Lab Assignment: Two-Sample t-Procedures

Maggie Schweihs

Sept. 25, 2017

## Exercise 1

An experiment was conducted to evaluate the effectiveness of a drug treatment for tapeworm in the stomachs of sheep. A random sample of 24 worm-infected lambs of approximately the same age was divided into two groups. Twelve of the lambs were injected with the drug and the remaining twelve were left untreated. After a 6-month period the lambs were slaughtered and the worm counts recorded.

### Part 1a

Load the dataset WormSheep from the DS705data package. Note that the package is loaded above at line 18. You just need the data() command. To see all the datasets in the package, type data(package='DS705data').

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

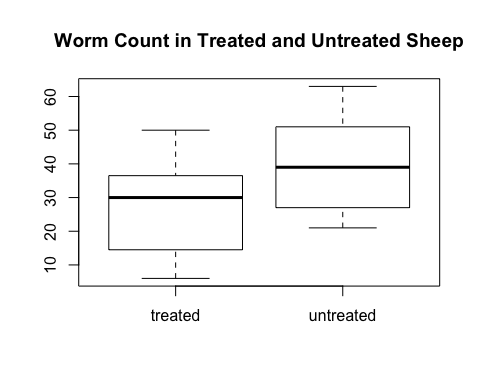
# Insert your R code here.  
data(WormSheep)

### Part 1b

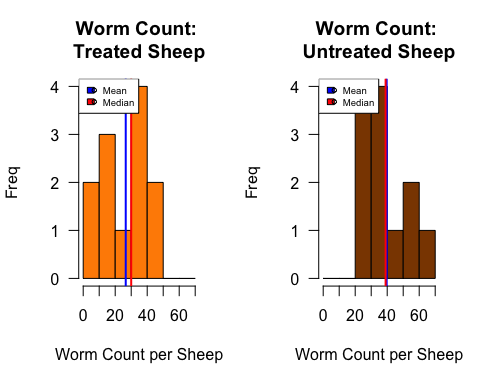
Create boxplots and histograms for each group (treated vs. untreated). Be sure that each plot is labeled appropriately.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

untreated <- WormSheep$worms[WormSheep$treated != "treated"]  
treatd <- WormSheep$worms[WormSheep$treated == "treated"]  
#par( mfrow = c( 3, 1 ) )  
boxplot(WormSheep$worms ~ WormSheep$treated, data = WormSheep,  
 main = "Worm Count in Treated and Untreated Sheep")



par( mfrow = c( 1, 2 ) )  
hist(treatd, breaks=c(0, 10, 20, 30, 40, 50, 60, 70), xlab = "Worm Count per Sheep",  
 ylab = "Freq",   
 main = "Worm Count: \nTreated Sheep", col = "darkorange", las=1)  
abline(v = mean(treatd), col = "blue", lwd = 2)  
abline(v = median(treatd), col = "red", lwd = 2)  
legend("topleft", c("Mean", "Median"), cex=0.6,  
 fill=c("blue", "red"), pch=21)  
hist(untreated, breaks=c(0, 10, 20, 30, 40, 50, 60, 70), xlab = "Worm Count per Sheep",   
 ylab = "Freq",   
 main = "Worm Count: \nUntreated Sheep",   
 col = "darkorange4", las=1)  
abline(v = mean(untreated), col = "blue", lwd = 2)  
abline(v = median(untreated), col = "red", lwd = 2)  
legend("topleft", c("Mean", "Median"),   
 cex=0.6, fill=c("blue", "red"), pch=21)



### Part 1c

Do the boxplots show any outliers?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

## The Boxplots do not show any outliers.

### Part 1d

Describe the shapes of the histograms for the sample data for each sample.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

## The histogram for "Treated Sheep" appears to have two observable humps indicating *possible* bimodal data. The histogram for "Untreated Sheep" appears to have a tail extending to the right indicating the data is *possibly* skewed to the right.

### Part 1e

Conduct an appropriate test to determine if the worm counts in each population can be considered as normally distributed. Provide the p-value and the conclusion of the test at a 5% level of significance.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

treatd\_stest <- shapiro.test(treatd)  
untreatd\_stest <- shapiro.test(untreated)  
treatd\_stest

##   
## Shapiro-Wilk normality test  
##   
## data: treatd  
## W = 0.95092, p-value = 0.6504

untreatd\_stest

##   
## Shapiro-Wilk normality test  
##   
## data: untreated  
## W = 0.94382, p-value = 0.5491

### Population 1: Treated Sheep

The sample was drawn from a normally distributed population.$

The sample was **not** drawn from a normally distributed. population.

Since 0.6504052, we do not reject the null hypothesis at the 5% significance level. There is not evidence to show that the sample was **not** drawn from a normally distributed population.

### Population 2: Untreated Sheep

The sample was drawn from a normally distributed population. The sample was **not** drawn from a normally distributed. population.

## Since 0.5491125, we do not reject the null hypothesis at the 5% significance level. There is not evidence to show that the sample was **not** drawn from a normally distributed population.

### Part 1f

Conduct an appropriate test to determine if the worm counts in each population can be considered to have equal variances. Provide the p-value and the conclusion of the test. Let .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1f -|-|-|-|-|-|-|-|-|-|-|-

ltest <- leveneTest(WormSheep$worms~WormSheep$treated)  
ltest\_pval <- ltest$`Pr(>F)`[1]  
ltest

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.1415 0.7104  
## 22

### Hypotheses

Population variances are equal.

Population variances are **not** equal.

### Conclusion

Since 0.7103962, we do not reject the null hypothesis at the 5% significance level. It is not unreasonabe to consider the underlying population variances to be equal.

### Part 1g

Conduct the test of your choice to determine if the population mean worm count for all sheep treated with the drug differs from the mean worm count for the population of untreated sheep. Let .

#### Step 1

Define the parameters in words in the context of the problem.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g.step1 -|-|-|-|-|-|-|-|-|-|-|-

Let be the population mean of sheep treated for worms.

Let be the population mean of sheep **not** treated for worms.

#### Step 2

State the null and alternative hypotheses for the test.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g.step2 -|-|-|-|-|-|-|-|-|-|-|-

#### Step 3

Use R to generate the output for the test you selected.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g.step3 -|-|-|-|-|-|-|-|-|-|-|-

ttest <- t.test(WormSheep$worms~WormSheep$treated, data = WormSheep, var.equal = TRUE, conf.level = 0.95)  
ttest

##   
## Two Sample t-test  
##   
## data: WormSheep$worms by WormSheep$treated  
## t = -2.2709, df = 22, p-value = 0.03329  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -25.031761 -1.134906  
## sample estimates:  
## mean in group treated mean in group untreated   
## 26.58333 39.66667

#### Step 4

State both a statistical conclusion at and interpret it in the context of the problem.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g.step4 -|-|-|-|-|-|-|-|-|-|-|-

Since 0.0332932, we reject the null hypothesis at the 5% significance level. There is sufficient evidence to claim that the population mean of the number of worms in sheep treated is different from the population mean for the number of worms in untreated sheep ( 0.0332932).

### Part 1h

Write an interpretation in the context of the problem for the 95% CI.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1h -|-|-|-|-|-|-|-|-|-|-|-

With 95% confidence, the population mean for worms found in sheep treated for worms is 1 to 25 worms less than in the group that was not treated for worms using the experimental drug.

### Part 1i

Did you use the separate-variance t-procedures or the pooled t-procedures? Justify your choice, including some discussion of how well the conditions for the hypothesis test and confidence interval procedures were met.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1i -|-|-|-|-|-|-|-|-|-|-|-

I used the pooled t-procedures since the following conditions were met:

* Independent and random samples from a population
* This information was given in the problem statement.
* Normally distributed population (verified by Shapiro-Wilk Test) +The results of each Shapiro-Wilk test indicated p-values much greater than ( 0.6504052 and 0.5491125) and also -values very close to 1 ($w\_ =$0.9509154 and $w\_ =$0.9438222). Due to these results, the null hypothesis was not rejected in either case and we accept that both samples come from a normally distributed population.
* Equal Variance (verified by the Levene Test) +The results of the Lavene Test indicated that at the significance level of 0.05, we could not reject the null hypothese and there was insufficient evidence to say that the samples came from populations with non-equal variance. So we maintain the assumption of equal variance which is necessary for conducting a pooled t-procedure.

## Exercise 2

Data was collected for a sample of college students at a university in the Midwest. One variable measured was the number of words per minute that they could type and also whether or not they had previously taken a typing course (Method 1) or if they were self-taught (Method 2).

### Part 2a

Load the dataset Typing from the DS705data package.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

data("Typing")  
head(Typing)

## Method Words  
## 1 prior course 28  
## 2 prior course 46  
## 3 prior course 36  
## 4 prior course 36  
## 5 prior course 28  
## 6 prior course 36

levels(Typing$Method)

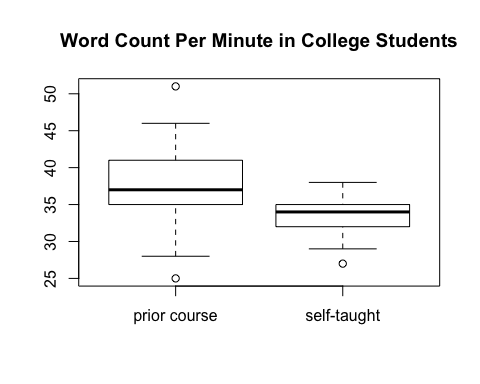
## [1] "prior course" "self-taught"

### Part 2b

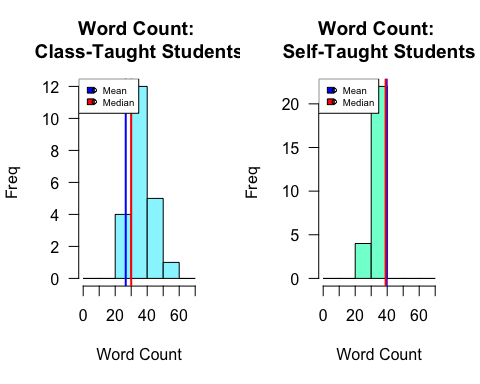
Create boxplots and histograms for each group (previous typing class vs. self-taught). Be sure that each plot is labeled appropriately.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

class <- Typing$Words[Typing$Method != "self-taught"]  
selftaught <- Typing$Words[Typing$Method == "self-taught"]  
#par( mfrow = c( 3, 1 ) )  
boxplot(Typing$Words ~ Typing$Method, data = Typing,  
 main = "Word Count Per Minute in College Students")



par( mfrow = c( 1, 2 ) )  
hist(class, breaks=c(0, 10, 20, 30, 40, 50, 60, 70), xlab = "Word Count",  
 ylab = "Freq",   
 main = "Word Count: \nClass-Taught Students", col = "cadetblue1", las=1)  
abline(v = mean(treatd), col = "blue", lwd = 2)  
abline(v = median(treatd), col = "red", lwd = 2)  
legend("topleft", c("Mean", "Median"), cex=0.6,  
 fill=c("blue", "red"), pch=21)  
hist(selftaught, breaks=c(0, 10, 20, 30, 40, 50, 60, 70), xlab = "Word Count",   
 ylab = "Freq",   
 main = "Word Count: \nSelf-Taught Students",   
 col = "aquamarine", las=1)  
abline(v = mean(untreated), col = "blue", lwd = 2)  
abline(v = median(untreated), col = "red", lwd = 2)  
legend("topleft", c("Mean", "Median"),   
 cex=0.6, fill=c("blue", "red"), pch=21)



### Part 2c

Do the boxplots show any outliers?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2c -|-|-|-|-|-|-|-|-|-|-|-

## The boxplot does show outliers. It indicates one high and one low extreme value among the course-taught students. There also appears to be an outlier on the low end among the self-taught students.

### Part 2d

Describe the shapes of the histograms for the sample data for each sample.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2d -|-|-|-|-|-|-|-|-|-|-|-

## The shapes of both histograms appear to be unimodal and symmetric, indicating a possible normal distribution. There is not a lot of variation among the groups, however, which may be a problem when evaluating normality.

### Part 2e

Conduct an appropriate test to determine if typing speed in each population can be considered as normally distributed. Provide the p-value and the conclusion of the test.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e -|-|-|-|-|-|-|-|-|-|-|-

class\_stest <- shapiro.test(class)  
selftaught\_stest <- shapiro.test(selftaught)  
class\_stest

##   
## Shapiro-Wilk normality test  
##   
## data: class  
## W = 0.96676, p-value = 0.6362

selftaught\_stest

##   
## Shapiro-Wilk normality test  
##   
## data: selftaught  
## W = 0.93509, p-value = 0.1025

### Population 1: Class-taught method of learning typing

The sample was drawn from a normally distributed population.

The sample was **not** drawn from a normally distributed. population.

Since 0.6361637, we do not reject the null hypothesis at the 5% significance level. There is not evidence to show that the sample was **not** drawn from a normally distributed population.

### Population 2: Self-taught method of learning typing

The sample was drawn from a normally distributed population.

The sample was **not** drawn from a normally distributed. population.

Since 0.1025237, we do not reject the null hypothesis at the 5% significance level. There is not evidence to show that the sample was **not** drawn from a normally distributed population.

### Part 2f

Conduct an appropriate test to determine if typing speed in each population can be considered to have equal variances. Provide the p-value and the conclusion of the test.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2f -|-|-|-|-|-|-|-|-|-|-|-

ltest <- leveneTest(Typing$Words ~ Typing$Method)  
ltest\_pval <- ltest$`Pr(>F)`[1]  
ltest

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 1 9.0646 0.004222 \*\*  
## 46   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Hypotheses

Population variances are equal.

Population variances are **not** equal.

### Conclusion

## Since 0.0042222, we **reject** the null hypothesis at the 5% significance level and accept the alternative hypothesis. That is, it is we accept the underlying population variances to be **un-equal**.

### Part 2g

Conduct the test of your choice to test that the population mean typing speed for all college students with a previous course in typing is higher than for those who were self-taught. Let .

#### Step 1

Define the parameters in words in the context of the problem.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g.step1 -|-|-|-|-|-|-|-|-|-|-|-

Let be the population mean words per minute of college students who learned typing skills in a course.

Let be the population mean words per minute of college students who self-taught themselves typing skills.

#### Step 2

State the null and alternative hypotheses for the test.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g.step2 -|-|-|-|-|-|-|-|-|-|-|-

#### Step 3

Use R to generate the output for the test you selected.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g.step3 -|-|-|-|-|-|-|-|-|-|-|-

ttest <- t.test(Typing$Words ~ Typing$Method, data = Typing, alternative = "greater", var.equal = FALSE, conf.level = 0.90)  
ttest

##   
## Welch Two Sample t-test  
##   
## data: Typing$Words by Typing$Method  
## t = 2.6505, df = 26.876, p-value = 0.006652  
## alternative hypothesis: true difference in means is greater than 0  
## 90 percent confidence interval:  
## 1.862042 Inf  
## sample estimates:  
## mean in group prior course mean in group self-taught   
## 37.00000 33.30769

#### Step 4

State both a statistical conclusion at and interpret it in the context of the problem.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g.step4 -|-|-|-|-|-|-|-|-|-|-|-

## Since 0.0066516, we reject the null hypothesis at the 10% significance level. There is sufficient evidence to claim that the population mean of words per minute typed by students who have had a prior course is higher than the population mean for students who are self-taught in typing ( 0.0066516).

### Part 2h

Write an interpretation in the context of the problem for a 90% confidence interval.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2h -|-|-|-|-|-|-|-|-|-|-|-

With 90% confidence, the population mean of words per minute (wpm) typed by students who have had a prior course is at least 1 (wpm) higher than students who are self taught.

### Part 2i

Did you use the separate-variance t-procedures or the pooled t-procedures? Justify your choice, including some discussion of how well the conditions for the hypothesis test and confidence interval procedures were met.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2i -|-|-|-|-|-|-|-|-|-|-|-

I used the separate variance t-procedures to compare the means.

I was able to use t-procedures because the following conditions were met: \* Independent and random samples from a population + This information was given in the problem statement. \* Normally distributed population (verified by Shapiro-Wilk Test) + The results of each Shapiro-Wilk test indicated p-values much greater than in both cases. Due to the results of Shapiro-Wilk, the null hypothesis was not rejected in either case and we accept that both samples come from a normally distributed population.

I used the separate variance t-procedures, rather than pooled, because the results of Levene's Test indicated un-equal variance at significance level of 0.10.

We could not use the pooled t-test because we have not met the condition of equal variance.