Statistical Inference Course Project Part 2

Johns Hopkins University - Coursera

by Massimo Malandra

Overview:

The "ToothGrowth" dataset describes the effect of Vitamin C on tooth growth in guinea pigs. As the description of the dataset itself suggests: the response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, (orange juice or ascorbic acid (a form of vitamin C and coded as VC).

Exploratory data analysis and summary of the data:

Data are structured in a data frame with 60 observations on 3 variables. The 3 variables represent the tooth length (len), the supplement type (supp) and the dose in milligrams per day (dose).

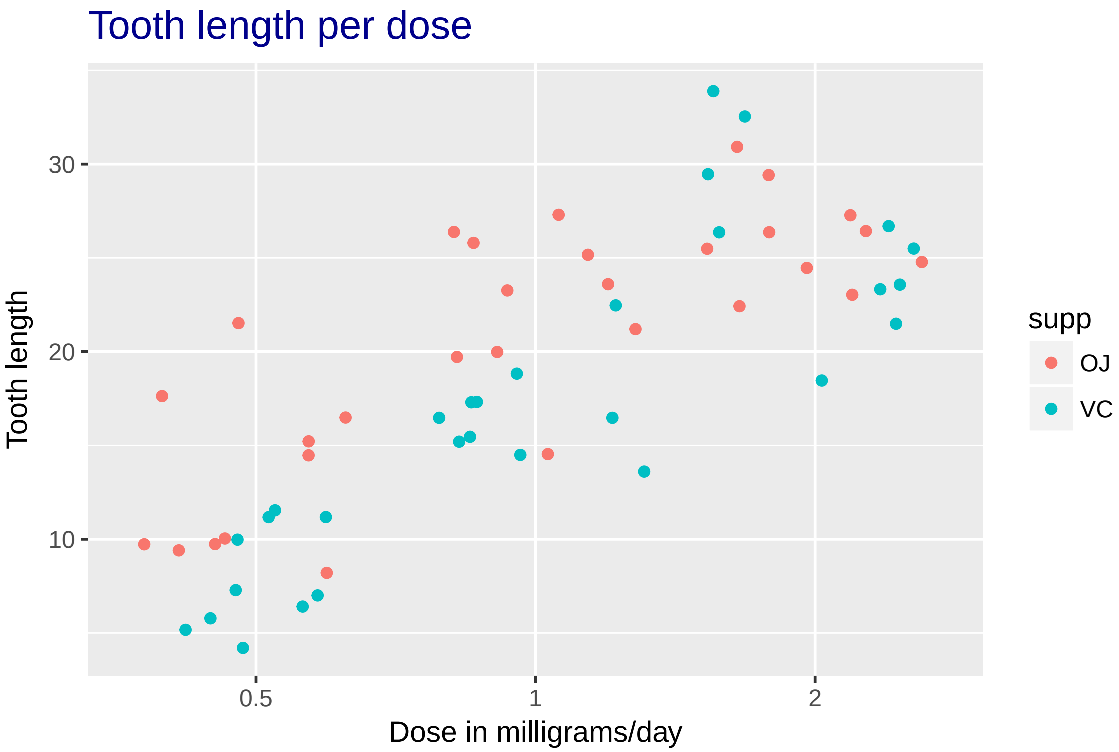
```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
##       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  
```

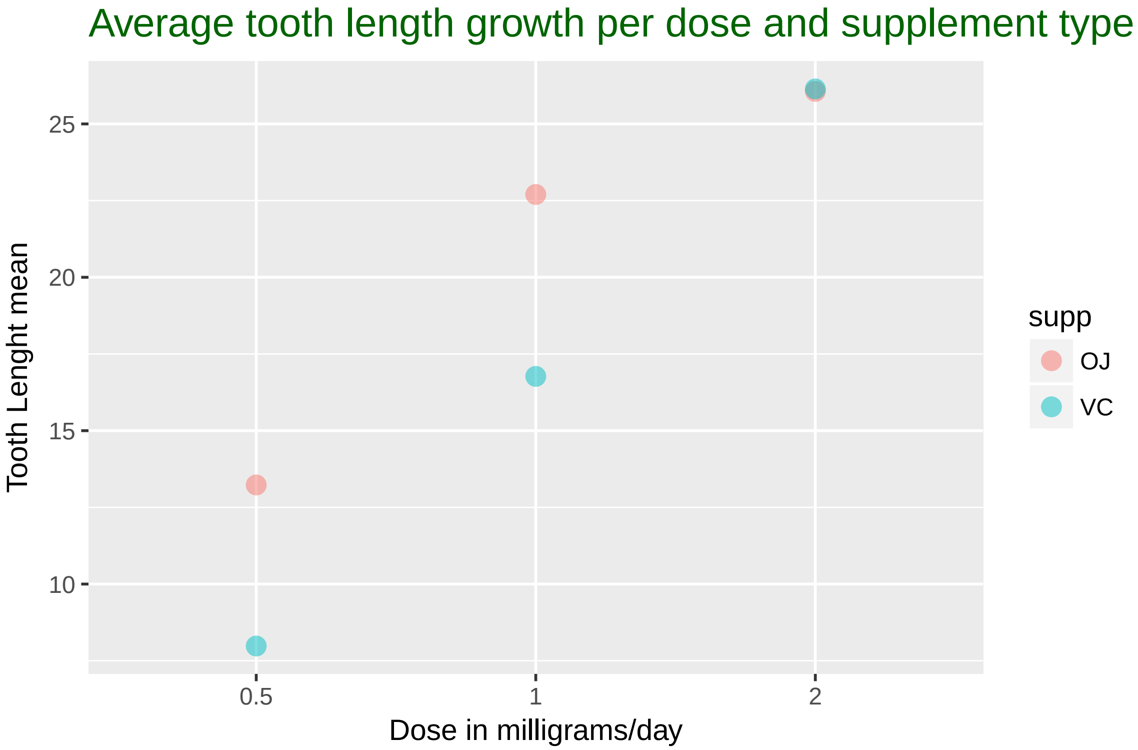
```
## Standard deviation of tooth length: 7.65
```

As we can see from the summary above, the dose variable has only 3 possible values (0.5, 1 and 2) which represent the level of vitamin C received by the animal in each dose. This suggests that we can convert the dose variable from integer to factor. From this first preliminary analysis we can see that the tooth length variable goes from a minimum of 4.20 to a maximum of 33.90, with a standard deviation of 7.65.

If we take a quick look at the dispersion of the data, adding a small amount of random variation to the loaction of each point, thanks to the jitter geom, we can see a few interesting properties. First of all it can be seen how the increase of dosage generates a greater impact on tooth length. Moreover, if we look at the difference between the 2 supplement type, we can see that at 0.5 and 1.0 level of dosage of the OJ points (light red) are quite separated from the VC points (light blue), suggesting a stronger impact of tooth length of OC rather than VC. The same thing cannot be said for the 2.0 level of dosage, for which it seems that the impact of OJ and VC is quite similar.


Tooth length per dose

The same conclusion can be strengthen if we plot the means of tooth growth per each supplment type, where it is very clear that the tooth growth means of OJ and VC at 2.0 level of dosage are almost the same.


Average tooth length growth per dose and supplement type

Hypotesis testing:

Based on the results seen above, let's state the following null hypothesis: the increase of supplement type OJ from 0.5 to 1.0 has a higher impact on tooth length. On the opposite, the alternative hypothesis will state that an increase from 0.5 to 1.0 does not produce a better impact. We use a one-sided t-test to examine whether we can reject the null hypothesis or not. In order to proceed to the calculation of the confidence intervals, we need to calculate the mean and the variance of the samples. We assume non equal variances between samples.

```
## The tooth length mean for 0.5 OJ dose is: 13.23
```

```
## The tooth length mean for 1.0 OJ dose is: 22.7
```

```
## The tooth length variance for 0.5 OJ dose is: 19.89
```

```
## The tooth length variance for 1.0 OJ dose is: 15.3
```

```
## With a 95% confidence, the interval is: 5.524 13.416
```

As the indicated interval does not include 0, we fail to reject the null hypothesis, hence we confirm the evidence that a greater dosage of OJ generates a higher impact on tooth growth.

In the same way, based on the results illustrated in the previous graphs, we can also formalize another null hypothesis: the impact of 2.0 dose of OJ on tooth growth is the same as the one generated by a 2.0 dose of VC. Hence the alternative hypothesis will state that the impact of OJ and VC at a 2.0 dose level is not the same, but one of the two is greater than the other. As we did in the previous hypothesis testing, we use a one-sided t-test and assume different variance between the two samples.

```
## The tooth length mean for 2.0 OJ dose is: 26.06
```

```
## The tooth length mean for 2.0 VC dose is: 26.14
```

```
## The tooth length variance for 2.0 OJ dose is: 7.05
```

```
## The tooth length variance for 2.0 VC dose is: 23.02
```

```
## With a 95% confidence, the interval is: -3.798 3.638
```

As the indicated interval does include 0, we fail to reject the null hypothesis, and hence conclude that the impact generated by a 2.0 dose level of OJ and VC is similar.

Appendix:

R code:

```r
library(datasets)
str(ToothGrowth)
summary(ToothGrowth)
cat(c("Standard deviation of tooth length:", round(sd(ToothGrowth$len), 2)))
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

library(ggplot2)
ggplot(data=ToothGrowth, aes(x=dose, y=len, colour=supp)) + geom_jitter() +
    xlab("Dose in milligrams/day") +
    ylab("Tooth length") +
    ggtitle("Tooth length per dose") +
    theme(plot.title = element_text(colour="DarkBlue", size=15))

means <- aggregate(data=ToothGrowth, len ~ dose + supp, mean)
colnames(means)[3] <- "mean"
```

```r
variances <- aggregate(data=ToothGrowth, len ~ dose + supp, var)
colnames(variances)[3] <- "variance"
variances$variance <- round(variances$variance, 2)

tableMeansVar <- cbind(means,variances)
tableMeansVar <- tableMeansVar[, -c(4, 5)]

ggplot(data=tableMeansVar, aes(x=dose, y=mean, colour=supp)) +
    geom_point(size = 3, alpha=0.5) +
    xlab("Dose in milligrams/day") +
    ylab("Tooth Lenght mean") +
    ggtitle("Average tooth length growth per dose and supplement type") +
    theme(plot.title = element_text(colour="DarkGreen", size=15))


oj05 <- ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 0.5, "len"]
oj10 <- ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 1.0, "len"]

cat(c("The tooth length mean for 0.5 OJ dose is:", round(mean(oj05), 2)))
cat(c("The tooth length mean for 1.0 OJ dose is:", round(mean(oj10), 2)))
cat(c("The tooth length variance for 0.5 OJ dose is:", round(var(oj05), 2)))
cat(c("The tooth length variance for 1.0 OJ dose is:", round(var(oj10), 2)))

testOj0510 <- t.test(oj10, oj05, lower.tail=TRUE, paired=FALSE, var.equal=FALSE)$conf
cat(c("With a 95% confidence, the interval is:", round(testOj0510, 3)))

oj20 <- ToothGrowth[ToothGrowth$supp == "OJ" & ToothGrowth$dose == 2.0, "len"]
vc20 <- ToothGrowth[ToothGrowth$supp == "VC" & ToothGrowth$dose == 2.0, "len"]

cat(c("The tooth length mean for 2.0 OJ dose is:", round(mean(oj20), 2)))
cat(c("The tooth length mean for 2.0 VC dose is:", round(mean(vc20), 2)))
cat(c("The tooth length variance for 2.0 OJ dose is:", round(var(oj20), 2)))
cat(c("The tooth length variance for 2.0 VC dose is:", round(var(vc20), 2)))

testOjVc20 <- t.test(oj20, vc20, lower.tail=TRUE, paired=FALSE, var.equal=FALSE)$conf
cat(c("With a 95% confidence, the interval is:", round(testOjVc20, 3)))
```