Tooth Growth Exploratory Data Analysis

Daniel Rodrigues Ambrosio

June 20th, 2015

1. Overview

This is the part 2 of the Project for the Statistical Inference course in Data Science Specialization track from Coursera.

The goal is to analyze the ToothGrowth data in the R datasets package, lLoad the ToothGrowth data and perform some basic exploratory data analysis.

The study shall provide a basic summary of the data and: 1. Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering) 2. State your conclusions and the assumptions needed for your conclusions. 3. Perform an exploratory data analysis of at least a single plot or table highlighting basic features of the data 4. Perform some relevant confidence intervals and/or tests 5. Assure that the results of the tests and/or intervals are interpreted in the context of the problem correctly 6. Describe the assumptions needed for the conclusions

1.1 Basis for the study

The data is set of 60 observations, 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) with each of two delivery methods (orange juice or ascorbic acid).

1.2 Environment

Being able to reproduce every step of a data analysis is a crucial aspect of the data science. That being said, all the libraries used as support for this analysis are listed below and so is the system information.

```
library(datasets)
library(ggplot2)

sessionInfo()
```

```
## R version 3.1.2 (2014-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_1.0.1
##
## loaded via a namespace (and not attached):
##  [1] colorspace_1.2-6 digest_0.6.8     evaluate_0.7     formatR_1.2
##  [5] grid_3.1.2       gtable_0.1.2     htmltools_0.2.6  knitr_1.10.5
```

```
##  [9] MASS_7.3-35      munsell_0.4.2    plyr_1.8.1       proto_0.3-10
## [13] Rcpp_0.11.4      reshape2_1.4.1   rmarkdown_0.3.10 scales_0.2.4
## [17] stringr_0.6.2    tools_3.1.2      yaml_2.1.13
```

2. Data Summady and Exploratory Data Analysis

Load the data and get a brief description of its content.

```
    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 ...
```

There are two factors for suplement: “OJ” and “VC”, but it is not possible to determine how many values for the dosage, so let us find out.

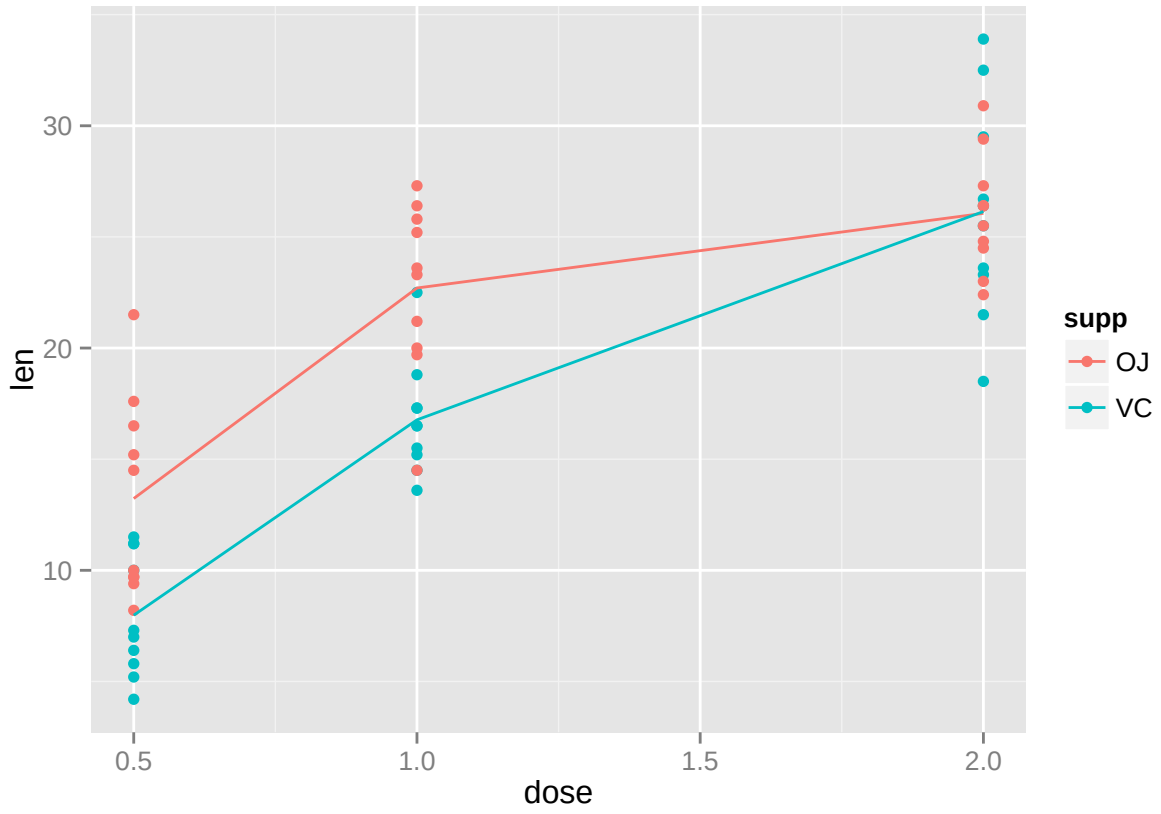
```
    unique(ToothGrowth$dose)
```

```
## [1] 0.5 1.0 2.0
```

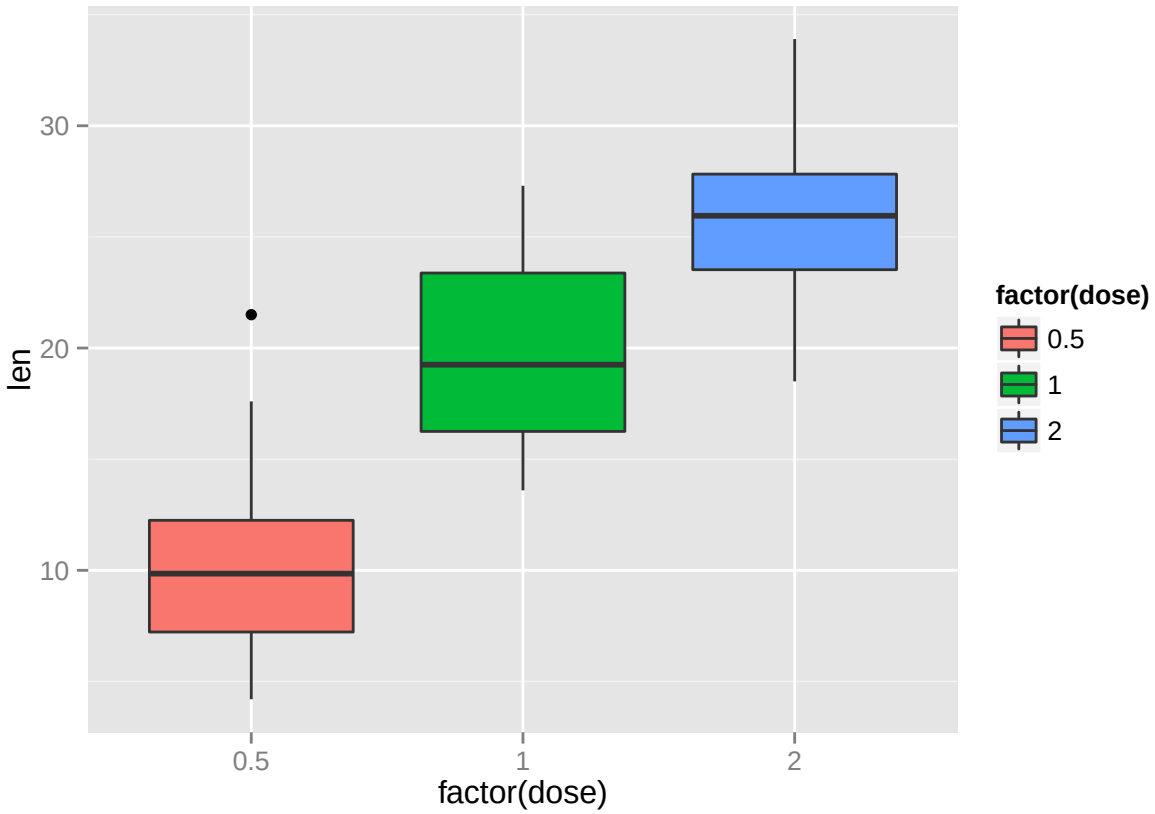
2.1 Visual Analysis

Let us explore visually the possible relation bewtween tooth length and Vitamin C dose using a Scatterplot with its average. Next we will use a boxplot to show the same relation.

```
    # Calculate the mean for every dose and supp
    avg <- aggregate(len~.,data=ToothGrowth,mean)
    g <- ggplot(aes(x=dose, y = len), data = ToothGrowth) +
        geom_point(aes(color = supp))
    g <- g + geom_line(data=avg,aes(group=supp,colour=supp))
    print(g)
```

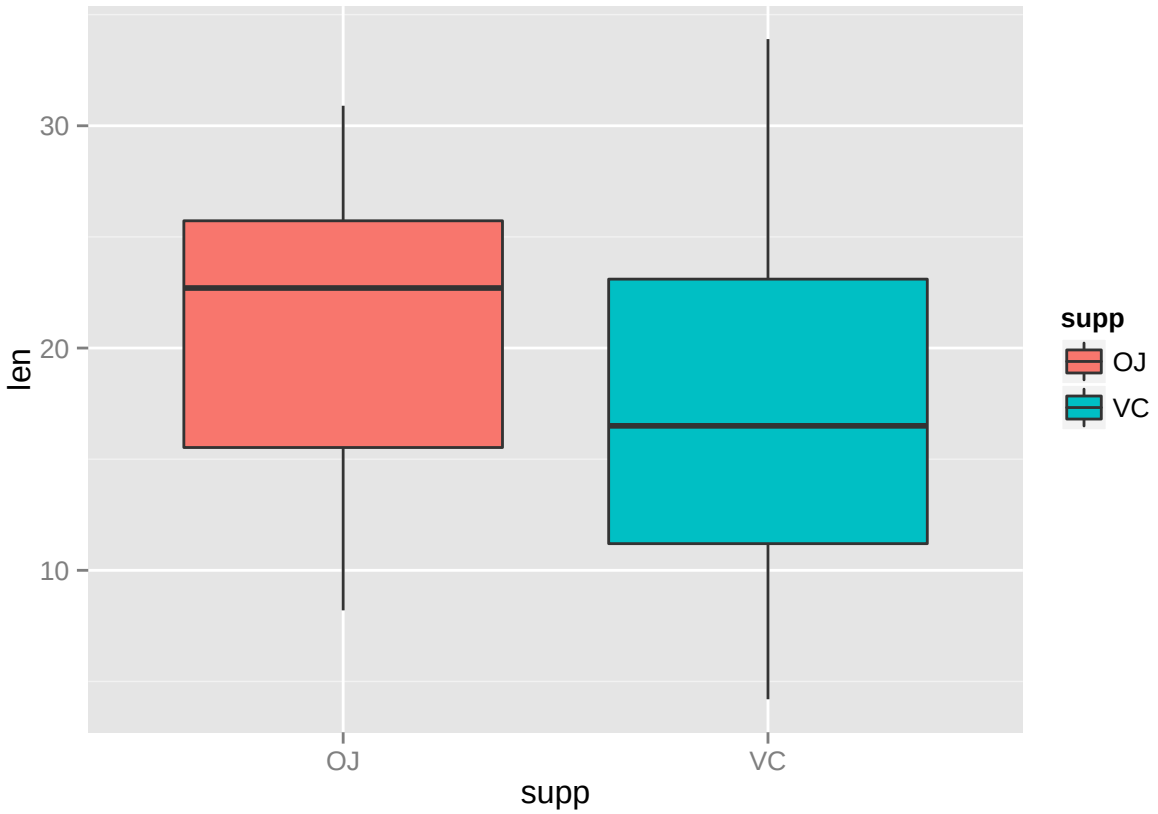



```
ggplot(aes(x = factor(dose), y = len), data = ToothGrowth) +
    geom_boxplot(aes(fill = factor(dose)))
```



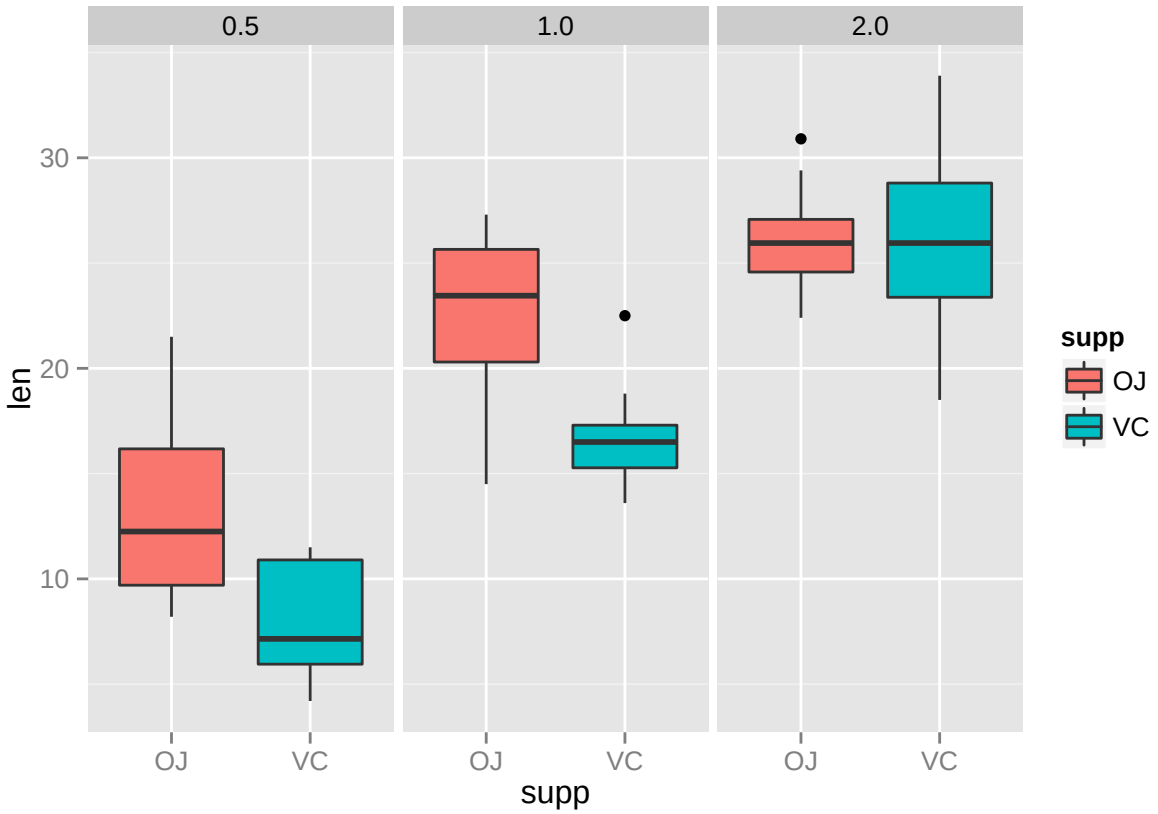
Now let us explore visually the possible relation between tooth length and the delivery methods using a boxplot.

```
    ggplot(aes(x = supp, y = len), data = ToothGrowth) +
        geom_boxplot(aes(fill = supp))
```



Now let us check what might be the relation between delivery methods at each dose level in a boxplot.

```
ggplot(aes(x = supp, y = len), data = ToothGrowth) +
    geom_boxplot(aes(fill = supp)) + facet_wrap(~ dose)
```



This initial exploratory data analysis show us that the dosage affects the tooth length - the larger the dosage, the longer the tooth.

The relation between supplement type howwever is not that obvious at this stage. When using Vitamin C as a supplement, the more vitaming given, the more the teeth grew. When the dosage is low, orange juice seems to correlate with longer teeth, but at higher dosages (2.0mg) there is no significant difference.

2.2 Numeric Analysis

```
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
```

Combining the dosage and the delivery method to analyse the statistical data for that combination.

```
by(ToothGrowth$len, INDICES = list(ToothGrowth$supp, ToothGrowth$dose), summary)
```

```
## : OJ
## : 0.5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.20    9.70   12.25   13.23   16.18   21.50
## ------------------------------------------------------------
## : VC
## : 0.5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.20    5.95    7.15    7.98   10.90   11.50
## ------------------------------------------------------------
## : OJ
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.50   20.30   23.45   22.70   25.65   27.30
## ------------------------------------------------------------
## : VC
## : 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   15.27   16.50   16.77   17.30   22.50
## ------------------------------------------------------------
## : OJ
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.40   24.58   25.95   26.06   27.08   30.90
## ------------------------------------------------------------
## : VC
## : 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.50   23.38   25.95   26.14   28.80   33.90
```

3. Confidence Intervals and Hypothesis Testing

The next two sections are for analyzing the data for correlation between the delivery method (Dosage and Supplement) and change in tooth growth.

3.1 Dosage as a Factor

```
    dose1 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
    dose2 <- subset(ToothGrowth, dose %in% c(0.5, 2.0))
    dose3 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))
    t.test(len ~ dose, paired = F, var.equal = F, data = dose1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```
    t.test(len ~ dose, paired = F, var.equal = F, data = dose2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

```
    t.test(len ~ dose, paired = F, var.equal = F, data = dose3)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

The confidence intervals ([-11.98, -6.276] for doses 0.5 and 1.0, [-18.16, -12.83] for doses 0.5 and 2.0, and [-8.996, -3.734] for doses 1.0 and 2.0) allow for the rejection of the null hypothesis and a confirmation that there is a significant correlation between tooth length and dose levels.

3.2 Supplement as a Factor

Analyzing the data for correlation between the delivery method and change in tooth growth:

```
    t.test(len ~ supp, paired = F, var.equal = F, data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

A confidence interval of [-0.171, 7.571] does not allow us to reject the null hypothesis (that there is no correlation between delivery method and tooth length).

4. Conclusions and Assumptions

4.1 Assumptions

In order to make conclusions with the data in this dataset, we must assume the following:

1. The poplulations are independent, the variances between populations are different and a random population was used
2. The population was comprised of similar guinea pigs, measurement error was accounted for with significant digits, and double blind research methods were used.
3. For the populations to be independent, 60 guinea pigs would have to be used so each combination of dose level and delivery method were not affected by the other methods.
4. To ensure double blind research methods are followed, the researchers taking the measurements must have been unaware of which guinea pigs were given which dose level or delivery method.
5. The guinea pigs must also be unaware that they are being given a specific treatment.

4.2 Conclusions

1. Supplement type has no effect on tooth growth.
2. Increasing the dose level leads to increased tooth growth.

5. Appendix

1. You can find the original RPub file used to build this document on Daniel Ambrosio's repository: RPub original document