Project Statistical Inference

JOHN HOPKINS UNIVERSTY THROUGH COURSERA. January 5th - February 2nd, 2015 Oscar Portillo

Project Description:

The project consists of two parts: a) simulation to explore inference and b) basic inferential data analysis.

B) Basic Inferential Data Analysis

In the second portion of this project, ToothGrowth data (dowloaded from the R datasets package) is used to perform basic inferential data analysis. The data describes the effect of vitamin C on tooth growth in Guinea Pigs.

Tasks:

1. Load the ToothGrowth data and perform some basic exploratory data analyses
2. Provide a basic summary of the data.
3. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.
4. State your conclusions and the assumptions needed for your conclusions.

Task 1: Download the ToothGrowth data and get the column names. It is observed that there are 3 columns: "len"-> tooth length, "supp" -> supplement type, and "dose" -> dose in milligrams

```
library(datasets)
data(ToothGrowth)
names(ToothGrowth)

## [1] "len"  "supp" "dose"
```

Take a look at the first 3 rows and last 3 rows. There are two delivery methods to supplemt Vitamin C: VC-> Ascorbic Acid and OJ -> Orange Juice

```
head(ToothGrowth,3); tail(ToothGrowth,3)

##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5

##     len supp dose
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

Review data variables. There are 60 observations of 3 variables. The 60 observations come from 10 pigs, 3 dose levels of Vitamin C (0.5, 1, and 2 mg) and 2 supplement types, so 10 x 3 x 2 = 60.

```
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 ...
```

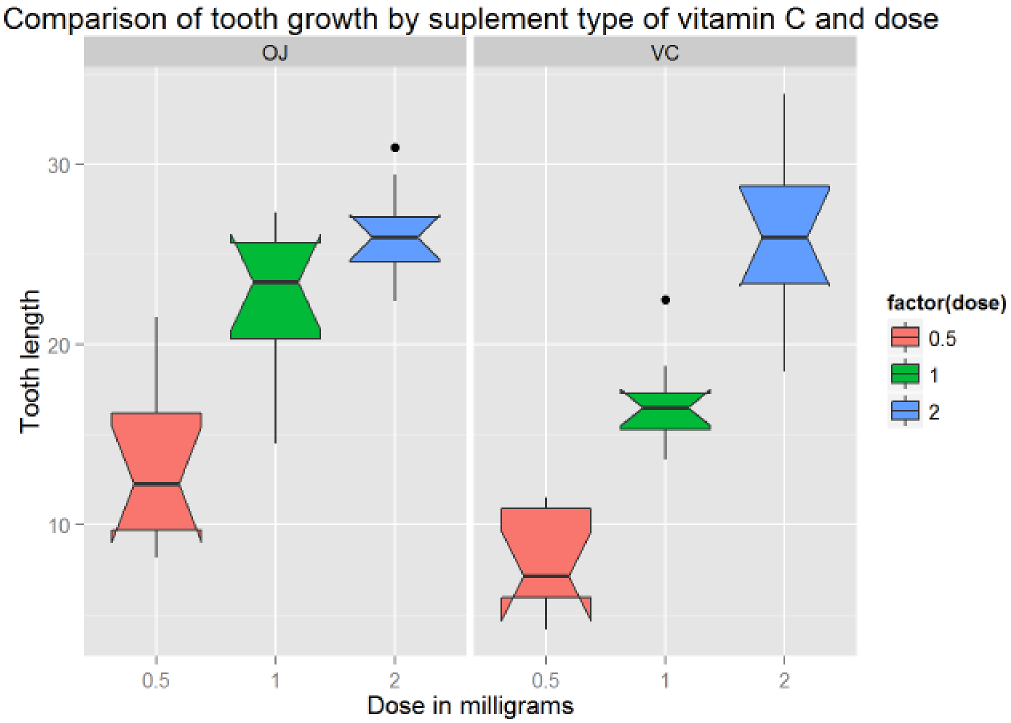
Task 2. Simple summary statistics of the toothgrowth data is shown below. It can be observed that the mean length of odontoblasts (teeth) is 18.81 units.

```
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
```

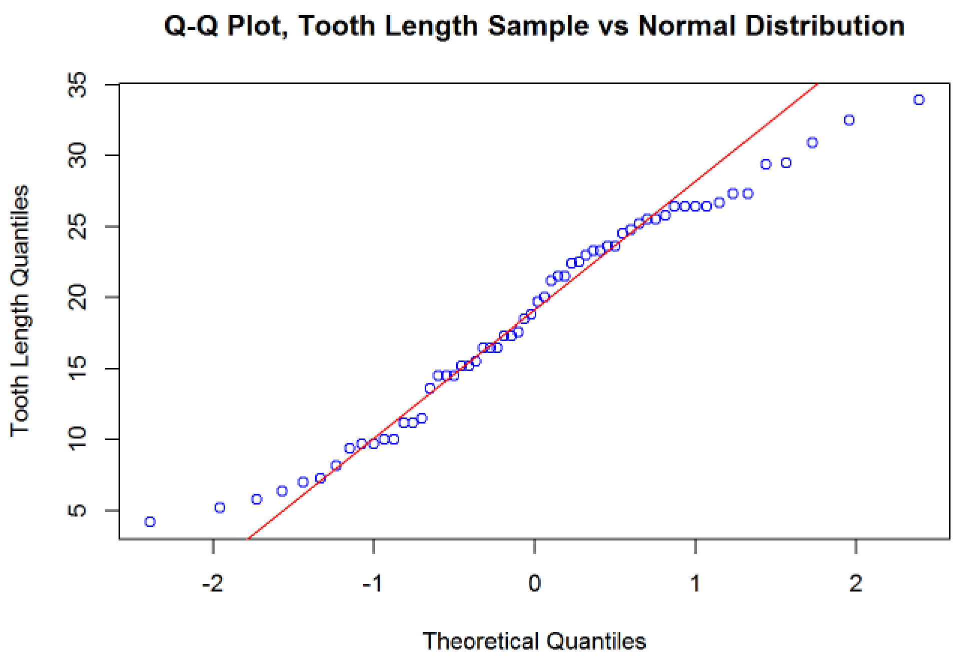
The boxplot below shows tha there is a significant positive correlation between the reponse on tooth length to the dose levels of Vitamin C for both delivery methods. As the dose increses, the tooth lenght increases as well. This type of plot provides a graphical view of the median, quartiles, and max. & min. of the data set. The small points in the box plot 1mg dose group and supplement VC as well as 2mg dose and supplement OJ shows a potential outlier that may need further investigation.

```
library(ggplot2)
graph1 <- ggplot(ToothGrowth, aes(x=factor(dose),y=len,fill=factor(dose)))
graph1 + geom_boxplot(notch=TRUE) + facet_grid(.~supp)  +
scale_x_discrete("Dose in milligrams") + scale_y_continuous("Tooth length") +
  ggtitle("Comparison of tooth growth by suplement type of vitamin C and dose")
```



Q-Q Plots are generated to investigate if the data is normally distributed. It can be observed some deviation in the upper and lower rangers of the data, but there is not evidence of severe skewness.

```
qqnorm(ToothGrowth$len, main="Q-Q Plot, Tooth Length Sample vs Normal Distribution",
ylab="Tooth Length Quantiles", xlab="Theoretical Quantiles", col=4)
qqline(ToothGrowth$len, col=2)
```



Calculation of Mean and Variance by supplement type and dose. It seems that the guinea pigs given orange juice had, on average, longer teeth than those given vitamin C, at the two lower dosage levels. However, at the highest dosage level, there is no visible difference between tooth lengths.

```
with(ToothGrowth, tapply(len, list(supp,dose), mean))

##       0.5     1     2
## OJ 13.23 22.70 26.06
## VC  7.98 16.77 26.14

with(ToothGrowth, tapply(len, list(supp,dose), var))

##        0.5         1         2
## OJ 19.889 15.295556  7.049333
## VC  7.544  6.326778 23.018222
```

Task 3: Tests. Use hypothesis test to investigate the difference in means for OJ and VC. First approach is to perform a t-Test with two runs: variance are equal and variance are unequal, see below. The data shows that the p-values of both equal and unequal variance tests are greater than 0.05 and the confidence intervals contain 0, therefore que cannot reject the null hypothesis. The data does not support the conclusion that there is a significant difference in tootlenght due to supplement method.

```
supplement.equalVariance <- t.test(len~supp, paired=F, var.equal=T, data=ToothGrowth)
supplement.unequalVariance <- t.test(len~supp, paired=F, var.equal=F, data=ToothGrowth)
supplement.testResult <- data.frame("p-value"=c(supplement.equalVariance$p.value,
supplement.unequalVariance$p.value),"Confidence Level-
Low"=c(supplement.equalVariance$conf[1],supplement.unequalVariance$conf[1]),"Confidence
Level-High"=c(supplement.equalVariance$conf[2],supplement.unequalVariance$conf[2]),
row.names=c("Equal Variance","Unequal Variance"))
supplement.testResult

##                     p.value Confidence.Level.Low Confidence.Level.High
## Equal Variance   0.06039337           -0.1670064              7.567006
## Unequal Variance 0.06063451           -0.1710156              7.571016
```

Statistical test to investigate if the variances of the groups OJ and VC can be assumed equal.The fuction var.test in R is used, see below. We can conclude that, based on the ratio of variances and confidence interval, variances are not equal.

```
var.test(ToothGrowth$len[ToothGrowth$supp=="VC"], ToothGrowth$len[ToothGrowth$supp=="OJ"])

## 
##  F test to compare two variances
## 
## data:  ToothGrowth$len[ToothGrowth$supp == "VC"] and ToothGrowth$len[ToothGrowth$supp == 
"OJ"]
## F = 1.5659, num df = 29, denom df = 29, p-value = 0.2331
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.745331 3.290028
## sample estimates:
## ratio of variances 
##           1.565937
```

Run one-way ANOVA that regresses toothlenght data (len) on supplement method (supp). The ANOVA results show that the effect of supplement is not significant at the .05 level (p-value from the data analysis is 0.06039). It should be noticed that this test assumes that the other factor (dose) has no effect.

```
ANOVA_result <- lm( len ~ supp, data=ToothGrowth)
summary( ANOVA_result )

## 
## Call:
## lm(formula = len ~ supp, data = ToothGrowth)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.7633  -5.7633   0.4367   5.5867  16.9367 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   20.663      1.366  15.127   <2e-16 ***
## suppVC        -3.700      1.932  -1.915   0.0604 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.482 on 58 degrees of freedom
## Multiple R-squared:  0.05948,    Adjusted R-squared:  0.04327 
## F-statistic: 3.668 on 1 and 58 DF,  p-value: 0.06039
```

Run full two-way ANOVA. In this analysis, we notice that the main effect of supplement type is now much more significant than it was before, under the assumption that the effects of dose were zero.

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

## 
## Call:
## lm(formula = len ~ supp * dose, data = ToothGrowth)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.2264 -2.8462  0.0504  2.2893  7.9386
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.550      1.581   7.304 1.09e-09 ***
## suppVC        -8.255      2.236  -3.691 0.000507 ***
## dose           7.811      1.195   6.534 2.03e-08 ***
## suppVC:dose    3.904      1.691   2.309 0.024631 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.083 on 56 degrees of freedom
## Multiple R-squared:  0.7296, Adjusted R-squared:  0.7151
## F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16
```

CONCLUSIONS: The dataset in R, ToothGrowth, was analyzed using some of the exploratory data and statistical tools learned in this course. The data set contains results of a study of the effects of vitamin C on the growth of incisors in guinea pigs. Hypothesis test were employed to assess if the data supports the contention that if vitamin C in orange juice is better for pig tooth growth than given vitamin as ascorbic acid. Linear models that regress the response variable (tooth length) on each factor separately (one-way ANOVA's) and on both factors (two-way ANOVA) were presented. T.test and var.test showed that the is not a significant difference in the mean of OJ and VC and that the variance cannot be assumed equal. The two-way ANOVA shows that the effect of supplement type is significant when considering the effect of dose. The assumptions made in the t-test for the difference between two means are: populations are normally distributed, samples are independent, populations have equal variance, sample size is small thus the critical value is t-value from the t-distribution. The assumptions for ANOVA analysis are: the population values for each combination of pairwise factor levels are normally distributed, the variance for each population are equal, the samples are independent, the data measure is interval or ratio level.

Reference 1.- http://en.wikipedia.org/wiki/Exponential_distribution

Note: This report consist of 5 pages, this fulfill the project requirement page length.