Course Project of Hypothesis Test

Xudong Zhang

September 10, 2014

1. Basic exploratory data analysis

```
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 ...
```

The dataset contains 60 observations and 3 variables. The variables are Tooth length(len), Supplement type(supp) and Dose in milligrams(dose).

Intuitively, tooth length depends on both supplement type and dose. The variable supp is a factor variable of {1, 2}, and the variable dose can be transformed into a factor variable of {0.5, 1, 2}

2. Basic summary

```
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  
```

```
names <- c("OJ", "VC", "d0.5", "d1", "d2", "OJ.d0.5", "VC.d0.5",
           "OJ.d1", "VC.d1", "OJ.d2", "VC.d2")
all_vec <- list(len_OJ, len_VC, len_all_d0, len_all_d1, len_all_d2,
                len_OJ_d0, len_VC_d0, len_OJ_d1, len_VC_d1, len_OJ_d2, len_VC_d2)
means <- round(sapply(all_vec, mean), 2)
std_devs <- round(sapply(all_vec, sd), 2)
rbind(names, means, std_devs)
```

```
##          [,1]    [,2]    [,3]    [,4]    [,5]   [,6]      [,7]     
## names    "OJ"    "VC"    "d0.5"  "d1"    "d2"   "OJ.d0.5" "VC.d0.5"
## means    "20.66" "16.96" "10.61" "19.73" "26.1" "13.23"   "7.98"   
## std_devs "6.61"  "8.27"  "4.5"   "4.42"  "3.77" "4.46"    "2.75"   
##          [,8]    [,9]    [,10]   [,11]  
## names    "OJ.d1" "VC.d1" "OJ.d2" "VC.d2"
## means    "22.7"  "16.77" "26.06" "26.14"
## std_devs "3.91"  "2.52"  "2.66"  "4.8"  
```

We first give a summary of distribution on each variable independently. We then plot the statistics of tooth length on different groups of supplement and dose in Figure2. The exact value is shown in the above table.

We can see that generally, the means of tooth length on OJ is larger that on VC, in terms of supplement. And as far as dose is concerned, increase in dose can lead to higher means of tooth length. We will verify the assumption in the hypothese testing section.

```
shapiro.test(ToothGrowth$len)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ToothGrowth$len
## W = 0.9674, p-value = 0.1091
```

Moreover, we perform the Shapiro-Wilk normality test on tooth length variable, and found that it's approximately normal distribute(greater than 5% significance level). Q-Q plot of tooth length on Figure 3 also proved the discovery.

3. Hypothesis testing

Here we all use one-sided t-test with two independent group on equal variance, and the significance level to set to be 0.05. We make 6 pairs of test: (1) supp OJ vs VC, (2) dose 1 vs 0.5, (3) dose 2 vs 1, (4) supp OJ vs VC on dose 0.5, (5) supp OJ vs VC on dose 1 and (6) supp OJ vs VC on dose 2.

```
t1 <- t.test(len_OJ, len_VC, alternative = "greater")
t2 <- t.test(len_all_d1, len_all_d0, alternative = "greater")
t3 <- t.test(len_all_d2, len_all_d1, alternative = "greater")
t4 <- t.test(len_OJ_d0, len_VC_d0, alternative = "greater")
t5 <- t.test(len_OJ_d1, len_VC_d1, alternative = "greater")
t6 <- t.test(len_OJ_d2, len_VC_d2, alternative = "greater")
d <- sapply(list(t1, t2, t3, t4, t5, t6), function(iter) {
  round (c(iter$stat[[1]], iter$p.value, iter$conf[1]), 3)})
cbind(c("t stat", "p-value", "lower conf"), d)
```

```
##      [,1]         [,2]    [,3]    [,4]    [,5]    [,6]    [,7]
## [1,] "t stat"     "1.915" "6.477" "4.9"   "3.17"  "4.033" "-0.046"
## [2,] "p-value"    "0.03"  "0"     "0"     "0.003" "0.001" "0.518"
## [3,] "lower conf" "0.468" "6.753" "4.174" "2.346" "3.356" "-3.133"
```

We list the t stat, p-value and corresponding lower bounds of 95% confidence interval(one sided) for each test. We found that for first 5 tests, the null hypothese is reject and the first set in the pair has larger mean. And the last t test(supp OJ vs VC on dose 2) cannot reject the null hypothesis.

From the above t tests, we come to the conclusion that (1) larger dose will leader to larger tooth length and (2) OJ supplement has better influence on tooth length than VC supplement.

Appendix

Figure 1. Plot data on each 2 out of 3 variables

```
plot(ToothGrowth)
```

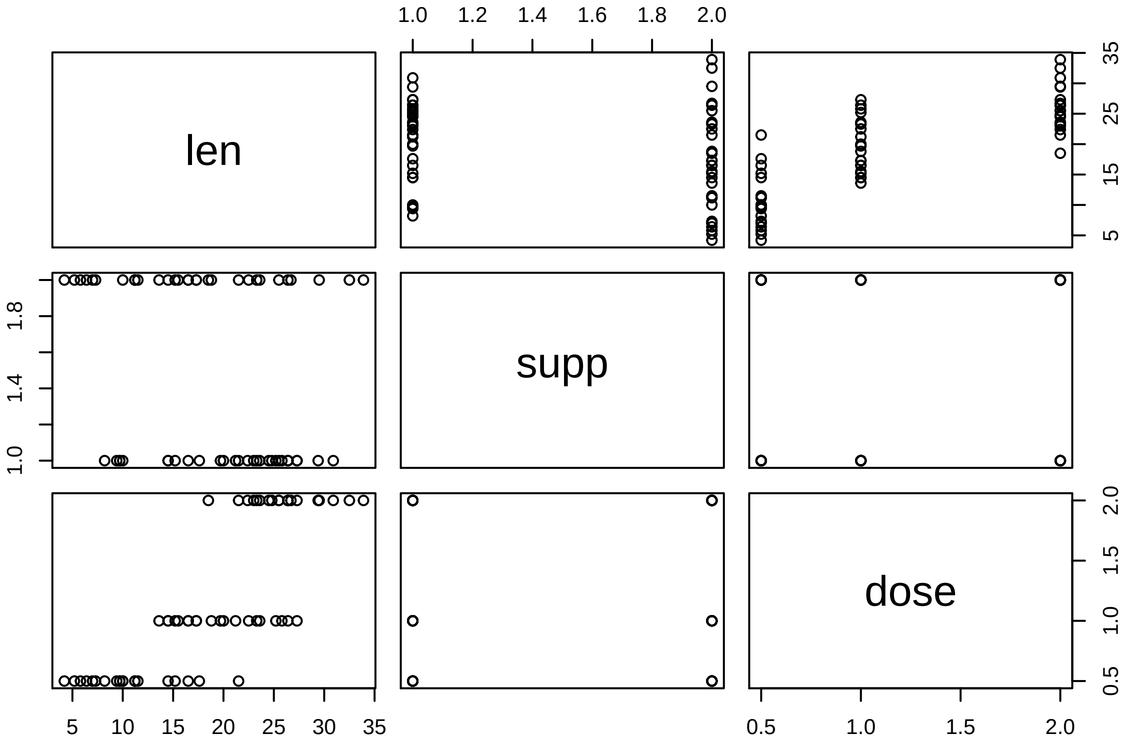
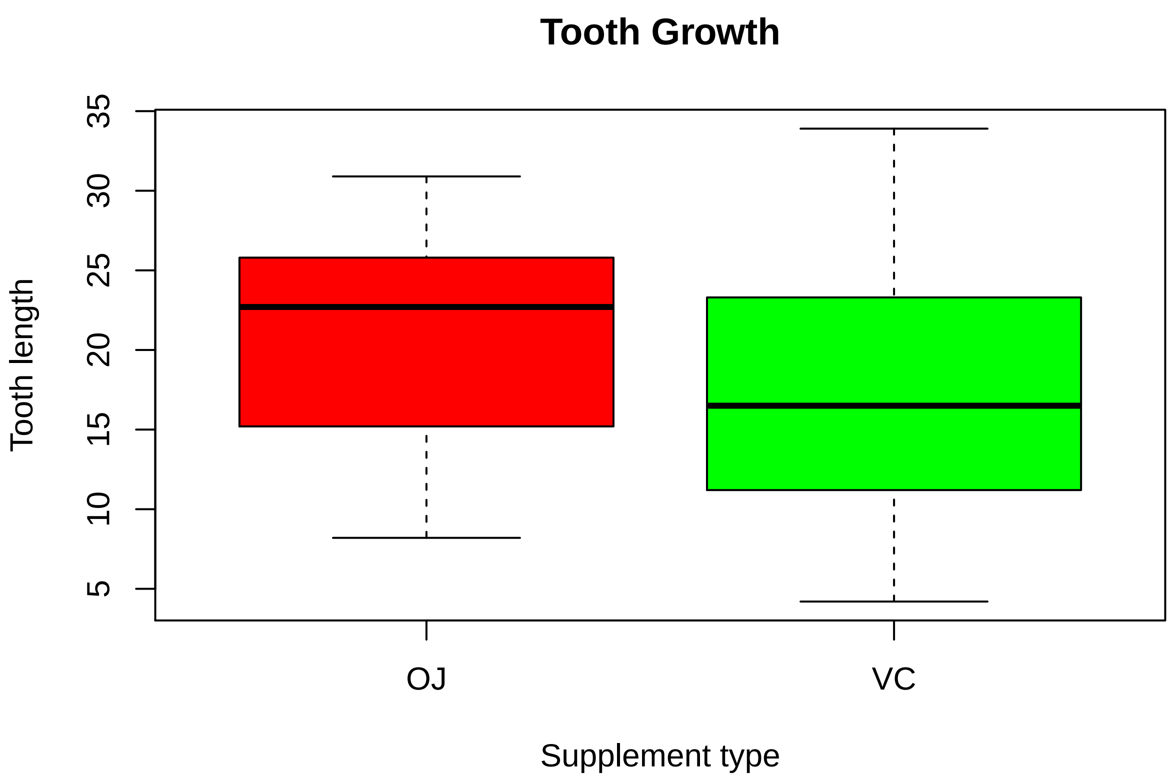
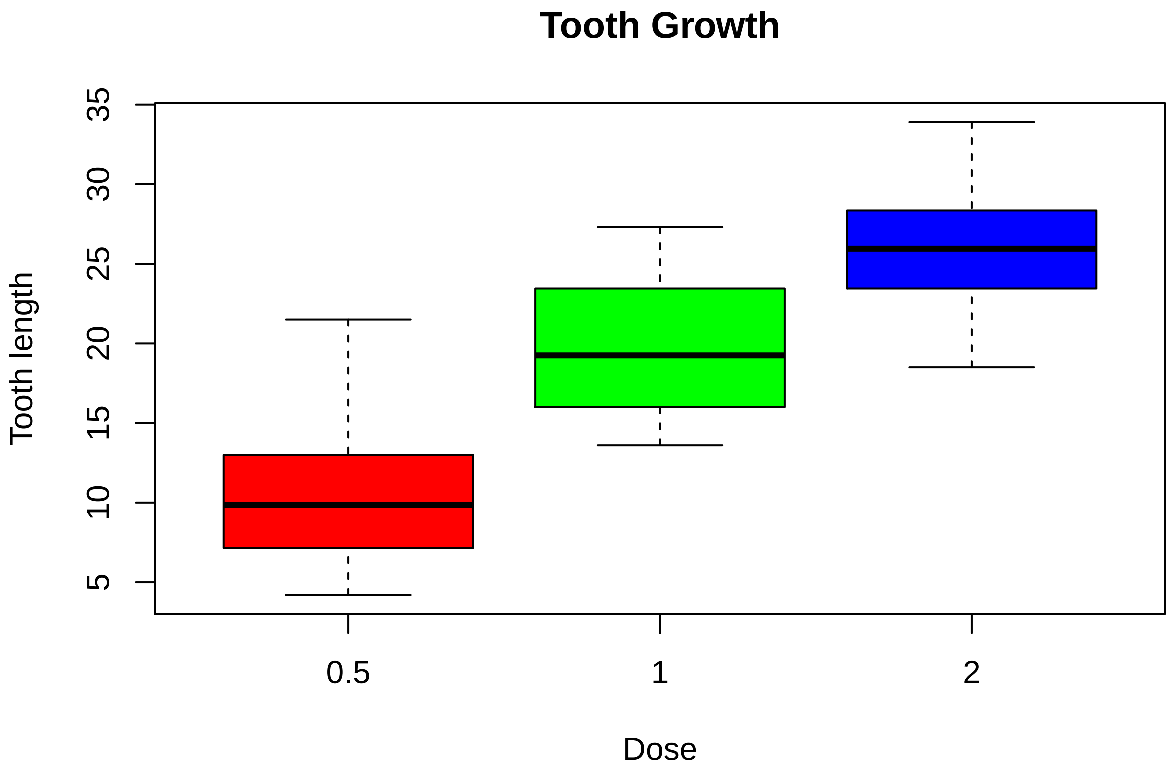


Figure 2. Box plot of len depending on different combination of value of supp and dose

```r
boxplot(len ~ supp, data = ToothGrowth, xlab = "Supplement type", ylab = "Tooth length",
        main = "Tooth Growth", col = c("red", "green"))
```




```r
boxplot(len ~ dose, data = ToothGrowth, xlab = "Dose", ylab = "Tooth length",
        main = "Tooth Growth", col = c("red", "green", "blue"))
```




```
boxplot(len ~ supp:dose, data = ToothGrowth, xlab = "Supplement and dose", ylab = "Tooth length",
        main = "Tooth Growth", col = c("red", "green"))
```

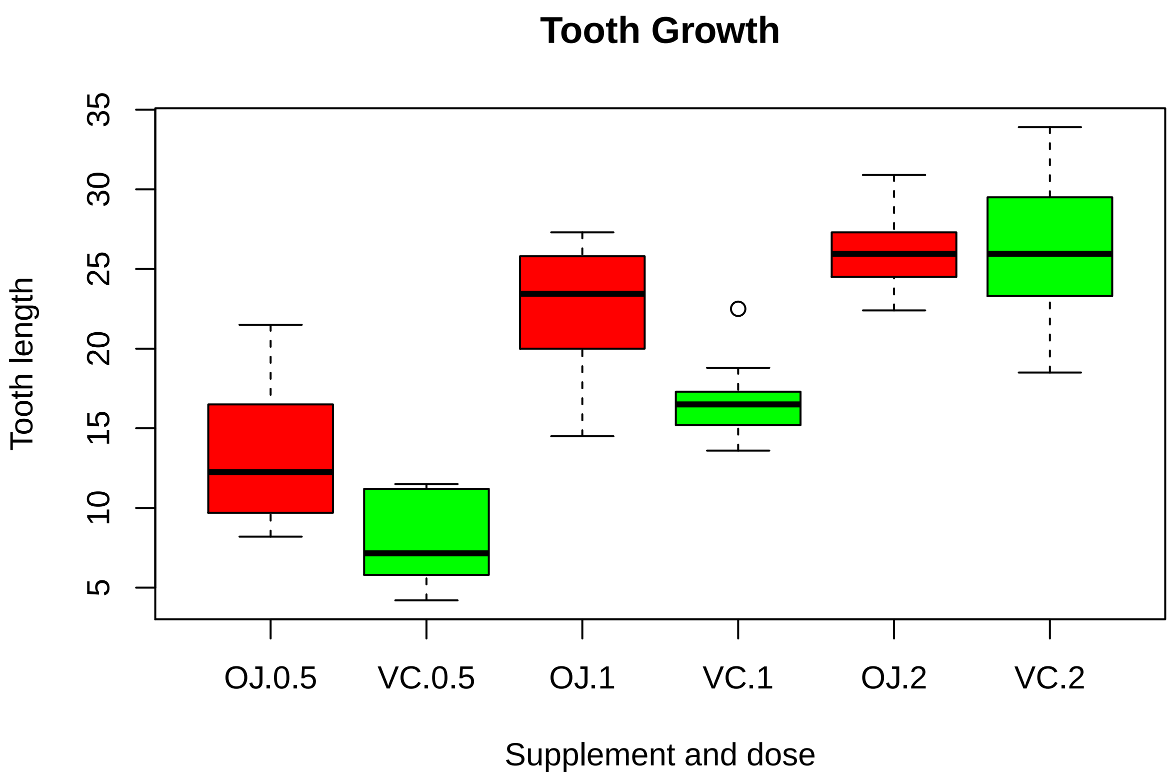
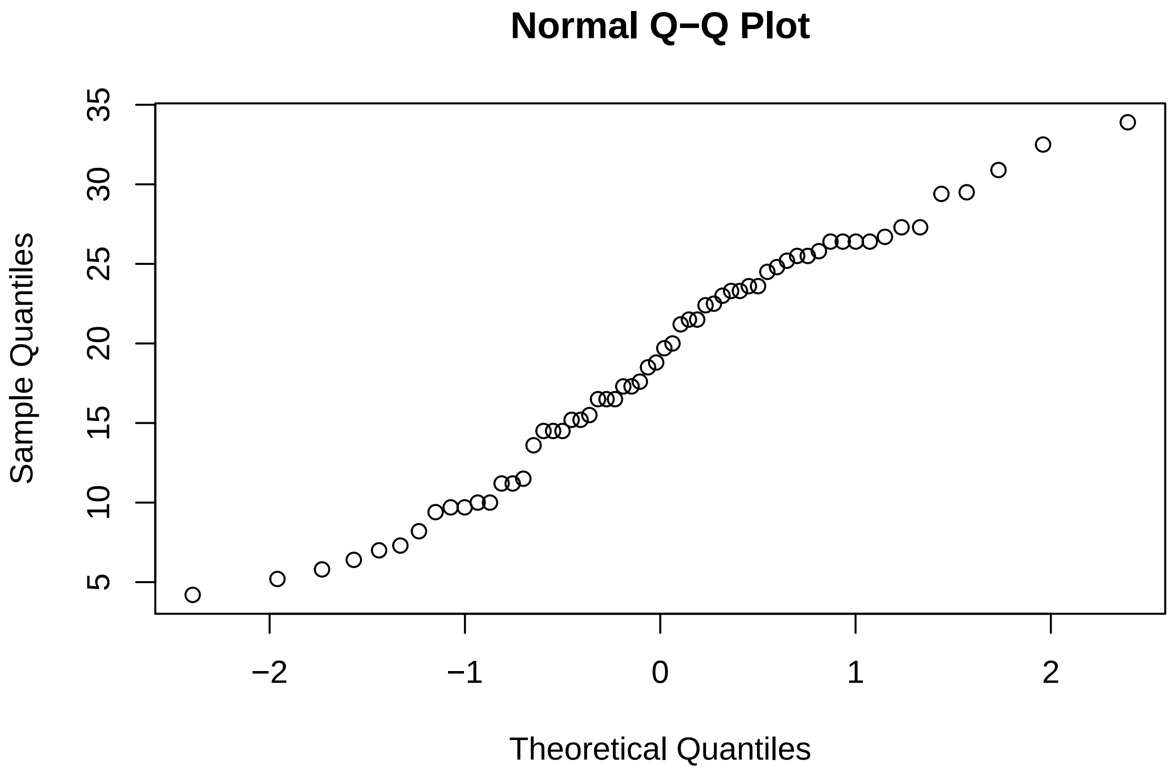


Figure 3 Q-Q Plot of Tooth Length

```
qqnorm(ToothGrowth$len)
```



Linear regression on the dataset

We also use fit the dataset into a regression model the verify the influence of supplement, dose on tooth length.

```
fit <- lm(len ~ factor(supp) * factor(dose) - 1, data = ToothGrowth)
summary(fit)
```

```
##
## Call:
## lm(formula = len ~ factor(supp) * factor(dose) - 1, data = ToothGrowth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.20  -2.72  -0.27   2.65   8.27
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## factor(supp)OJ                    13.23       1.15   11.52  3.6e-16 ***
## factor(supp)VC                     7.98       1.15    6.95  5.0e-09 ***
## factor(dose)1                      9.47       1.62    5.83  3.2e-07 ***
## factor(dose)2                     12.83       1.62    7.90  1.4e-10 ***
## factor(supp)VC:factor(dose)1      -0.68       2.30   -0.30    0.768
## factor(supp)VC:factor(dose)2       5.33       2.30    2.32    0.024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.63 on 54 degrees of freedom
## Multiple R-squared:  0.971,  Adjusted R-squared:  0.968
## F-statistic:  303 on 6 and 54 DF,  p-value: <2e-16
```

The analysis indicate that, given 5% significant level, supp OJ has larger influence than VC, and the larger the dose, the larger the tooth length.