# Hypothesis tests, t-test review

KEY

Wed Mar 10 13:45:24 2021

## Importing the data

Note that this comes after the summary command in the online instructions, but we will put it here for the sake of repeatability of the analysis:

```
library(readxl)

onesample <- read_excel("review_data.xls","onesample")
```

## Summarizing a data set

The onesample data set has just a single variable, called Body.temperature. Using the summary() command on the onesample data set gives you a numerical summary of Body.temperature.

```
summary(onesample)

##   Body.temperature
##   Min.   :97.30
##   1st Qu.:98.20
##   Median :98.40
##   Mean   :98.36
##   3rd Qu.:98.70
##   Max.   :98.90
```

**Question: which of the reported summary statistics are measures of central tendency? (write your answer below the questions)**

Median and mean.

## One sample t-tests

One-sample t-tests test hypotheses about the population mean for a single sample of data. We have data on body temperature, and we want to know if it is consistent with the typical value used for normal human body temperature (98.6 degrees F), or not.

**Question: what is the null hypothesis for a test of our data having a mean body temperature of 98.6 degrees F? Use symbolic notation (i.e. Ho: ...)**

Ho: mu = 98.6 (not x-bar, which is the sample mean)

The one-sample t-test assumes normality. Test normality here:

```
shapiro.test(onesample$Body.temperature)

##
##  Shapiro-Wilk normality test
##
## data:  onesample$Body.temperature
## W = 0.93494, p-value = 0.1131
```

**Question: do you pass the normality test? How do you know?**

Yes, p is greate than 0.05.

Whenever you are analyzing data, it's important to look at its distributed so that you can see whether the data is clearly not normally distributed, spot data entry errors, and look for "outliers", which are data points that could have a large effect on the outcome of your analysis. First do a strip chart:

```
stripchart(onesample$Body.temperature, xlab = "Temperature")
```
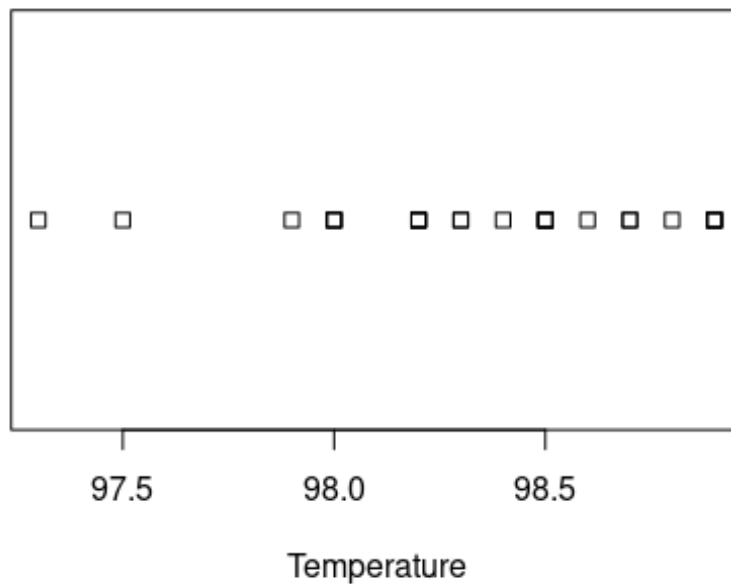


Then a boxplot:

```
boxplot(onesample$Body.temperature, ylab = "Temperature")
```
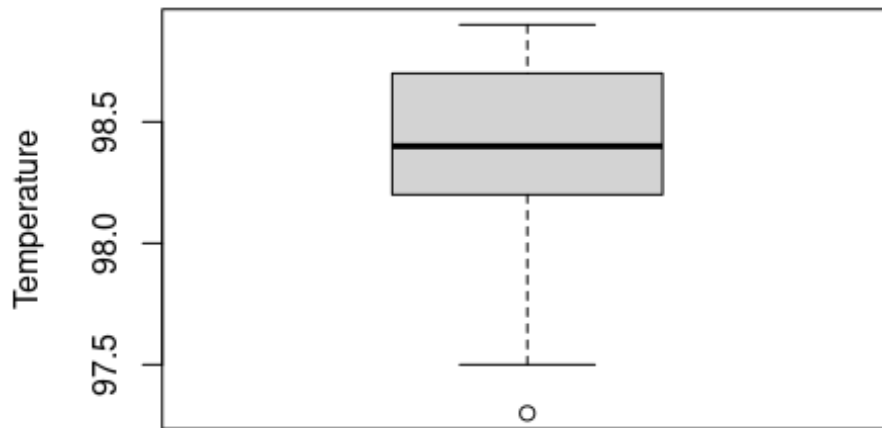
**Question: are there any extreme values in your plots? How do you know?**

Yes, the smallest value is, because there is a circle below the lower whisker.

Now conduct the t-test:

```
t.test(onesample$Body.temperature, mu = 98.6)

##
##   One Sample t-test
##
## data:  onesample$Body.temperature
## t = -2.8817, df = 24, p-value = 0.008205
## alternative hypothesis: true mean is not equal to 98.6
## 95 percent confidence interval:
##   98.18125 98.53075
## sample estimates:
## mean of x
##     98.356
```

**Question: how are the degrees of freedom calculated?**

For a one-sample t-test degrees of freedom are equal to sample size (n) minus 1, df = 25 - 1 = 24.

**Question: what does the t-value tell you?**

It is the number of standard errors between the sample mean and the hypothetical value of 98.6.

**Question: do you reject or retain the null hypothsis? How do you know?**

Yes, the null is rejected because p is less than 0.05.

**Question: what does your result tell you about mean body temperature? Do the results pertain to the sample mean, or the population mean?**

The result indicates that the mean of this population is different from 98.6 degrees. We know the mean of the sample is not 98.6 because we calculated its value as 98.356. The hypothesis test pertains to the mean of the population that this sample came from, and we are concluding the population does not have a mean of 98.6 degrees F.

**Question: is there a homogeneity of variances assumption for a one-sample t-test? Why or why not?**

No, because there is only one sample of data. HOV pertains to two or more variances, we have only one.

## Two-sample t-tests

We have data comparing heights of plants that were either self-pollinated, or pollinated with pollen from another individual. We wish to compare the mean heights between these two treatment groups.
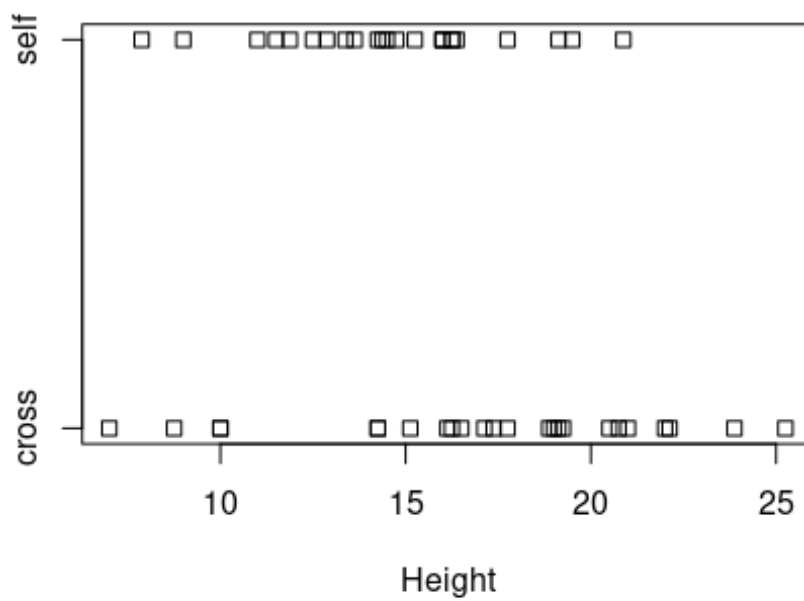
**Question: what is the null hypothesis for this test?**

Ho: mu1 = mu2

Import the data from worksheet "twosample" of file "review_data.xls":

```
twosample <- read_excel("review_data.xls","twosample")
```
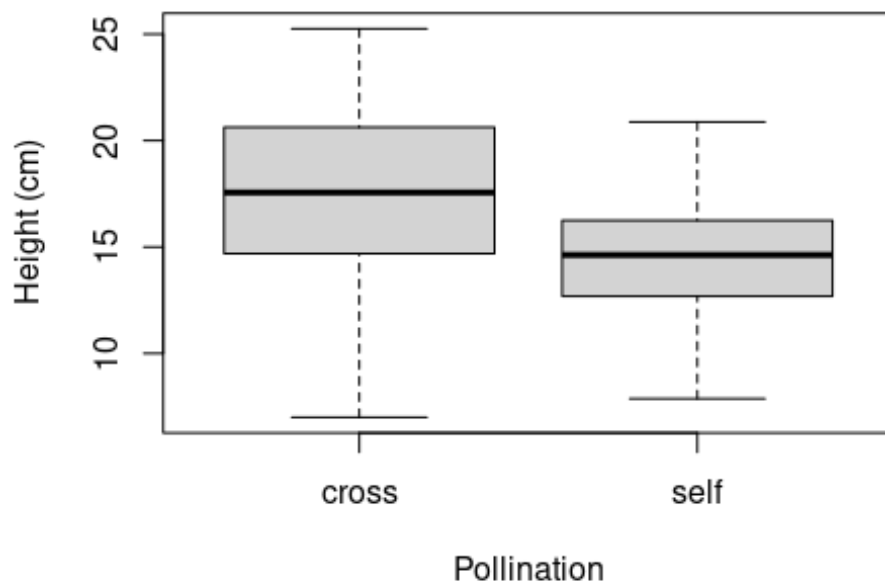
Before analysis look at your data with a stripchart:

```
stripchart(Height ~ Pollination, data = twosample)
```

and a boxplot:

```
boxplot(Height ~ Pollination, data = twosample, ylab = "Height (cm)")
```

**Question: based on the boxplots, are the data skewed (and if so, are they positively or negatively skewed)?**

No, the median is about in the middle of the box for both groups. If the median was not in the middle the data would be skewed in the direction of the wider half of the box.

Test for normality:

```
with(twosample, tapply(Height, Pollination, shapiro.test))

## $cross
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95884, p-value = 0.4155
##
##
## $self
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.98327, p-value = 0.9476
```

**Question: do we pass normality for both groups? Why did we have to split the data into the two groups to test for normality?**

Yes we do, because both groups have p > 0.05. We had to split them because if the means are in different locations the combination of both groups could be bimodal, rather than normal, when each group alone is normally distributed.

Test for HOV:

```
with(twosample, bartlett.test(Height, Pollination))

##
##  Bartlett test of homogeneity of variances
##
## data:  Height and Pollination
## Bartlett's K-squared = 3.6947, df = 1, p-value = 0.05458
```

**Question: do we pass the HOV test? How do you know?**

Yes we do, because p > 0.05 (just barely)

Now do the t-test:

```
t.test(Height ~ Pollination, var.equal = TRUE, data=twosample)
```

```
## 
##  Two Sample t-test
## 
## data:  Height by Pollination
## t = 2.2124, df = 46, p-value = 0.03195
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2305639 4.8840194
## sample estimates:
## mean in group cross  mean in group self
##            17.17708            14.61979
```

**Question: is the difference between the two group means statistically significant? How do you know?**

Yes, because it is less than 0.05.

**Question: based on these results, does outcrossing increase plant height? How can you tell (is the p-value enough to know, or do you need to look at additional information)?**

The p-value indicates that the difference between the means is not 0, but it doesn't indicate which group is bigger. The sample means show that the crossed plants are taller on average, which tells us the direction of the difference.

## Paired t-test

You have data on CO2 uptake by a single set of plants exposed to two different ambient CO2 concentrations. Since the same plants are used for each treatment you need to use a paired t-test to account for the lack of independence of measurements for the same plant at each treatment level.

Import the data from sheet "paired" in the file "review_data.xls":
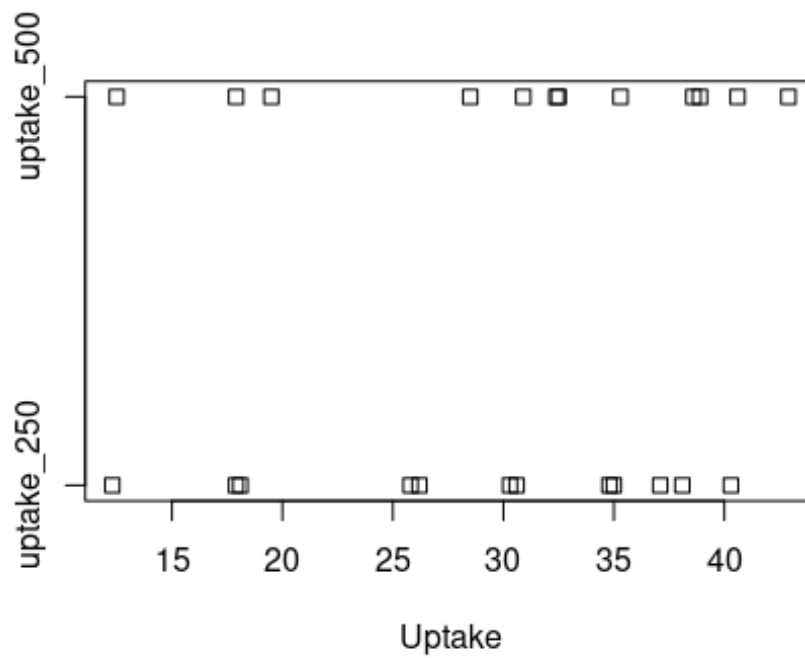
```
paired <- read_excel("review_data.xls","paired")
```

Stack the data and rename the columns so that you can plot the raw data values:

```
stack(paired, c(2,3)) -> co2.stacked

colnames(co2.stacked) <- c("Uptake", "Treatment")
```
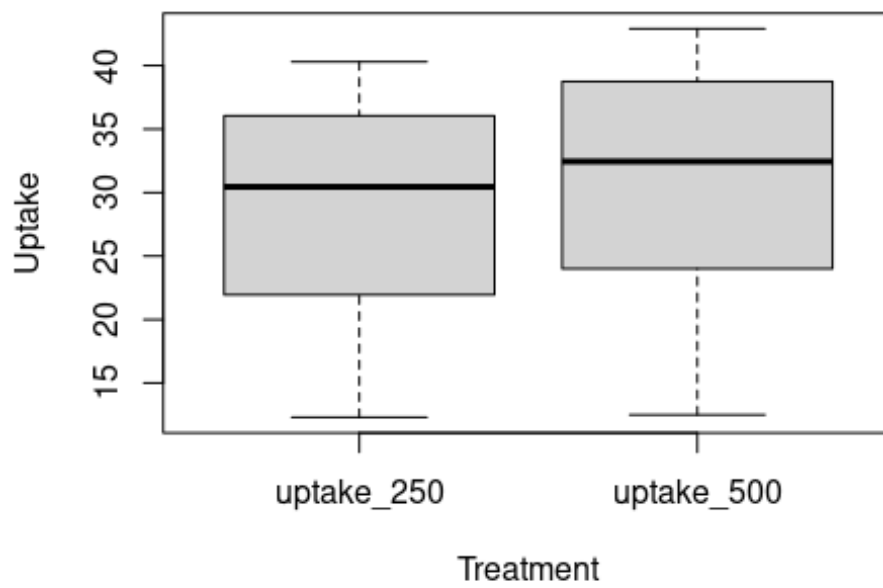
Look at the data using a strip chart:

```
stripchart(Uptake ~ Treatment, data = co2.stacked)
```

and using a boxplot:

```
boxplot(Uptake ~ Treatment, data = co2.stacked)
```

**Question: is there a clear separation in the data values between the groups, or is there a lot of overlap in uptake amounts between them?**
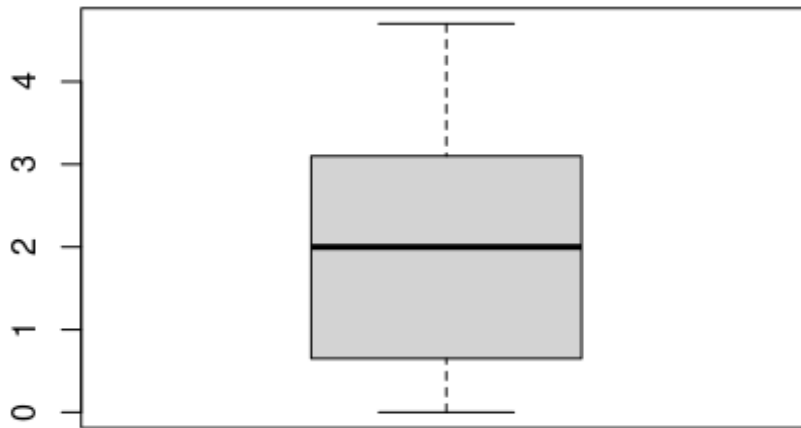
No, the groups are not clearly different, there is a great deal of overlap in the data values.

The raw data are good to look at, but for a paired t-test it is the differences between paired of values that are used. Calculate the differences:

```
with(paired, uptake_500 - uptake_250) -> paired$diffs
```

Make a boxplot of the differences:

```
boxplot(paired$diffs)
```



**Question: are the differences mostly positive or mostly negative? Or are they a roughly even mix of positive and negative?**

They are all positive.

Check normality of the differences:

```
shapiro.test(paired$diffs)

##
##  Shapiro-Wilk normality test
##
```

```
## data:  paired$diffs
## W = 0.96012, p-value = 0.7856
```

**Question: do you pass the normality test for differences between the pairs of measurements?**

Yes we pass normality, because p > 0.05.

Now do the paired t-test:

```
with(paired, t.test(uptake_250, uptake_500, paired = TRUE))

##
##  Paired t-test
##
## data:  uptake_250 and uptake_500
## t = -4.6595, df = 11, p-value = 0.0006943
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -2.944739 -1.055261
## sample estimates:
## mean of the differences
##                      -2
```

**Question: what do you conclude about the effect of ambient concentration on uptake of CO2? Is the difference statistically significant (how do you know)? And, how do you know if increased ambient CO2 increases uptake or decreases it?**

The p-value is less than 0.05, so the differences are significantly different from zero. We know that there is a greater uptake at 500 than 250 because the test subtracted uptakes at 500 from uptakes at 250 and got a mean difference of -2.

**Question: explain how it was possible for the two sets of data values that were so broadly overlapping to produce differences that were so clearly, consistently positive? How did using an analysis of the differences help detect the effect of ambient CO2 on uptake in this experiment?**

The difference between 250 and 500 was small compared with the large amount of variation in uptake among the different plants, but the difference was consistent. In other words, a small plant took up less CO2 than a large one, but all plants increased their CO2 uptake when the ambient CO2 was higher. This small, consistent increase is what was tested, and it was highly statistically significant.