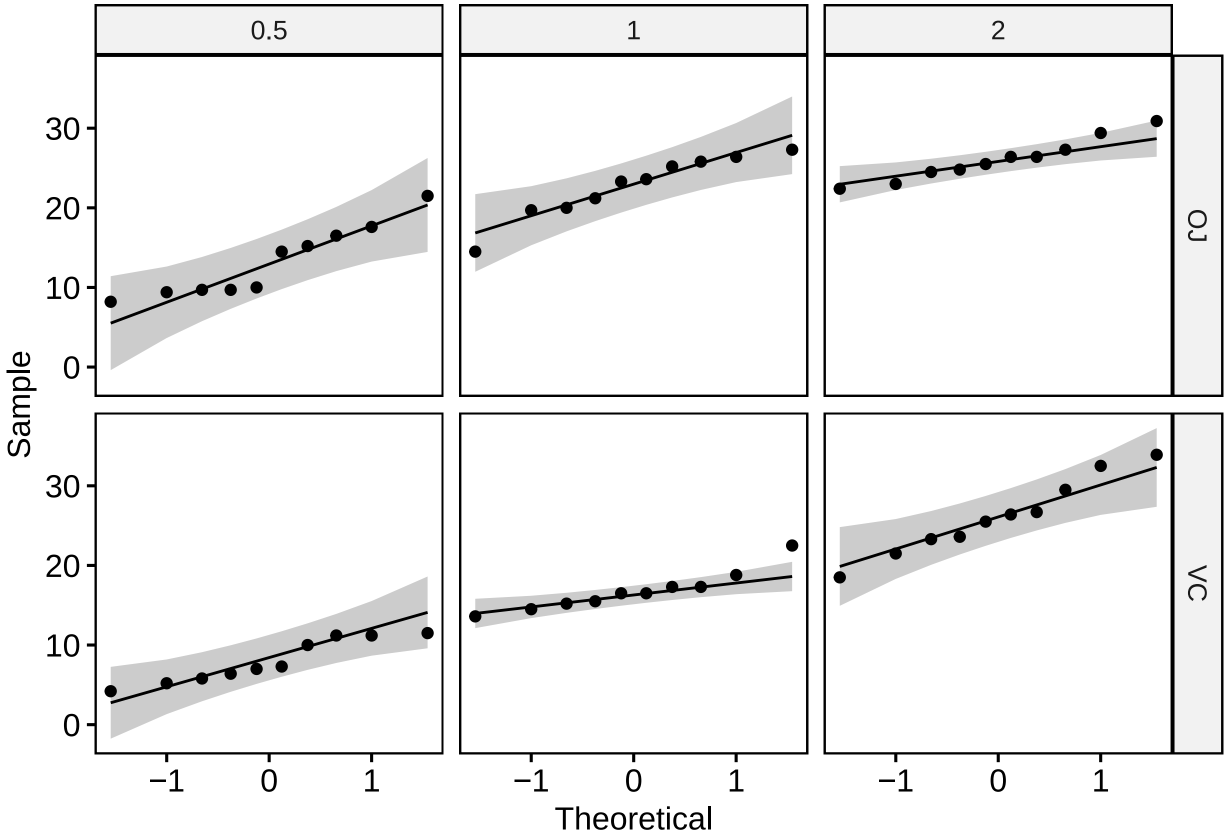
```
## # A tibble: 6 x 5
##   supp   dose variable statistic     p
##   <fct> <dbl> <chr>        <dbl> <dbl>
## 1 OJ      0.5 len          0.893 0.182
## 2 OJ      1   len          0.927 0.415
## 3 OJ      2   len          0.963 0.815
## 4 VC      0.5 len          0.890 0.170
## 5 VC      1   len          0.908 0.270
## 6 VC      2   len          0.973 0.919
```

You can also create QQ plots for each group. QQ plot draws the correlation between a given data and the normal distribution.

```
#qq_data <- tooth_raw %>%
     #   pivot_longer(names_to = len, values_to = c(dose, supp))

#qq_data

ggqqplot(tooth_raw, x = "len", facet.by = c("supp", "dose"))
```



All the points fall approximately along the (45-degree) reference line, for each group. So we can assume normality of the data.

5. Homogeneity of variances. the variance of the outcome variable should be equal in each group. Recall that, the Welch t-test does not make this assumptions. - Not O.K. This can be done using the Levene's test. If the variances of groups are equal, the p-value should be greater than 0.05

```
tooth_raw %>% levene_test(len ~ supp)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     1    58      1.21 0.275
```

The p-value of the Levene's test is not significant, suggesting that there is no significant difference between the variances of the two groups. Therefore we can use the standard t-test.