Basic Inferential Data Analysis

Clément Brutti-Mairesse

27/04/2020

About the dataset

We make an analysis on an experiment made on 60 pigs, trying to find the effect of `Orange Juice` and `Ascorbic Acid` on tooth growth. The metric is the length of odontoblast, evaluated in milligrams/day.

Exploratory Data Analysis

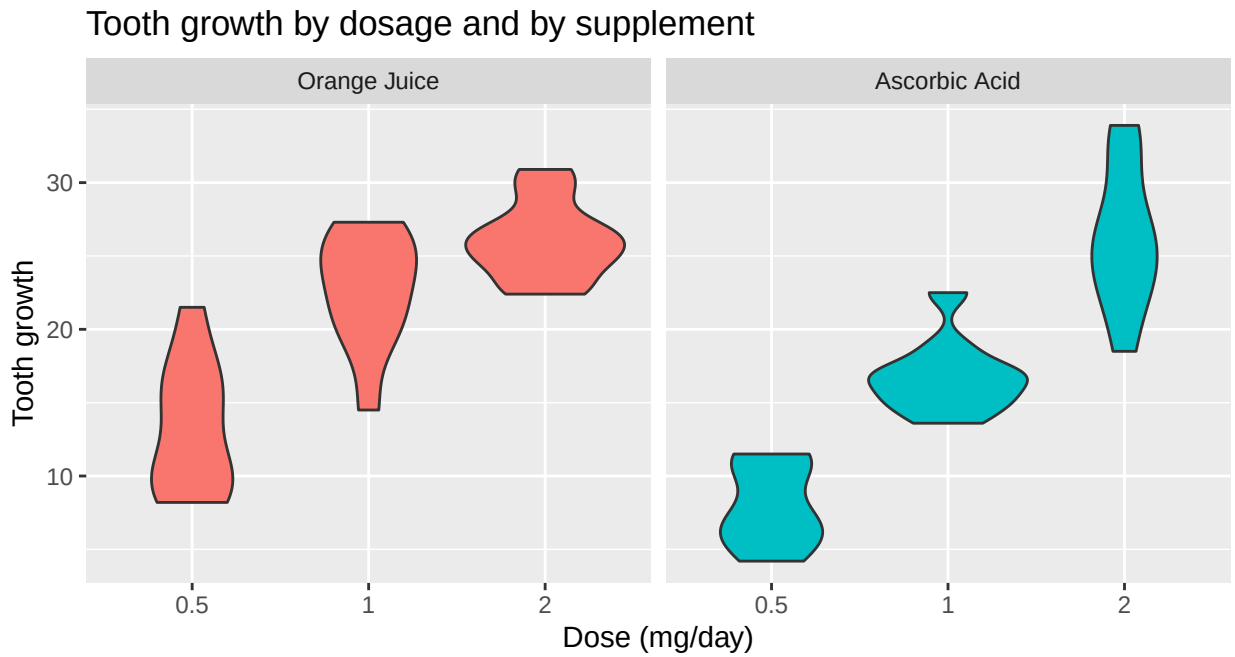
Here is a table with the count, variance, quartiles of each group.

```
data <- ToothGrowth
levels(data$supp) <- c("Orange Juice", "Ascorbic Acid")
groupData <- data %>% group_by(supp, dose)
kable(groupData %>% summarise(n = n(),  mean = mean(len), variance = var(len),
                              q25 = quantile(len)[2],
                              median = quantile(len)[3],
                              q75 = quantile(len)[4],
                              skewness = skewness(len),
                              kurtosis = kurtosis(len)))
```

supp	dose	n	mean	variance	q25	median	q75	skewness	kurtosis
Orange Juice	0.5	10	13.23	19.889000	9.700	12.25	16.175	0.4381154	-1.3697072
Orange Juice	1.0	10	22.70	15.295556	20.300	23.45	25.650	-0.6804992	-0.6799774
Orange Juice	2.0	10	26.06	7.049333	24.575	25.95	27.075	0.3685108	-1.0857354
Ascorbic Acid	0.5	10	7.98	7.544000	5.950	7.15	10.900	0.1330745	-1.8068950
Ascorbic Acid	1.0	10	16.77	6.326778	15.275	16.50	17.300	0.9255236	0.0762364
Ascorbic Acid	2.0	10	26.14	23.018222	23.375	25.95	28.800	0.1605264	-1.2320527

We can already see that groups do not have the same variance.

```
ggplot(data, aes(x=factor(dose), y=len)) +
 facet_grid(.~supp) +
 geom_violin(aes(fill = supp), show.legend = FALSE) +
 labs(title="Tooth growth by dosage and by supplement",
      x="Dose (mg/day)",
      y="Tooth growth")
```



Growth analysis

We now compare the six different groups with themselves. We suppose this is a randomized experiment, and that the groups are independents and that the variance is **not** constant between groups. We use a 95% T confidence interval to compare the groups. Here is a reminder of the detail of the t confidence interval calculation with an unequal variance, (we use `t.test(, var.equal=FALSE)`)

$$\bar{Y} - \bar{X} \pm t_{df} \times (\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y})^{1/2}$$

with t_{df} : t quantile and df equals to

$$df = \frac{(S_x^2/n_x + S_y^2/n_y)^2}{(S_x^2/n_x)^2/(n_x - 1) + (S_y^2/n_y)^2/(n_y - 1)}$$

Here is a table summing up the confidence intervals. Rows are compared with columns. For example: *Orange Juice 2mg/day* (OJ 2) is `15.8 to 20.36` mg/day better than *Ascorbic Acid 0.5mg/day* (AA 0.5).

	OJ 0.5	OJ 1	OJ 2	AA 0.5	AA 1	AA 2
OJ 0.5	NA	[-14.62;-4.325]	[-16.7;-8.956]	[1.263;9.237]	[-6.706;-0.374]	[-18.26;-7.562]
OJ 1	[4.325;14.62]	NA	[-7.271;0.5509]	[12;17.44]	[1.952;9.908]	[-6.848;-0.03218]
OJ 2	[8.956;16.7]	[-0.5509;7.271]	NA	[15.8;20.36]	[6.833;11.75]	[-4.329;4.169]
AA 0.5	[-9.237;-1.263]	[-17.44;-12]	[-20.36;-15.8]	NA	[-12.03;-5.55]	[-22.36;-13.96]
AA 1	[0.374;6.706]	[-9.908;-1.952]	[-11.75;-6.833]	[5.55;12.03]	NA	[-13.33;-5.405]
AA 2	[7.562;18.26]	[0.03218;6.848]	[-4.169;4.329]	[13.96;22.36]	[5.405;13.33]	NA

Similarly, this a table sum up the p-Values in percentage

	OJ 0.5	OJ 1	OJ 2	AA 0.5	AA 1	AA 2
OJ 0.5	NA	0.24%	0.0037%	1.5%	3.2%	0.04%
OJ 1	0.24%	NA	8.4%	6.6e-05%	0.82%	4.8%
OJ 2	0.0037%	8.4%	NA	2.4e-06%	0.0013%	97%
AA 0.5	1.5%	6.6e-05%	2.4e-06%	NA	0.017%	0.00043%
AA 1	3.2%	0.82%	0.0013%	0.017%	NA	0.046%
AA 2	0.04%	4.8%	97%	0.00043%	0.046%	NA

Conclusions

We can see that Ascorbic Acid had relevant improvement as the dose increase. In opposition as the dose of Orange Juice increase the growth does not increase significantly. Comparing the two supplements with equal dose does not show a relevant difference of growth either. For example the confidence interval comparing *Orange Juice 2mg/day* (OJ 2) with *Ascorbic Acid 0.5mg/day* (AA 2) is *-4.329:4.169* it contains zero, therefore the difference is not significant with a 95% confidence interval.