Lab 2 – Summary statistics, graphics, and the t-test

FANR 6750

Richard Chandler and Bob Cooper
University of Georgia

Recap

Last week we covered:

- Vectors
- Data frames
- Indexing
- Importing and exporting data
- Saving and loading workspaces

Today's Topics

1. Introduction
2. Graphics
3. t tests
   - Two-sample t test
   - Equality of variance test
   - Paired t-test

Types of t-tests

One sample

- Does the mean (μ) differ from some value of interest?
- One or two-tailed

Two sample

- Do the two means (μ_1 and μ_2) differ from one another?
- One or two-tailed
- Paired or unpaired

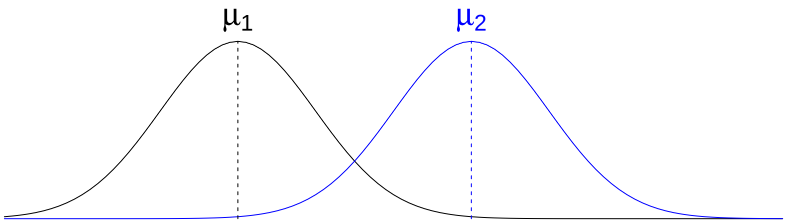
Two-sample, unpaired, two-tailed scenario

We have 2 samples of data and we want to know if they came from the same population.

The problem is that the true population means (μ_1, μ_1) are unknown.

Under the assumption that the variances of the two populations are equal, the relevant hypotheses are:

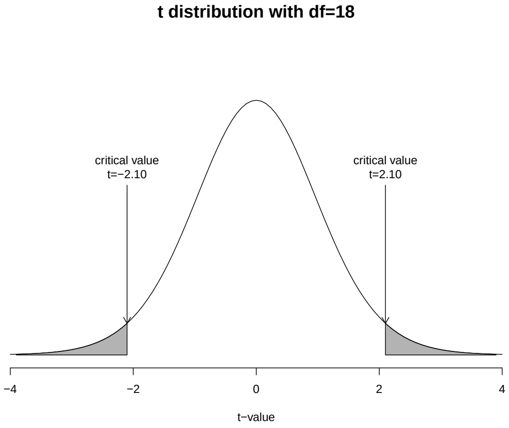
- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu_1 \neq \mu_2$



Key points

If the two sample means $(\bar{y_1}, \bar{y_2})$ are very different and the standard error of the difference in means is small, the t statistic will be far from zero.

If the t statistic is more extreme than the critical values, you reject the null hypothesis (H_0).



Exercise I

(1) Open **R** and set the working directory to a convenient location on your computer. Do this using the drop down menu or the `setwd` function.

(2) Put the file `treedata.csv` into your working directory.

(3) Create a new **R** script and import `treedata.csv`. Name your object `treedata`.

(4) Use the indexing methods we covered last time to create 2 objects: `yL` is the tree density data for the first 10 experimental units (low elevation), and `yH` is the tree density data for the last 10 units (high elevation).

(5) Compute the mean, variance, and standard deviation of the 2 samples.

Today's Topics

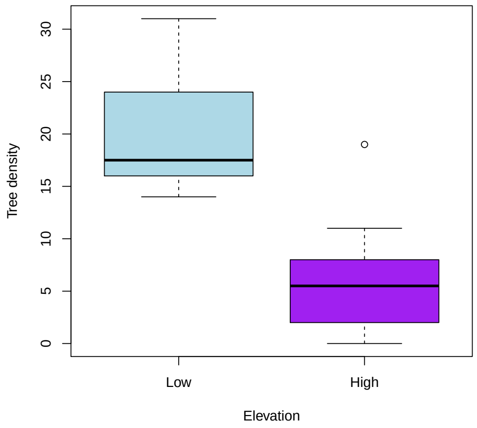
1 Introduction

2 Graphics

3 *t* tests

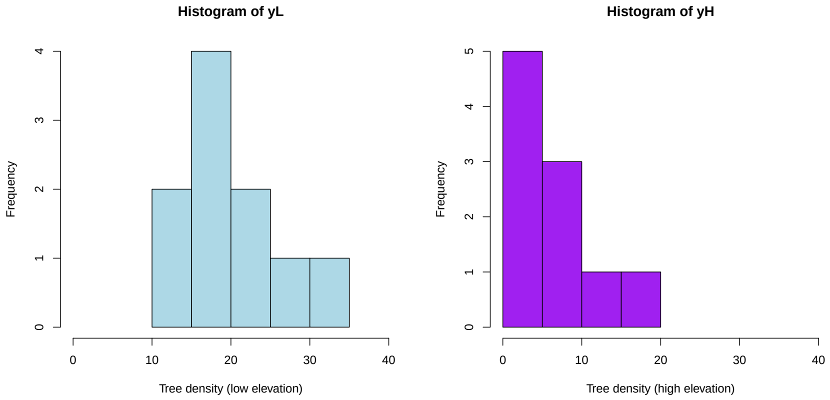
Boxplots

```
boxplot(yL, yH, xlab="Elevation", names=c("Low", "High"),
        ylab="Tree density", col=c("lightblue", "purple"))
```



Histograms

```
hist(yL, xlab="Tree density (low elevation)", col="lightblue",
     xlim=c(0, 40))
hist(yH, xlab="Tree density (high elevation)", col="purple",
     xlim=c(0, 40))
```



Exercise II

Create the same boxplots and histograms as before, but change the colors of the boxplots and the number of break points in the histograms.

Today's Topics

1. Introduction
2. Graphics
3. t tests

Two-sample t-test with equal variances

Step 1: Compute the t statistic[1]:

$$t = \frac{(\bar{y_L} - \bar{y_H}) - (\mu_L - \mu_H)}{\sqrt{s_p^2/n_L + s_p^2/n_H}}$$

where s_p^2 is the pooled variance:

$$s_p^2 = \frac{(n_L - 1)s_L^2 + (n_H - 1)s_H^2}{n_L + n_H - 2}$$

Step 2: Compare t statistic to critical values

Critical value for 1-tailed test $t_{\alpha=0.05,18} = -1.73 \text{ or } 1.73$

Critical values for 2-tailed test $t_{\alpha=0.05,18} = -2.10 \text{ and } 2.10$

[1] Remember, H_0 states that $\mu_L - \mu_H = 0$.

Do it by hand in R

Step 1: Compute the t statistic:

```r
mean.L <- mean(yL)
mean.H <- mean(yH)
s2.L <- var(yL)
s2.H <- var(yH)
n.L <- length(yL) # length returns the number of elements in a vector
n.H <- length(yH)
s2.p <- ((n.L-1)*s2.L + (n.H-1)*s2.H)/(n.L+n.H-2)
SE <- sqrt(s2.p/n.L + s2.p/n.H)
t.stat <- (mean.L - mean.H) / SE
t.stat

## [1] 5.404896
```

Step 2: Compare t statistic to critical values (two-tailed)

```r
alpha <- 0.05
## NOTE: qt returns critical values. No need to use "t tables"
critical.vals <- qt(c(alpha/2, 1-alpha/2), df=n.L+n.H-2)
critical.vals

## [1] -2.100922  2.100922
```

Conclusion: Reject H_0 because 5.4 is more extreme than the critical values.

Let R do all the work – Option 1

Provide the data as two vectors, one for each sample.

```r
t.test(yH, yL, var.equal=TRUE,
       paired=FALSE, alternative="two.sided")

##
##  Two Sample t-test
##
## data:  yH and yL
## t = -5.4049, df = 18, p-value = 3.898e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.580772  -8.619228
## sample estimates:
## mean of x mean of y
##       6.1      20.2
```

Make sure you set `var.equal=TRUE`. Otherwise, **R** will assume that the variances of the two populations are unequal.

Let R do all the work – Option 2

Provide the data as a data.frame, with one column for treeDensity and one column for Elevation.

```r
t.test(treeDensity ~ Elevation, data=treedata, var.equal=TRUE,
       paired=FALSE, alternative="two.sided")

##
##  Two Sample t-test
##
## data:  treeDensity by Elevation
## t = -5.4049, df = 18, p-value = 3.898e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.580772  -8.619228
## sample estimates:
## mean in group High  mean in group Low
##                6.1               20.2
```

This second option returns identical results, but it is preferred because the notation is much more similar to the notation used to fit ANOVA models.

Test equality of variances using var.test

The standard 2 sample t-test assumes that the variances are equal. Here's how you can test this assumption:

```
var.test(yL, yH)

##
##  F test to compare two variances
##
## data:  yL and yH
## F = 1.1499, num df = 9, denom df = 9, p-value = 0.8386
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2856132 4.6293987
## sample estimates:
## ratio of variances
##           1.149877
```

Suppose the samples are paired

The Caterpillar Data from class

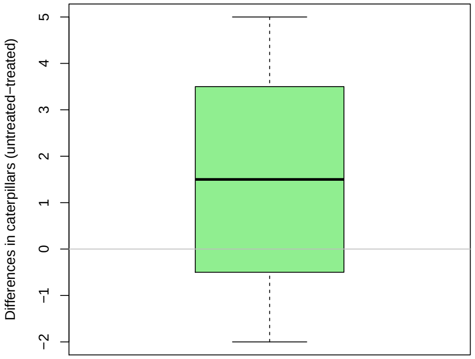
```
location <- 1:12
untreated <- c(23,18,29,22,33,20,17,25,27,30,25,27)
treated <- c(19,18,24,23,31,22,16,23,24,26,24,28)
```

For paired t-tests, we are interested in the mean of the differences. Focusing on the differences allows us to account for extraneous sources of variation among sample pairs.

```
diff <- untreated-treated
diff

##  [1]  4  0  5 -1  2 -2  1  2  3  4  1 -1

mean(diff) ## Estimate of the mean of the differences

## [1] 1.5
```

Is the mean of the differences > 0?

```
boxplot(diff, col="lightgreen", size='small',
        ylab="Differences in caterpillars (untreated-treated)")
abline(h=0, col="grey")
```



Paired t-test

Recall: Paired t-test is the same as a one-sample t-test on the differences. The hypothesis *in the Caterpillar example* is one-tailed:

- $H_0 : \mu_d \leq 0$
- $H_A : \mu_d > 0$

Step 1. Calculate the standard deviation of the differences.

$$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Step 2. Calculate the test statistic.

$$t = \frac{\bar{y} - 0}{s_d/\sqrt{n}}$$

Step 3. Compare to critical value.

Assignment

Create a script to do the following:

(1) Do a paired t test on the caterpillar data without using the `t.test` function. Use only the functions `mean`, `sd`, and possibly `length`.

(2) Do the paired t test again, but this time using the `t.test` function.

- You will need to use the `"paired"` argument

(3) Do a standard (*unpaired*) two-sample t test using the `t.test` function.

(4) Add a comment to the end of your script to interpret your results and explain why the results differ from the paired vs unpaired analyses. Also list the null and alternative hypotheses for each test.

Upload your script[2] to ELC before next week's lab.

- The script must be self-contained. In other words, you should be able to copy and paste the entire thing into the **R** console, and it should return the correct results.

Read pp. 127–131 in "Introductory Statistics with R"

[2] Or upload an Rmarkdown (.Rmd) file