Peer-graded Assignment Part 2: Inferential Data Analysis

Rogelio Caballero

May 27, 2017

Abstract

What follows is an exploratory analysis of the ToothGrowth dataset and some statistical tests performed on it as part of the Statistical Inference Course of Johns Hopkins University at Coursera.

Exploring *ToothGrowth*

`ToothGrowth` contains information about the effects of vitamine C on the growth of odontoblasts of 60 guinea pigs. Let's load it:

```
data("ToothGrowth")
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```
nrow(ToothGrowth)
```

```
## [1] 60
```

As we can see is a dataframe with 60 observations of 3 variables. `len` is the length of the odontoblasts (I haven't been able to find the units), `supp` is the type of supplement though which the vitamine C is given to the subjects and `dose` is given in mg/day. Let's take a look at the supplements and the doses:

```
levels(ToothGrowth$supp)
```

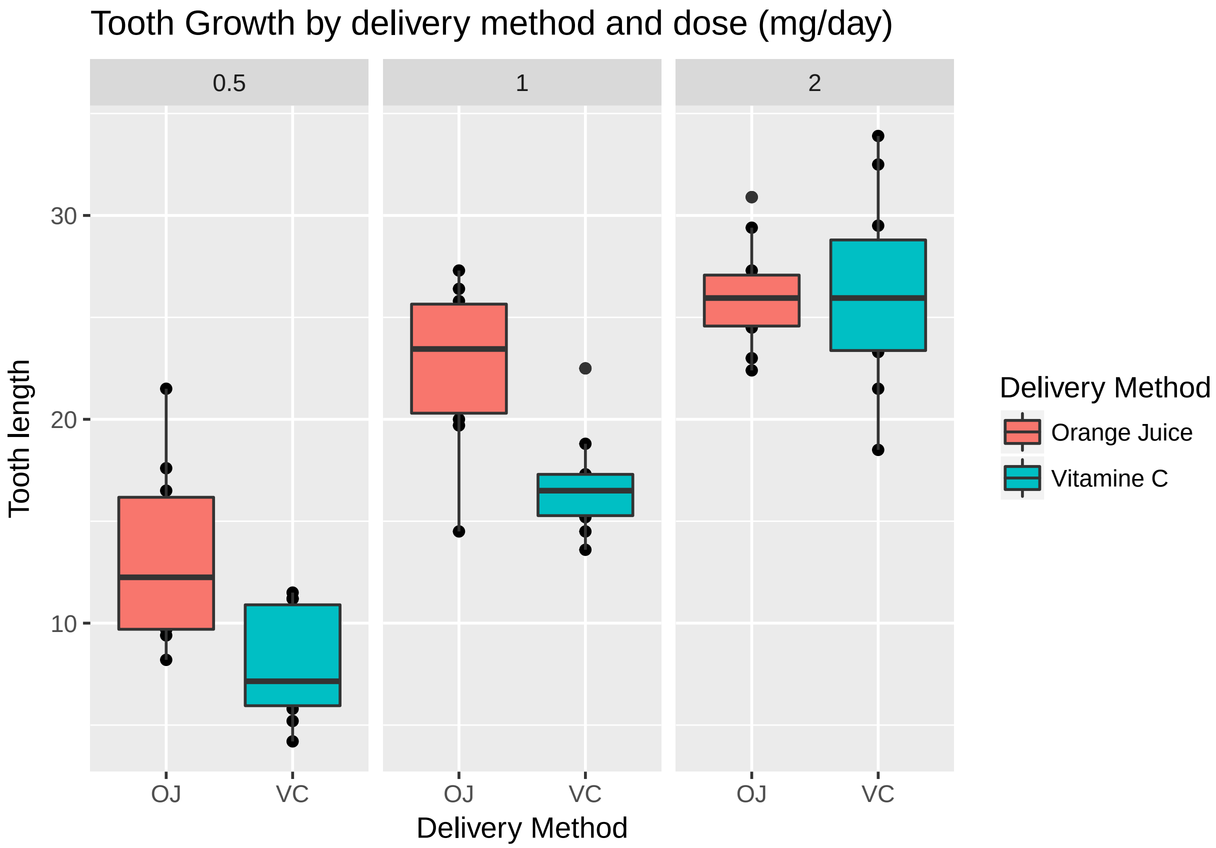
```
## [1] "OJ" "VC"
```

"OJ" stands for "orange juice" while "VC" stands for "vitamine C", these are the delivery methods though which the doses were given to the pigs. We can also check all the values of `dose`:

```
levels(as.factor(ToothGrowth$dose))
```

```
## [1] "0.5" "1"   "2"
```

Now let's see some boxplots to get an idea about the behavior of `len` varies with the other parameters. As we can see for doses of $0.5mg/day$ and $1mg/day$ the `"OJ"` delivery method *seems* to produce greater tooth growth than `"VC"`. For the $2.0mg/day$ dose there seems to be no significant difference. We will test that hypothesis in the second part of this exercise:



Observe that boxes show the median and quantiles of the data while in hypothesis testing we are more interested in means and standard deviations. Let's take a look at those:

```
df <- aggregate(len~supp + dose, data = ToothGrowth,
         FUN = function(x)c(mn = mean(x), sd = sd(x)))
```

```
xt <- xtable(do.call(data.frame, c(df, check.names = FALSE)))
print(xt, type = "latex")
```

	supp	dose	len.mn	len.sd
1	OJ	0.50	13.23	4.46
2	VC	0.50	7.98	2.75
3	OJ	1.00	22.70	3.91
4	VC	1.00	16.77	2.52
5	OJ	2.00	26.06	2.66
6	VC	2.00	26.14	4.80

As we can see the only dose for which the results of both delivery methods seem to be equal is $2mg/day$: each mean lies less than one standard deviation of the other.

Hypothesis testing

We are going to run some t-tests to see whether the difference in tooth growth for different delivery methods is significant.

```
a <- subset(ToothGrowth, supp == "OJ" & dose == 0.5)$len
b <- subset(ToothGrowth, supp == "VC" & dose == 0.5)$len
```

```r
t.test(a,b)$conf.int
```

```
## [1] 1.719057 8.780943
## attr(,"conf.level")
## [1] 0.95
```

As we can see 0 lies outside the 95% confidence interval which suggests that we can reject the null hypothesis (mean growth is the same for both supplements). Let's see the confidence intervals of the other doses:

```r
t.test(subset(ToothGrowth, supp == "OJ" & dose == 1)$len,
       subset(ToothGrowth, supp == "VC" & dose == 1)$len)$conf.int
```

```
## [1] 2.802148 9.057852
## attr(,"conf.level")
## [1] 0.95
```

```r
a1 <- subset(ToothGrowth, supp == "OJ" & dose == 2)$len
b1 <- subset(ToothGrowth, supp == "VC" & dose == 2)$len
t.test(a1, b1)$conf.int
```

```
## [1] -3.79807  3.63807
## attr(,"conf.level")
## [1] 0.95
```

We can reject the null hypothesis in favor of greater tooth growth for orange juice for doses of $1mg/day$. In the case of $2mg/day$, as we suspected, 0 lies inside the 95% confidence interval. As a consequence we fail to reject the null hypothesis.

Let's compare the *power* of our tests for the $0.5mg/day$ and $2.0mg/day$ doses:

```r
power.t.test(n = 10, delta = mean(a) - mean(b),
             sd = sqrt((var(a)+var(b))/2), alternative = "one.sided", type = "two.sample")$power
```

```
## [1] 0.9195986
```

```r
power.t.test(n = 10, delta = mean(a1) - mean(b1),
             sd = sqrt((var(a1)+var(b1))/2), alternative = "one.sided", type = "two.sample")$power
```

```
## [1] 0.04558198
```

As we can see, in the case of $2mg/day$ power is low due to the fact that the difference between the means is small (small effect size).

Conclusions

For $0.5mg/day$ and $1.0mg/day$ the data supports the hypothesis that orange juice produces greater tooth growth as a supplement. In the case of $2.0mg/day$ dose, the data supports the hypothesis of equal mean tooth growth for both supplements. The low power of the test for the latter dose is consistent with the hypothesis testing.

An important assumption that we've made is that data is symmetric around the mean in order to apply Welch t-tests. This is not completely accurate since tooth growth cannot be negative, while a variable that strictly follows a t-distribution takes negative values to maintain symmetry.