Tooth Growth Analysis

Normand Desmarais

January 27, 2016

In this project, we are going to analyze the `ToothGrowth` dataset, which is the result of a study on the effect of vitamin C on tooth growth in guinea pigs. Researchers were interested in three variables:

- `dose`: the dose of vitamin C given to the guinea pig
- `supp`: the supplement type - OJ for orange juice, VC for ascorbic acid
- `len`: and of course how it affect the tooth length

1. Exploratory Analysis

Let's first have a look at the data with `head` and `tail`:

```
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
```

```
tail(ToothGrowth)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

We can see that the data seem to be ordered in terms of supplement type and dose. Let's have a better look at the structure of the dataset with the command `str`:

```
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 ...
```

We now know that there is 60 observations, `len` and `dose` are numbers, while `supp` is a factor. To have an idea of how the 60 observations are distributed in terms of `supp` and `dose`, let's print a `table` of these two variables:

```
table(ToothGrowth$dose, ToothGrowth$supp)
```

```
##
##       OJ VC
##   0.5 10 10
##   1   10 10
##   2   10 10
```

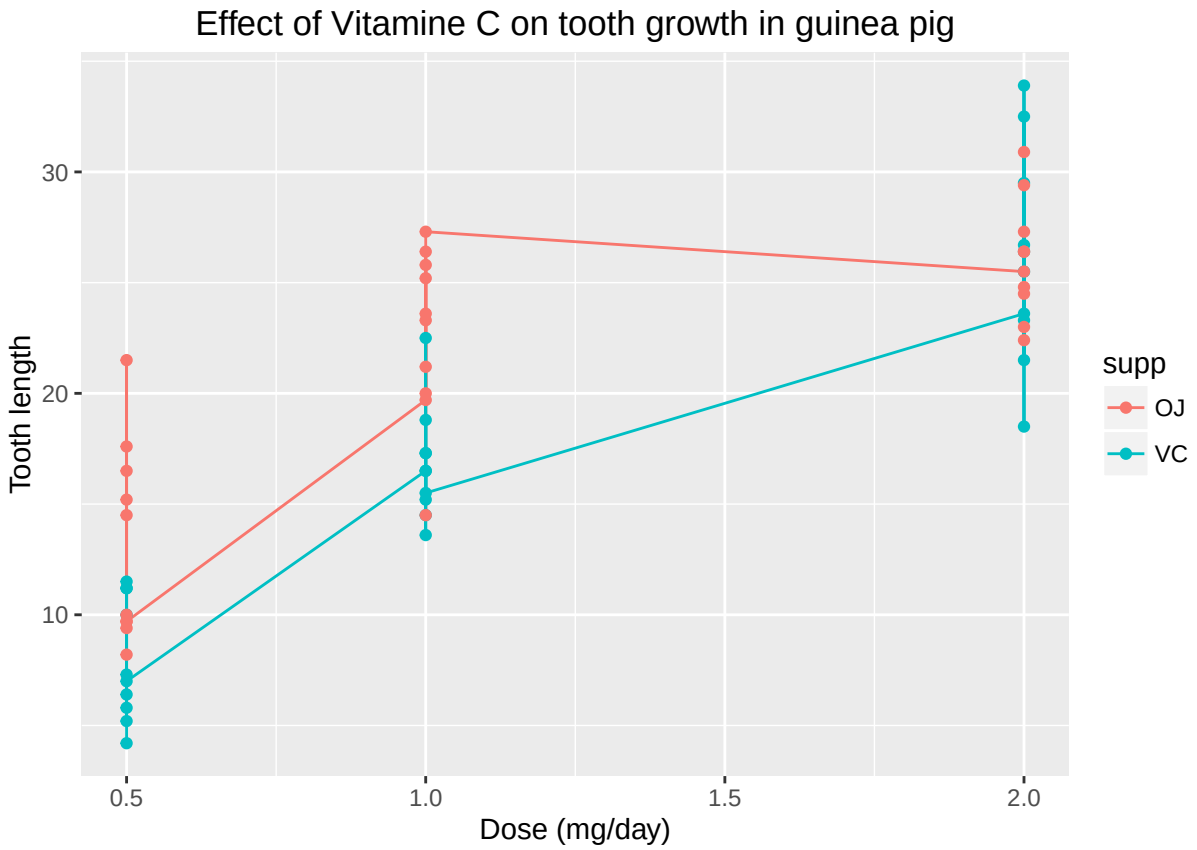
Nice! We now see that there is only three dose values: $0.5, 1$ and 2 mg/day. Furthermore, each supplement type have 30 observations divided in three groups of 10 observations for each dose. Let's now print a `summary` to see what else we can learn:

```
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
```

As seen before, there is 30 observations per `supp`. We can also see that the `len` are widely distributed with the mean being slightly smaller than the median. Finally, let's print a simple lines & dots ggplot to have a better picture of the dataset.

```
g <- ggplot(ToothGrowth, aes(x = dose, y = len, colour = supp))
g <- g + geom_line() + geom_point()
g <- g + xlab("Dose (mg/day)")
g <- g + ylab("Tooth length")
g <- g + ggtitle("Effect of Vitamine C on tooth growth in guinea pig")
g
```



There seems to be an indication that OJ is a better supplement than VC (at least for the 0.5 and 1.0 mg/day doses). It is also pretty clear that tooth length increases with the dose. Although, in the case of $VC/2.0$ mg/day, tooth length are widely spread, henceforth it is not as clear whether or not in this case OJ performs better. Let's formulate our observations into hypothesis and test them statistically.

2. Hypothesis testing

We have the following hypothesis:

- H_a: OJ is a better supplement than VC for tooth growth in guinea pigs: $\left(\overline{len}_{OJ} - \overline{len}_{VC}\right) > 0$.

We want to test it against the null hypothesis:

- H_0: OJ and VC supplements have the same effect: $\left(\overline{len}_{OJ} - \overline{len}_{VC}\right) = 0$.

We won't use confidence intervals since it would be the same as testing $\left(\overline{len}_{OJ} - \overline{len}_{VC}\right) \neq 0$, while our point is to show that OJ is better than VC, not that one or the other is better. Furthermore, we will test separately for each dose in order to answer the question:

- Are OJ Vitamin C supplement better than VC in each cases?

We will perform the calculations and present a summary of the results in a conclusion along with assumptions made to reach this conclusion.

2.1. Dose $= 0.5$ mg/day

Let's separate the dataset in two groups (based on the supplement value) and then filter the observations to only keep the right dose ($= 0.5$). We can then feed this result to a `t.test`. We are assuming a t-distribution as well as unequal variances. We are using the "greater" alternative of the test since we want to test the case where $(\overline{len}_{OJ} - \overline{len}_{VC}) > 0$.

For the conclusion, we are interested in grabbing the lowest confidence value to compare it to the difference between the supplement means. We also need the p-value, to inform us on how confident we could be in rejecting (or failing to) the null hypothesis.

```
# make the two groups
g1 <- filter(ToothGrowth, supp == "VC" & dose == 0.5)$len
g2 <- filter(ToothGrowth, supp == "OJ" & dose == 0.5)$len

# test
t_test <- t.test(g2, g1, paired = FALSE, var.equal = FALSE, alternative = "greater")

# grab what we need for the conclusion
conf_05 <- round(t_test$conf.int[1], 3)
diff_mean_05 <- round(t_test$estimate[1] - t_test$estimate[2], 3)
pvalue_05 <- round(t_test$p.value, 3)

# print the test
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  g2 and g1
## t = 3.1697, df = 14.969, p-value = 0.003179
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.34604     Inf
## sample estimates:
## mean of x mean of y
##     13.23      7.98
```

The p-value is far lower than 5%, hence we can safely reject the null hypothesis.

2.2. Dose $= 1.0$ mg/day

Let's do the same thing for dose $= 1.0$ mg/day.

```
g1 <- filter(ToothGrowth, supp == "VC" & dose == 1)$len
g2 <- filter(ToothGrowth, supp == "OJ" & dose == 1)$len

t_test <- t.test(g2, g1, paired = FALSE, var.equal = FALSE, alternative = "greater")

conf_10 <- round(t_test$conf.int[1], 3)
diff_mean_10 <- round(t_test$estimate[1] - t_test$estimate[2], 3)
pvalue_10 <- round(t_test$p.value, 3)

t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  g2 and g1
## t = 4.0328, df = 15.358, p-value = 0.0005192
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.356158      Inf
## sample estimates:
## mean of x mean of y
##     22.70     16.77
```

Again, the p-value is far lower than 5%, hence we can safely reject the null hypothesis in that case too.

2.3. Dose = 2.0 mg/day

Let's now evaluate the test for dose = 2.0 mg/day.

```
g1 <- filter(ToothGrowth, supp == "VC" & dose == 2)$len
g2 <- filter(ToothGrowth, supp == "OJ" & dose == 2)$len

t_test <- t.test(g2, g1, paired = FALSE, var.equal = FALSE, alternative = "greater")

conf_20 <- round(t_test$conf.int[1], 3)
diff_mean_20 <- round(t_test$estimate[1] - t_test$estimate[2], 3)
pvalue_20 <- round(t_test$p.value, 3)

# also grab the standard deviation in that case (for later use)
sd_VC <- round(sd(g1), 3)
sd_OJ <- round(sd(g2), 3)

t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  g2 and g1
## t = -0.046136, df = 14.04, p-value = 0.5181
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.1335     Inf
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

The p-value is far larger than 5% (p-value $\sim$ 52%)! Hence we've failed to reject the null hypothesis in that case.

3.0. Conclusion

Before reviewing the result, a word about assumptions. In our analysis, we have made the following assumptions:

- Tooth lengths are independent and identically distributed variables.
- The tooth length distribution for a given supplement type and dose follows a t-distribution.
- For a given dose value, tooth length have unequal variance between the two supplement types.
- Of course, the two groups (OJ and VC) were not paired (but this is more a fact than an assumption).

The results of the analysis are summarized in the following table; where "95% **mean-difference**" represents the (absolute) value above which ($\overline{len}_{OJ} - \overline{len}_{VC}$) must be greater than in order to reject the null hypothesis.

dose	$\overline{len}_{OJ} - \overline{len}_{VC}$	95% mean-difference	p-value	H_0
0.5	5.25	2.346	0.003	Reject
1.0	5.93	3.356	0.001	Reject
2.0	-0.08	-3.133	0.518	Fail to reject

For the two small doses (0.5 and 1.0 mg/day) we've successfully rejected the null hypothesis with a very small p-value (less than 0.5% while we only needed less than 5% to pass the test). Hence, in both cases, we are confident to say that OJ Vitamin C supplements are far better than VC ones. But, when we've reached 2.0 mg/day, we've totally failed to reject the null hypothesis. The difference between the two means is very small (-0.08). Hence, inverting the comparison would lead to a similar result (not shown here). The p-value in that case is approximately 52%, showing us how strongly we have failed to reject the null hypothesis.

In this last case, we would be interested into comparing the standard deviation between the two sets. We thus find the OJ supplement standard deviation to be 2.655, while the VC one is 4.798. Hence, although their mean are comparable, there is far more variance in the case of the VC supplement compare to the OJ. In other words, if one uses the VC supplement with a dose of 2.0 mg/day, one would have to expect to have as many terrible results (in terms of tooth growth) as good ones (relatively to the OJ supplement).

The final conclusion is that for small doses per day, OJ is a better Vitamin C supplement than VC, but for a dose of 2.0 mg/day, there is on average no difference between the two. Yet we still recommend to stick with OJ since the latter has a narrower variance, leading to more "reliable" and consistent results.