Basic inferential data analysis - Course Project - Statistical Inference

attila.toth86

23 Sep 2015

This is the project for the statistical inference class. In this, I use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

This document aims to fulfill requirements of second task.

Loading Data

I will use `ToothGrowth` data in the R datasets package to my further analyzes.

I load `ToothGrowth` data into my working environment and insert their content to a `TG` data frame.

```
data(ToothGrowth)
TG <- ToothGrowth
```

I looked up R documentation (`?ToothGrowth`) for more descriptive information about this dataset and found the following description:

> The Effect of Vitamin C on Tooth Growth in Guinea Pigs: The response is the 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).

Format of the dataset:

- `len`: numeric, tooth length
- `supp`: factor, supplement type (VC or OJ)
- `dose`: numeric, dose in miligrams

According to the description above, my best guess was to assign *VC* to delivery method of Vitamin C via ascorbic acid and *OJ* to orange juice supplement type.

Exploratory data analyses & summary

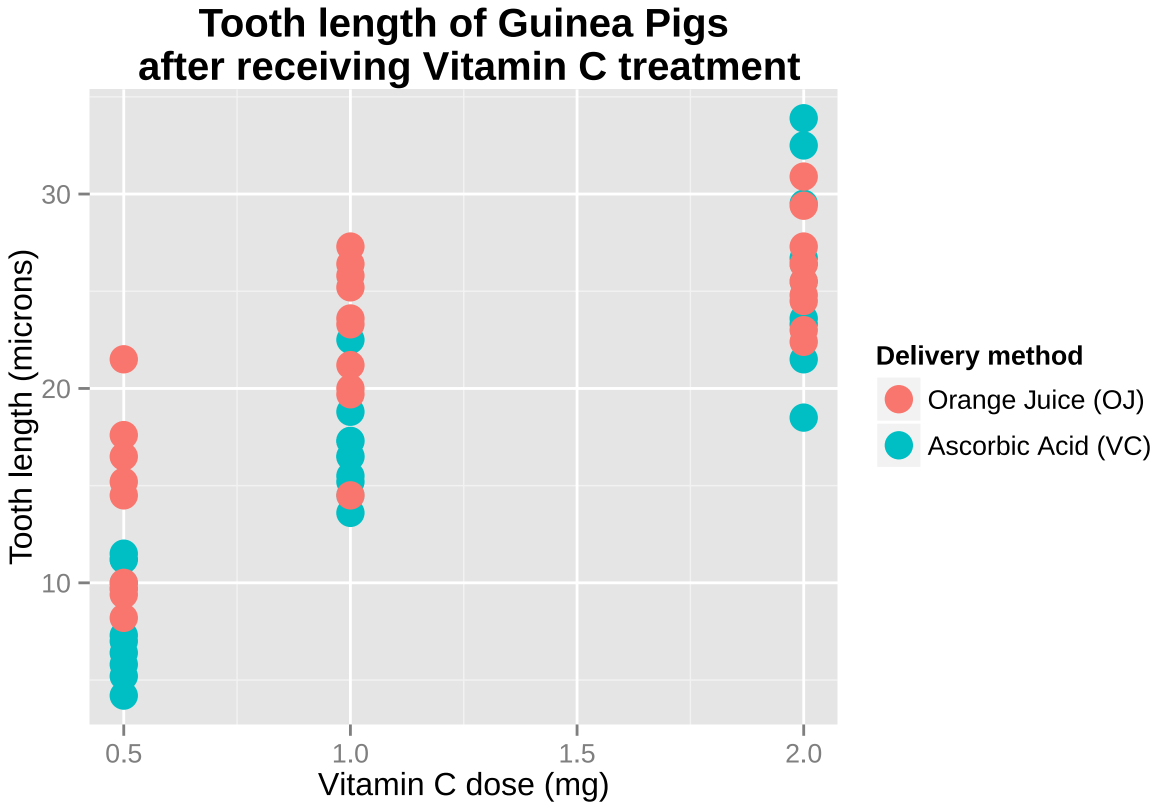
By looking at the data, description (at least the reference to format) seems to be inline with what we see:

```
str(TG)
```

```
## '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 ...
```

To visualize what we have, I simply plotted our dataframe.

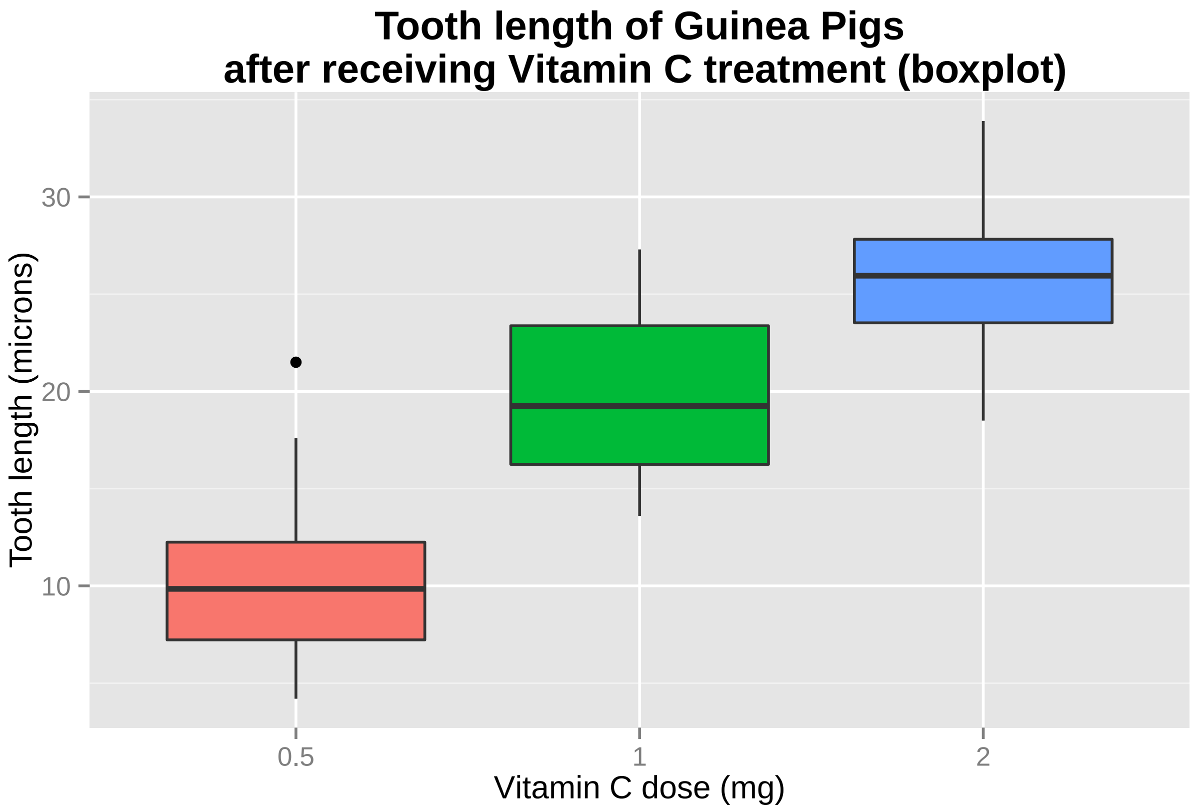
```r
# Rename factors for meaningful names
levels(TG$supp)[levels(TG$supp)=="OJ"] <- "Orange Juice (OJ)"
levels(TG$supp)[levels(TG$supp)=="VC"] <- "Ascorbic Acid (VC)"

# Using ggplot2 package to create figures
library(ggplot2)

ggplot() +
geom_point(data = TG,
           aes(x = dose, y= len, color = supp),
           size = 5
           ) +
ylab("Tooth length (microns)") +
xlab("Vitamin C dose (mg)") +
ggtitle("Tooth length of Guinea Pigs\n after receiving Vitamin C treatment") +
theme(plot.title=element_text(face="bold", size=15)) +
guides(color=guide_legend(title="Delivery method")) +
guides(size=FALSE)
```



This figure above shows that there might be a correlation between measured tooth length and dosed vitamin C. I tend to say that subjects that received higher amount of vitamin C most likely have grown longer teeth regardless of delivery method.

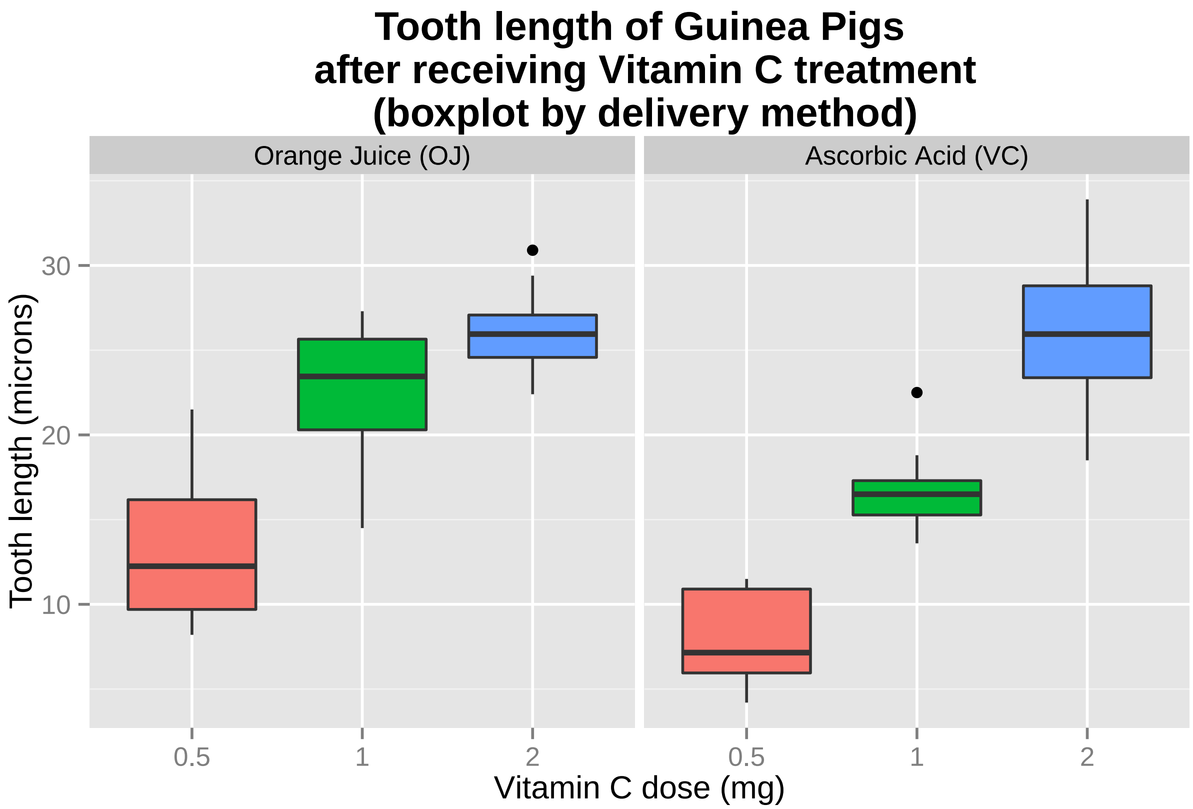
To support my previous statement, let's look at the following boxplot that shows the average of tooth length by each group:

```
ggplot() +
geom_boxplot(data = TG, aes(x = factor(dose), y= len, fill=factor(dose))) +
ggtitle("Tooth length of Guinea Pigs\n after receiving Vitamin C treatment (boxplot)") +
theme(plot.title=element_text(face="bold", size=15)) +
ylab("Tooth length (microns)") +
xlab("Vitamin C dose (mg)") +
guides(fill=FALSE)
```



If I separated the population by delivery method, the trend wouldn't break (i.e. higher dosage of vitamin C ~ greater tooth length) but a notable difference can be observed in lower level of dosage of Vitamin C:

```
ggplot() +
geom_boxplot(data = TG, aes(x = factor(dose), y= len, fill=factor(dose))) +
facet_grid(.~supp) +
ggtitle("Tooth length of Guinea Pigs\n after receiving Vitamin C treatment\n (boxplot by delivery metho
theme(plot.title=element_text(face="bold", size=15)) +
ylab("Tooth length (microns)") +
xlab("Vitamin C dose (mg)") +
guides(fill=FALSE)
```



Apparently, delivery method might be a factor in tooth growth in lower doses.

Confidence Intervals & Tests

In order to step further in my analysis I brake the entire population into 6 groups by Vitamin C dosage & delivery method.

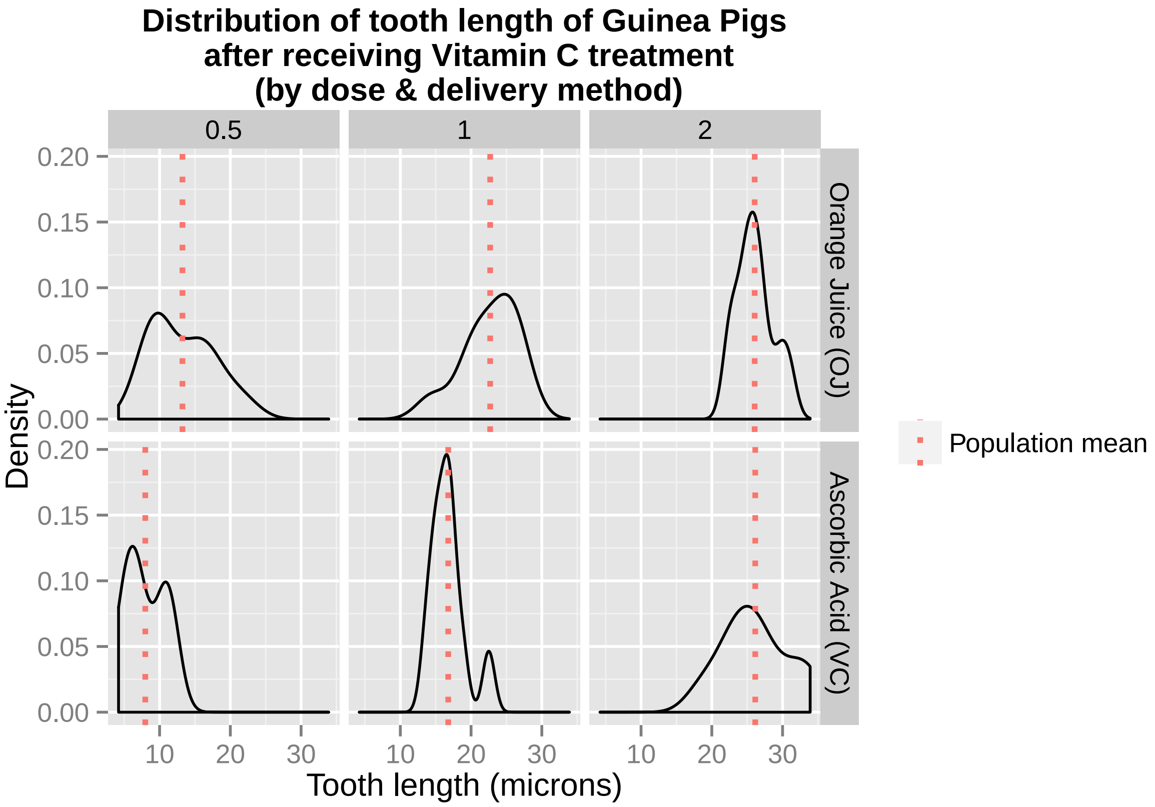
I created some summary statistics by delivery method & dose:

```
TGsum <- cbind(
               aggregate(TG$len, by=list(TG$supp, TG$dose), FUN=mean),
               aggregate(TG$len, by=list(TG$supp, TG$dose), FUN=sd)[3])
colnames(TGsum) <- c("supp","dose","Mean","StandardDeviation")
```

Delivery method	Dose	Mean of tooth length	Standard Deviation of tooth length
Orange Juice (OJ)	0.5	13.23	4.4597085
Orange Juice (OJ)	1.0	22.7	3.9109533
Orange Juice (OJ)	2.0	26.06	2.6550581
Ascorbic Acid (VC)	0.5	7.98	2.7466343
Ascorbic Acid (VC)	1.0	16.77	2.5153087
Ascorbic Acid (VC)	2.0	26.14	4.7977309

Figure below shows the distribution of each group (together with their mean - red-dotted line):

```
ggplot() +
geom_density(data=TG, aes(x=len)) +
geom_vline(data=TGsum,
           aes(x=Mean, xintercept=Mean, color="red"),
           linetype=3,
           size=1,
           show_guide=TRUE) +
facet_grid(supp ~ dose) +
ylab("Density") +
xlab("Tooth length (microns)") +
ggtitle("Distribution of tooth length of Guinea Pigs\n after receiving Vitamin C treatment\n (by dose &
theme(plot.title=element_text(face="bold", size=12)) +
guides(color=guide_legend(title=NULL)) +
scale_colour_discrete(labels="Population mean")
```



To compare the mean of tooth length between the (independent) groups receiving Vitamin C via different delivery method over each dosage, I used unpaired t confidence interval (t-test):

```
# Separating groups
OJ05 <- TG[TG$supp=="Orange Juice (OJ)" & TG$dose==0.5,1]
OJ1 <- TG[TG$supp=="Orange Juice (OJ)" & TG$dose==1,1]
OJ2 <- TG[TG$supp=="Orange Juice (OJ)" & TG$dose==2,1]
VC05 <- TG[TG$supp=="Ascorbic Acid (VC)" & TG$dose==0.5,1]
VC1 <- TG[TG$supp=="Ascorbic Acid (VC)" & TG$dose==1,1]
VC2 <- TG[TG$supp=="Ascorbic Acid (VC)" & TG$dose==2,1]

# Getting test statistics
t.test(VC05, OJ05, paired = FALSE, var.equal = FALSE)
```

```r
t.test(VC1, OJ1, paired = FALSE, var.equal = FALSE)
t.test(VC2, OJ2, paired = FALSE, var.equal = FALSE)
```

Dosage	T Confidence Interval	P-Value
0.5	-8.7809427, -1.7190573	0.0063586
1.0	-9.0578518, -2.8021482	0.0010384
2.0	-3.6380705, 3.7980705	0.9638516

Conclusions

For doses of 0.5 & 1 mg my test statistic show that there is significant difference (on 5% significance level) between average tooth length in observed groups with different delivery method.

- For these doses null hypothesis (true difference in means is equal to 0) must be rejected because p-value shows less than 1% probability of obtaining more extreme results than what was actually observed.

In contrast with previous doses, groups treated with 2 mg of Vitamin C don't show significant differences whether they received treatments via ascorbic acid or orange juice.

- For this dose null hypothesis can not be rejected because of the high p-value.

Apparently, on lower doses the way of how Vitamin C is delivered matters. For groups that received treatment via orange juice showed more intensive growth compared to groups that were treated via ascorbic acid. This difference in higher amount disappeared, no significant difference could be revealed. However, the highest growth has been measured in the groups of 2 mg treatment.