Tooth Growth dataset analysis

Course Project for Statistical Inference Coursera Class

Leon Duplay

16 July 2015

Overview

In this document, we will investigate the ToothGrowth dataset by performing some basic exploratory data analysis and some statistical inference. By using confidence intervals and/or hypothesis tests to compare tooth growth by `supp` and `dose`, we will study the impact of these two variables on tooth growth and present our conclusions, including the assumptions made.

Exploring the dataset

In this section, we will load the ToothGrowth dataset and perform some exploratory data analysis, before performing our hypothesis testing in the next section.

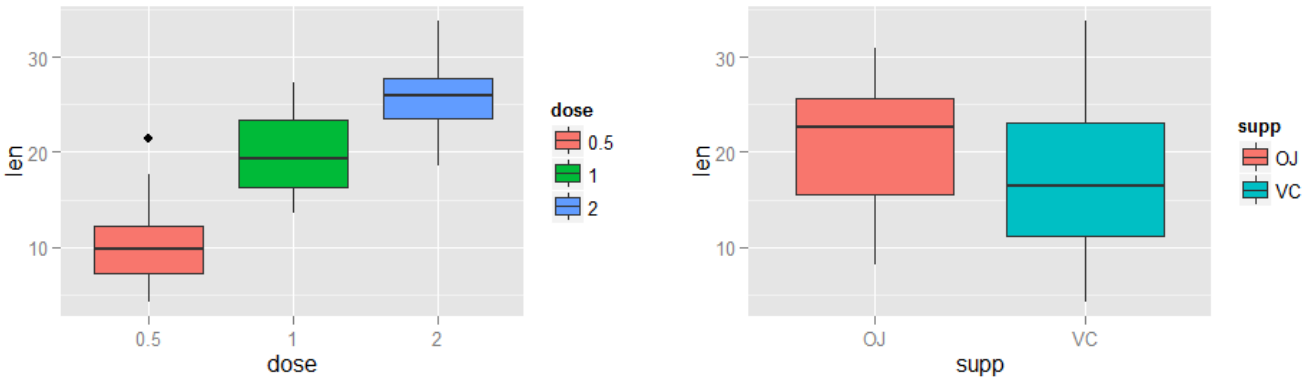
```
# Libraries
library(datasets)
library(ggplot2)
library(grid)
library(gridExtra)

# Load data & basic data info
data(ToothGrowth)
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
summary(ToothGrowth)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

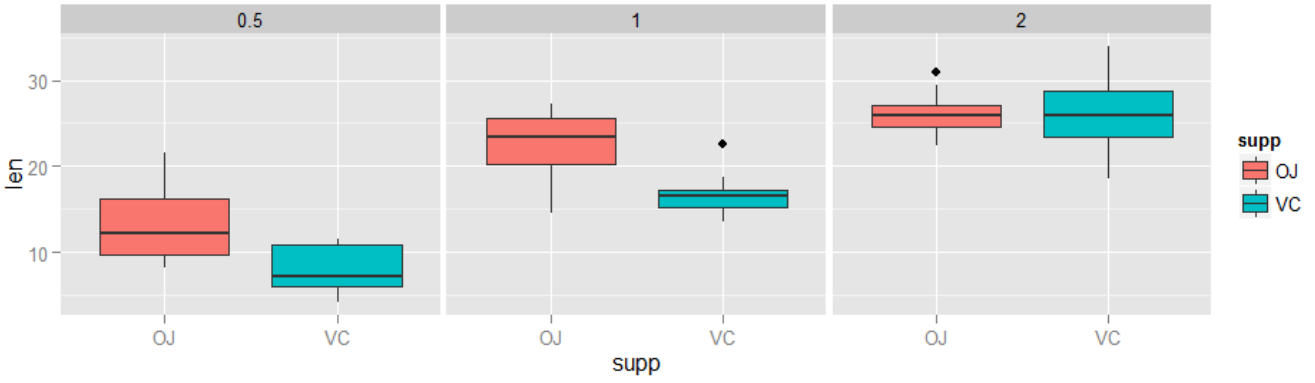
The data represents 60 observations, with the length of teeth in each of 10 guinea pigs after following three dose levels of Vitamin C (0.5, 1 and 2 mg) with each of two delivery methods (orange juice OJ or ascorbic acid AC).

We'll now explore the data using some simple boxplots to get an idea if `supp` and `dose` have an impact on tooth length and in which way.

```
# Plot Length vs dose & supp
g1 <- ggplot(aes(x = dose, y = len), data = ToothGrowth) + geom_boxplot(aes(fill = dos
e))
g2 <- ggplot(aes(x = supp, y = len), data = ToothGrowth) + geom_boxplot(aes(fill = sup
p))
grid.arrange(g1, g2, ncol=2)
```




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



In the graphs above, we can see a clear trend: the larger the dosage, the longer the tooth. However, the effectiveness of delivery method is not clear, both OJ and VC have roughly the same performance (with OJ having a slightly higher mean).

Looking at both variables together gives an interesting insight: it seems that at lower doses OJ is more effective than VC in terms of tooth length, but at high dosage (2mg), the performance is the same.

Confidence Intervals and Hypothesis Testing

The objective of this section is to use confidence intervals and hypothesis testing to prove or disprove the null hypothesis (dosage/supplement have no impact on teeth length). To do so, we will use unpaired T-tests between factors of `dose` and `supp`.

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

```
##
##  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
```

In this T-test testing for supplements, while the means are different **(20.66, 16.96)**, we cannot disprove the null hypothesis as the probability under the null hypothesis of obtaining these means is over 5% (p value is **0.061**). Therefore, we can conclude that supplement type **does NOT** have a statistically significant impact on teeth length within the given sample.

In order to use T tests on the `dose`, we must split the dataset into pairs of dosages.

```
    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))
    t1 <- t.test(len ~ dose, paired = F, var.equal = F, data = dose1)
    t2 <- t.test(len ~ dose, paired = F, var.equal = F, data = dose2)
    t3 <- t.test(len ~ dose, paired = F, var.equal = F, data = dose3)
```

The p values for the T-tests are **1.268300710^{-7}, 4.39752510{-14}**, and **1.906429510^{-5}**. Since these values are all under 0.05, we can disprove the null hypothesis and conclude that dosage **does** have a statistically significant impact on teeth length within the given sample.

Assumptions

For all these hypothesis tests, we are assuming that our guinea ping subjects represent a population that is IID (independant and identically distributed), and that:

- The guinea pigs are repesentative of the whole population and follow a random sample.
- Tooth length shows a normal distribution
- Observations are independent of each other

Appendix

This analysis was completed with the below system:

```
sessionInfo()
```

```
## R version 3.1.3 (2015-03-09)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] gridExtra_0.9.1 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] gtable_0.1.2     htmltools_0.2.6  knitr_1.10.5     labeling_0.3
##  [9] magrittr_1.5     MASS_7.3-42      munsell_0.4.2    plyr_1.8.3
## [13] proto_0.3-10     Rcpp_0.11.6      reshape2_1.4.1   rmarkdown_0.7
## [17] scales_0.2.5     stringi_0.5-5    stringr_1.0.0    tools_3.1.3
## [21] yaml_2.1.13
```