Inference for Paired Data

Josh Jarvey

05/29/2020

Create a Word docx from this R Markdown file for the following exercise. Submit the R markdown file and resulting Word docx file.

## Exercise 1

To reduce ankle injuries, restrictive appliances such as taping and spatting (applying tape over the shoe and sock) have been employed. As part of a study at UWL, subjects also completed a 5-point Likert-type scale survey regarding their perceptions of the movement of each ankle appliance during exercise.

Researchers would like to compare the central values for perceptions of the movement of taped ankles compared to spatted ankles using and to estimate the difference with 90% confidence.

### Part 1a

Load the data set AnkleMovement.rda from the DS705 package.

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

#load package and the dataset.  
library(DS705data)  
data("AnkleMovement")

### Part 1b

Create a new variable of the differences, with the perceptions of the spatted ankle (spat) subtracted from the perceptions of the taped ankle (tape).

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

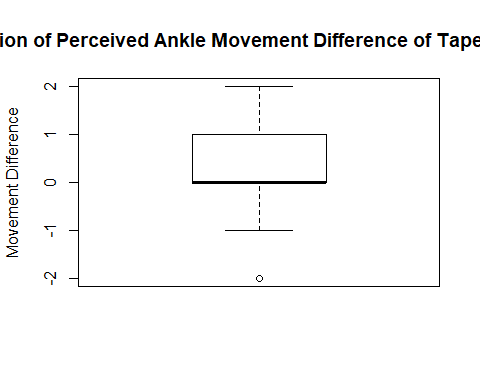
#creating a vector of the differences: tape - spat. will access normality of this difference to determine the appropiate test.   
movement\_difference = AnkleMovement$tape - AnkleMovement$spat

### Part 1c

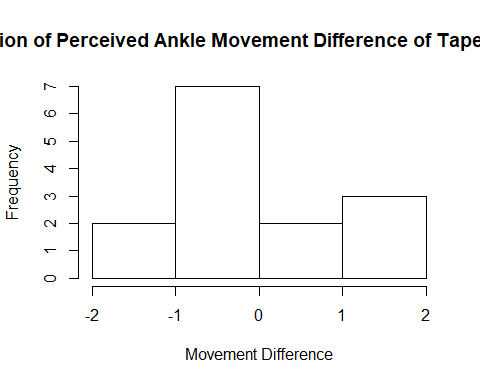
Create a boxplot and histogram for the sample of differences.

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

#creating a boxplot of the difference in perceived movement.  
boxplot(movement\_difference, main = "Distribution of Perceived Ankle Movement Difference of Tape vs. Spatting", ylab = "Movement Difference")



#creating a histogram of the difference in perceived movement.   
hist(movement\_difference,main = "Distribution of Perceived Ankle Movement Difference of Tape vs. Spatting", xlab = "Movement Difference")



### Part 1d

Comment on the suitability of this data for the paired t-test, the Wilcoxon signed rank test, and the sign test.

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

The sample shows an outlier, and the median is quite skewed to one side, however the whiskers of the boxplot are of equal size. Because of this, I lean more toward the t-test with this sample over the Wilcoxon Rank Sign test since the sample does not appear symmetric with the median being skewed.

### Part 1e

Because the choice of test is somewhat unclear, as happens often in real life, try all three tests to compare the central values for subject’s perceptions of the movement of taped ankles compared to spatted ankles using .

Do the t-test first:

#### Step 1

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

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step1 -|-|-|-|-|-|-|-|-|-|-|-

mu\_d: The mean difference in percieved ankle movement score between Tape vs. Spatting for UWL athletes.

#### Step 2

State the null and alternative hypotheses for the test using the symbol you defined previously.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step2 -|-|-|-|-|-|-|-|-|-|-|-

H\_0: mu\_d = 0 H\_a: mu\_d != 0

#### Step 3

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

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step3 -|-|-|-|-|-|-|-|-|-|-|-

#completing the paired t.test(). We just want to test that the perceived movement is different from each other, so two-sided.  
t.test(AnkleMovement$tape,AnkleMovement$spat, alternative = "two.sided", paired = TRUE)

##   
## Paired t-test  
##   
## data: AnkleMovement$tape and AnkleMovement$spat  
## t = 1.1613, df = 13, p-value = 0.2664  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.3072679 1.0215536  
## sample estimates:  
## mean of the differences   
## 0.3571429

#### Step 4

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

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step4 -|-|-|-|-|-|-|-|-|-|-|-

At a 90% level of significance, there is not enough evidence to suggest that there is a difference between the percieved movement score of taped ankles vs. spatted ankles for UWL athletes (p-value = 0.2664).

#### Step 5

Write an interpretation in the context of the problem for the 90% CI for the population mean difference.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step5 -|-|-|-|-|-|-|-|-|-|-|-

At a 90% level of significance, the population mean difference of percieved ankle movement scores among tape and spatting lie within -0.31 to 1.02 units for UWL athletes, thus we found no significant difference.

#### Step 6

Perform the Wilcoxon Signed Rank Test.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step6 -|-|-|-|-|-|-|-|-|-|-|-

#completing the paired wilcoxon rank sign test. We just want to test that the perceived movement is different from each other, so two-sided.   
wilcox.test(AnkleMovement$tape, AnkleMovement$spat, alternative = "two.sided", conf.level = 0.90, conf.int = TRUE, paired = TRUE)

## Warning in wilcox.test.default(AnkleMovement$tape, AnkleMovement$spat,  
## alternative = "two.sided", : cannot compute exact p-value with ties

## Warning in wilcox.test.default(AnkleMovement$tape, AnkleMovement$spat,  
## alternative = "two.sided", : cannot compute exact confidence interval with ties

## Warning in wilcox.test.default(AnkleMovement$tape, AnkleMovement$spat,  
## alternative = "two.sided", : cannot compute exact p-value with zeroes

## Warning in wilcox.test.default(AnkleMovement$tape, AnkleMovement$spat,  
## alternative = "two.sided", : cannot compute exact confidence interval with  
## zeroes

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: AnkleMovement$tape and AnkleMovement$spat  
## V = 20.5, p-value = 0.2981  
## alternative hypothesis: true location shift is not equal to 0  
## 90 percent confidence interval:  
## -0.5000113 2.0000000  
## sample estimates:  
## (pseudo)median   
## 0.9999682

At a 90% level of significance, there is not enough evidence to suggest that there is a difference between the percieved movement score of taped ankles vs. spatted ankles for UWL athletes (p-value = 0.2981).

At a 90% level of significance, the population mean difference of percieved ankle movement scores among tape and spatting lie within -0.50 to 2.00 units for UWL athletes, thus we found no significant difference.

#### Step 7

Perform the sign test.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step7 -|-|-|-|-|-|-|-|-|-|-|-

#load the signmedian.test package so i can use the test.  
library(signmedian.test)  
  
 #completing the signmedian test. We just want to test that the difference in perceived movement has a majority (>0.50) of positive sign differences in the data set.   
signmedian.test(movement\_difference, mu=0, alternative = "greater", conf.level = 0.90)

##   
## Exact sign test  
##   
## data: movement\_difference  
## #(x>0) = 5, mu = 0, p-value = 0.2266  
## alternative hypothesis: the median of x is greater than mu  
## 87.5 percent confidence interval:  
## -1 2  
## sample estimates:  
## point estimator   
## 0

At a 90% level of significance, there is not enough evidence to suggest that there is a difference between the percieved movement score of taped ankles vs. spatted ankles for UWL athletes (p-value = 0.2266).

At a 90% level of significance, the population mean difference of percieved ankle movement scores among tape and spatting lie within -1.00 to 2.00 units for UWL athletes, thus we found no significant difference.

#### Step 8

Construct a bootstrap confidence interval at a 90% level of confidence for the mean difference in population mean perception of movement for taped and spatted ankles. Use a bootstrap sample size of 5000. Compare this interval with the results of the 90% *t*-interval.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step8 -|-|-|-|-|-|-|-|-|-|-|-

#creating an auxillary function for calculating the mean of the data thats passed in.  
bootMeanPaired <- function(x,i){  
 return(mean(x[i]))  
}  
  
 #load the boot strap package.  
library(boot)

## Warning: package 'boot' was built under R version 3.6.3

#simulate 5000 sampling distributions using the difference data, and calculating the mean to generate a sampling distribution.   
boot.object <- boot(movement\_difference, bootMeanPaired, R = 5000)  
  
 #generating the confidence interval bca.   
boot.ci(boot.object, type = "bca", conf = 0.90)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 5000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = boot.object, conf = 0.9, type = "bca")  
##   
## Intervals :   
## Level BCa   
## 90% (-0.2143, 0.7857 )   
## Calculations and Intervals on Original Scale

At a 90% level of significance, the population mean difference of percieved ankle movement scores among tape and spatting lie within -0.21 to 0.79 units for UWL athletes, thus we found no significant difference.

#### Step 9

Compare the results of the three hypothesis tests and also whether or not the 90% bootstrap interval agrees with the result of each test. Which procedure should be reported and why?

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e.step9 -|-|-|-|-|-|-|-|-|-|-|-

All three hypothesis tests agree with each other, as well as the bootstrapped version - there doesnt appear to be a significance in the mean population mean difference of percieved ankle movement scores among tape and spatting for UWL athletes. Because they all agree with each other, I will report the t.test() since this is the most widely recognized test.

## Exercise 2

In a nationwide study of insurance claims (in dollars) filed in the previous year, a random sample of 125 claims was selected from all claims for vehicles classified as small, meaning the gross vehicle weight rating (GVWR) is less than 4500 pounds.

For each claim, the insurance company’s estimate for the claim was also provided.

The data frame SmallAuto.rda contains the claims and estimates for each vehicle class.

### Part 2a

Load the data SmallAuto from the DS705 package.

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

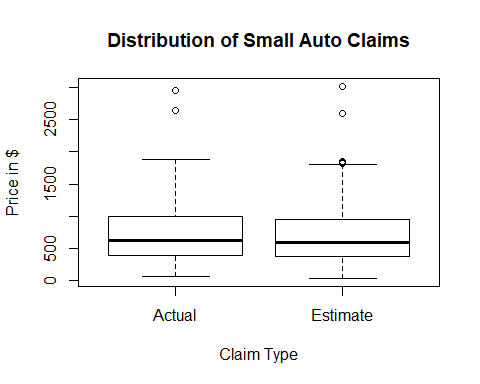
#load in the dataset.  
data("SmallAuto")

### Part 2b

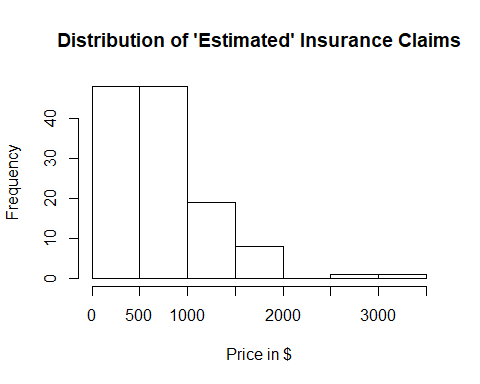
Construct histograms and boxplots for both the estimated claims and actual for the small class of vehicle. Describe the shapes of these distributions.

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

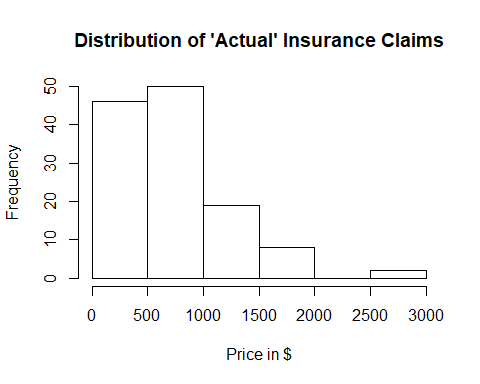
#data is stacked, so pulling each category into its own vector  
estimated = SmallAuto[SmallAuto$Category == "Estimate",1]  
actual = SmallAuto[SmallAuto$Category == "Actual",1]  
  
 #create a boxplot of the claim observations by category.   
boxplot(SmallAuto$Claim~SmallAuto$Category, main = "Distribution of Small Auto Claims", ylab = "Price in $", xlab = "Claim Type")



#generating histograms of the estimated and actual datasets.   
hist(estimated,main = "Distribution of 'Estimated' Insurance Claims", xlab = "Price in $")



hist(actual,main = "Distribution of 'Actual' Insurance Claims", xlab = "Price in $")



The data in both categories display a skew to the right, with the whisker on the boxplot being more heavily weighted to the right. The data also have a moderate amount of outliers.

### Part 2c

Create a new variable of the differences for small vehicles, with the difference being the estimated claim amount minus the actual claim amount. The estimated claim amounts in the first half of the vector are paired with the actual claim amounts in the second half of the vector so that row 1 and row 126 form a pair, rows 2 and 127, etc.

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

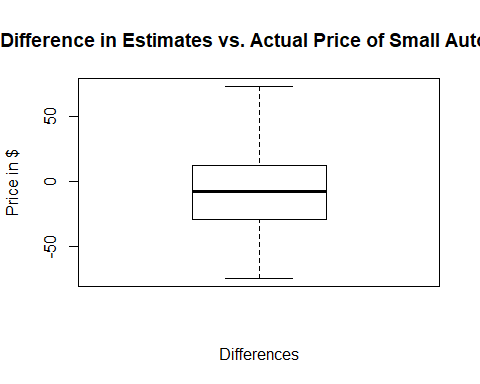
#since this is paired data, calculate the difference between estimated and actual to review the shape of the data and access its normality.   
claim\_difference = estimated - actual

### Part 2c

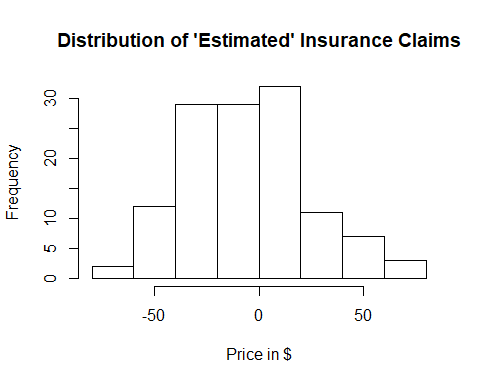
Create a boxplot, histogram, and normal probability plot for the sample of differences. Also, obtain the P-value for a Shapiro-Wilk normality test.

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

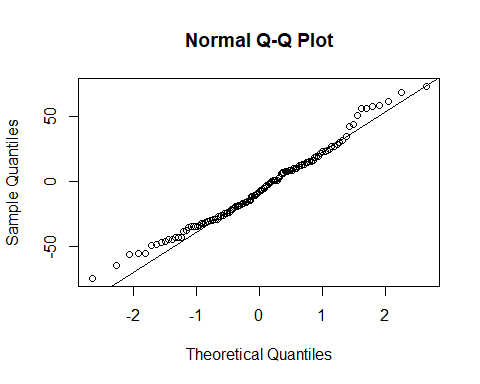
#create a boxplot of the differences between estimate and actual  
boxplot(claim\_difference, main = "Distribution of Difference in Estimates vs. Actual Price of Small Auto Insurance Claims", ylab = "Price in $", xlab = "Differences")



#create a histogram of the differences between estimate and actual  
hist(claim\_difference,main = "Distribution of 'Estimated' Insurance Claims", xlab = "Price in $")



#create a QQ Plot of the differences between estimate and actual  
qqnorm(claim\_difference)  
qqline(claim\_difference)



#test normality of the differences between estimate and actual  
shapiro.test(claim\_difference)

##   
## Shapiro-Wilk normality test  
##   
## data: claim\_difference  
## W = 0.98264, p-value = 0.1093

### Part 2d

Comment on the shape of the distribution of differences and the suitability of this data for the paired *t*-test, the Wilcoxon signed rank test, and the sign test. Which test will you use?

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

The boxplot and histogram show evidence that tend toward normality - the IQR of the boxplot appears evenly spaced with the median line in the center and the whiskers are generally the same shape. The histogram has a relatively mound shaped appearance. The QQ Plot shows a slight skew to the right, which does show slightly on the boxplot. Running the shapiro-wilks test returns a p-value of 0.1093, thus there is not enough evidence to reject that the sample follows a normal distribution.

With all that said, I feel comfortable going forward with the paired t-test since it meets all the criteria: randomly selected, and the differences are normally distributed.

### Part 2e

Conduct an appropriate test to see if the population central values for the estimated claim amount is less than for the actual claim amounts for vehicles in the small class using .

#### Step 1

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

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e.step1 -|-|-|-|-|-|-|-|-|-|-|-

mu\_d: The mean difference in insurance claims between estimated cost vs. actual cost for all small vehicles (GVWR < 4500 lbs) in the nation.

#### Step 2

State the null and alternative hypotheses for the test using the symbol you defined previously.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e.step2 -|-|-|-|-|-|-|-|-|-|-|-

H\_0: mu\_d >= 0 H\_a: mu\_d < 0

#### Step 3

Use R to generate the output for the test you selected. Pay close attention to the order of subtraction done behind the scenes in R.

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e.step3 -|-|-|-|-|-|-|-|-|-|-|-

#complete the paired t-test. estimated > actual at 0.95 significance level.   
t.test(estimated,actual,alternative = "less" ,conf.level = 0.95, paired = TRUE)

##   
## Paired t-test  
##   
## data: estimated and actual  
## t = -2.1537, df = 124, p-value = 0.0166  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -1.318328  
## sample estimates:  
## mean of the differences   
## -5.718576

#### Step 4

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

#### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e.step4 -|-|-|-|-|-|-|-|-|-|-|-

At a 95% significance level, there is enough evidence to support the claim that the mean of estimated costs is less than the mean of actual costs for insurance claims of small vehicles (GVWR < 4500) in the nation (p-value = 0.0166)

### Part 2f

Write an interpretation in the context of the problem for a 95% two-sided confidence interval.

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

#complete the paired t-test. estimated != actual at 0.95 significance level.   
t.test(estimated,actual,alternative = "two.sided" ,conf.level = 0.95, paired = TRUE)

##   
## Paired t-test  
##   
## data: estimated and actual  
## t = -2.1537, df = 124, p-value = 0.03319  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -10.9739093 -0.4632425  
## sample estimates:  
## mean of the differences   
## -5.718576

With 95% confidence, the mean claim cost for estimates is -10.97 to -0.46 dollars less than for actual costs for small vehicles (GVWR < 4500) in the nation.

## Exercise 3

The data frame AutoIns is very similar to Small Auto.

In a nationwide study of insurance claims filed in the previous year, a random sample of 125 claims was selected from all claims for vehicles classified as small, meaning the gross vehicle weight rating (GVWR) is less than 4500 pounds A separate sample of 125 claims for vehicles classified as standard, meaning the GVWR is between 4500 and 8500 pounds.

For each claim, the insurance company’s estimate for the claim was also provided.

The data frame AutoIns.rda contains the claims and estimates for each vehicle class. The variables in the data frame are defined as follows:

claim.small = the actual claim amount in dollars for a vehicle in the small class

est.small = the estimated claim amount in dollars for a vehicle in the small class

claim.standard = the actual claim amount in dollars for a vehicle in the standard class

est.standard = the estimated claim amount in dollars for a vehicle in the standard class

### Part 3a

Load the data AutoIns from the DS705 package and look at the structure of the data in the file.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3a -|-|-|-|-|-|-|-|-|-|-|-

#load in the dataset.   
data("AutoIns")

### Part 3b

Is the data “stacked” or “side-by-side” (“tall” or “wide”)?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3b -|-|-|-|-|-|-|-|-|-|-|-

The data appears to be wide side-by-side since row 1 represents not only two different vehicles, but for each vehicle the actual and estimate costs of the claim.

### Part 3c

Which pairs of variables in the data frame are independent of each other? You can use the variable names to identify them.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3c -|-|-|-|-|-|-|-|-|-|-|-

Because each row contains two vehicles, the claim.small and claim.standard are independent of each other. The est.small and est.standard are also independent of each other.

### Part 3d

Which pairs of variables in the data frame are paired (matched pairs)? You can use the variable names to identify them.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3d -|-|-|-|-|-|-|-|-|-|-|-

The paired data is the claim.small and est.small since that represents one vehicle. The next pair is the claim.standard and est.standard since this is the other vehicle in the dataset.