Inferential Data Analysis

Andrew Witherspoon

11/11/2018

```
library(datasets)
#ToothGrowth is now available as a dataframe
```

For this exercise, we will use the **ToothGrowth** data, which is in the R datasets package.The R documentation gives a description of the data:

"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)."

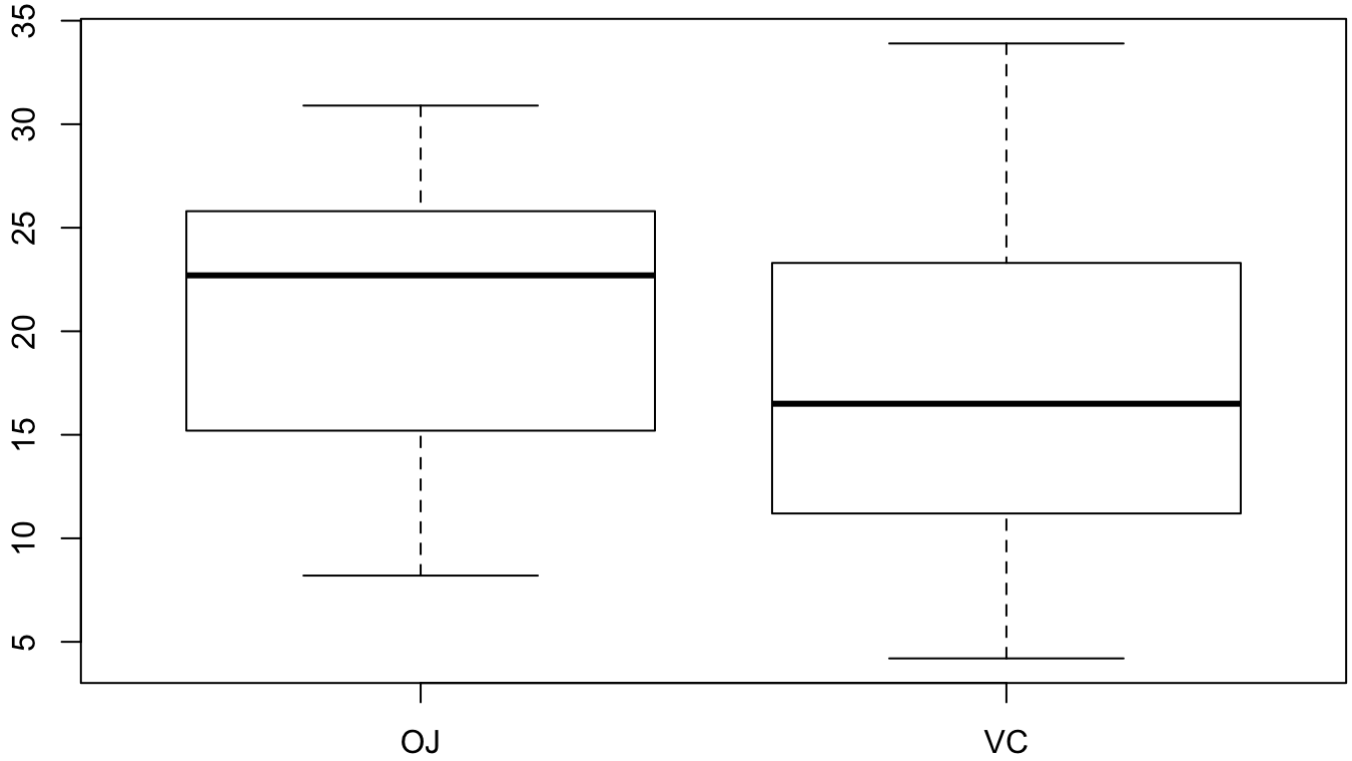
```
library(knitr)
kable(table(ToothGrowth[,2:3]), align = "l")
```

	0.5	1	2
OJ	10	10	10
VC	10	10	10

As the table above shows, there are 10 observations for each combination of dose, and supplement type.

Let's start by comparing the mean tooth length of guinea pigs treated with orange juice to the tooth length of those treated with ascorbic acid:

```
par(mar=c(2, 2, .75, .1))
boxplot(len ~ supp, data = ToothGrowth)
```



Our null hypothesis is that the mean tooth length of the OJ group not different than the mean tooth length of the VC group. We'll run a t-test to see if we can reject this hypothesis:

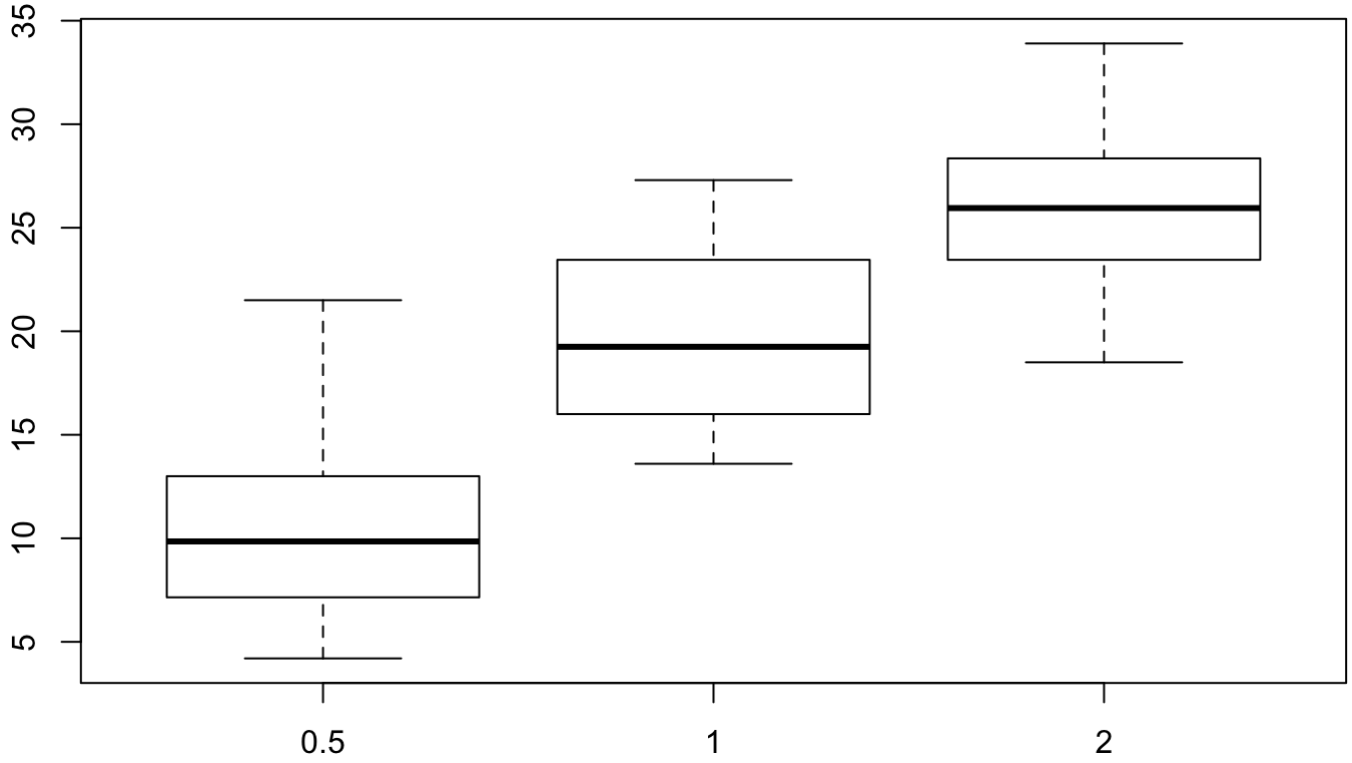
```
OJ <-  ToothGrowth$len[ToothGrowth$supp=="OJ"]
VC <-  ToothGrowth$len[ToothGrowth$supp=="VC"]
t.test(OJ, VC, alternative = "two.sided", conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  OJ and VC
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

The 95% confidence interval contains the value 0, therefore we cannot reject the null hypothesis. The data does not show that tooth length is affected by supplement type.

Let's do the same type of hypothesis testing based on dose:

```
par(mar=c(2, 2, .75, .1))
boxplot(len ~ dose, data = ToothGrowth)
```




```
dose0.5 <-  ToothGrowth$len[ToothGrowth$dose==0.5]
dose1 <-  ToothGrowth$len[ToothGrowth$dose==1]
dose2 <-  ToothGrowth$len[ToothGrowth$dose==2]
```

A dose of 0.5 and a dose of 2 have the largest mean difference, so let's use these values for our hypothesis testing. Once again, our null hypothesis will be that the mean of dose0.5 and the mean of dose2 are not different.

```
t.test(dose0.5, dose2, alternative = "two.sided", conf.level = .95)
```

```
##
##  Welch Two Sample t-test
##
## data:  dose0.5 and dose2
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean of x mean of y
##    10.605    26.100
```

We can reject the null hypothesis, as the confidence interval does not contain 0. With 95% confidence, the data shows that the dose does have an effect on tooth length.

These two sample t-tests make several assumptions about the populations they are sampled from. One is that the population data (tooth length), would follow a normal probability distribution. Another is that the variances of the two populations (e.g., the dose0.5 and the dose2 populations) are not equal. We also assume that the two populations are independent (not paired observations, or otherwise related), and that each guinea pig selected for the sample is randomly selected for its sample with equal probability. There are no obvious violations of these assumptions.