Basic inferential data analysis

Josh McNamara

7/9/2019

From the Rstudio guide:

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).

I will do some exploratory analysis on this.

Load the data and set up

```
library(ggplot2)
data("ToothGrowth")
df <- ToothGrowth
df$dose <- as.factor(df$dose)
```

Data summary

Have a look at the data, see what the variables are, see how they relate.

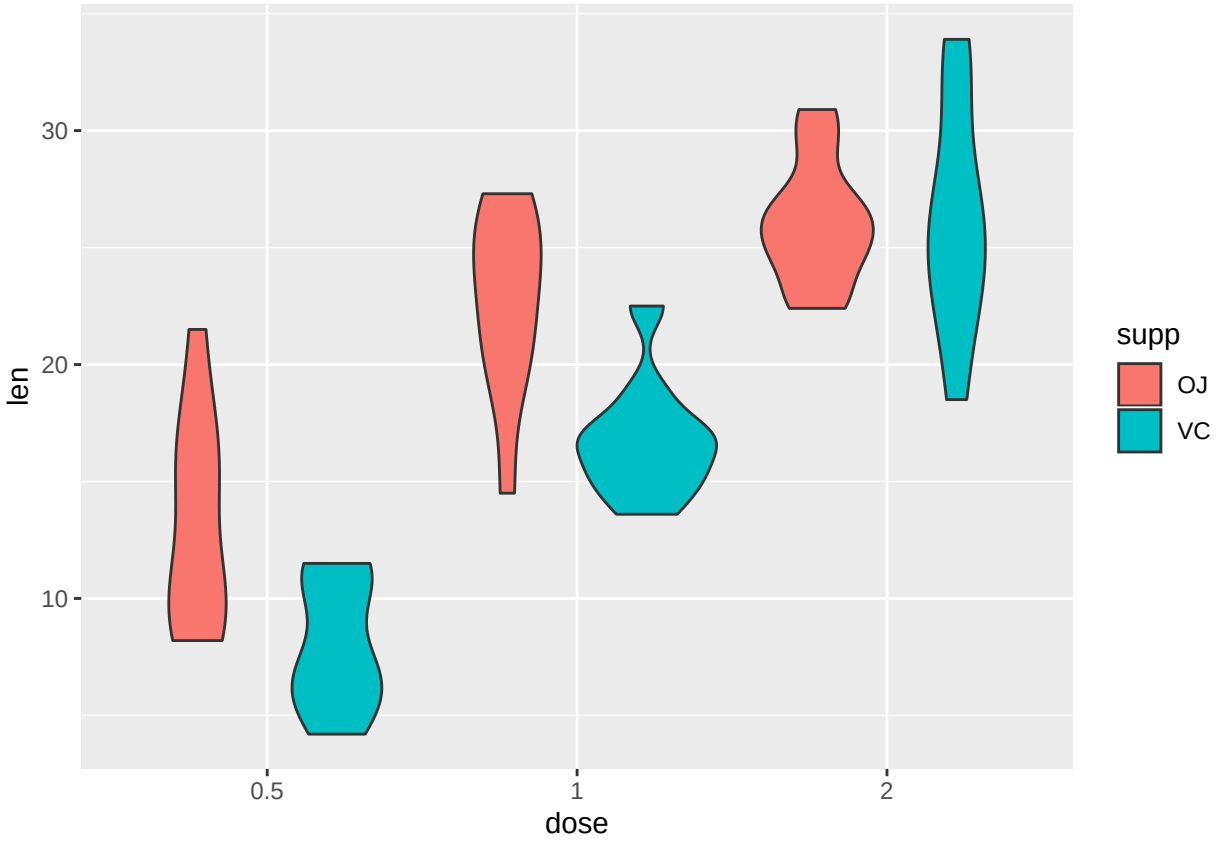
```
# Summary of the data
summary(df)
```

```
##       len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

```
str(df)
```

```
## '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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
ggplot(df, aes(x = dose, y=len, fill=supp)) + geom_violin()
```



There are two major conditions (Vitamin C and Orange juice dosing) taken thorough dose escalation for several animals.

It seems that OJ might be a more potent dosing method for a 1 mg dose. Let's test this.

Does the delivery method impact the response at a 1 mg dose?

I will perform a two-sample t-test to gauge whether these two dosing measures may pertain to different distributions. I hypothesize that the orange juice has a significantly greater effect on the length variable than vitamin C alone.

```
# My hypothesis is that for a dose of 1 mg, orange juice causes more growth than vitamin C alone

oj <- df$len[df$dose==1 & df$supp=='OJ']
vc <- df$len[df$dose==1][df$supp=='VC']
t.test(x = oj, y = vc, conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  oj and vc
## t = 1.8736, df = 20.234, p-value = 0.0755
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3336249  6.2636249
```

```
## sample estimates:
## mean of x mean of y
##    22.700    19.735
```

Conclusion

The t sample statistic is within the 95% confidence interval. Thus, we fail to reject the null hypothesis that the treatments outcomes pertain to the same distribution. This test assumed that the data are normally distributed. If they are not, the t-test is not a valid test.