R Tutorial for STAT 350 Lab 6

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

1. T Procedures for Matched Pairs

The same function, proc t.test(), is used for both the single sample inference, two independent samples and matched pairs. The diagnostics are performed on the difference vector. I have provided the code to generate the vector; however, the code for the plots is the same as before.

Example 1: (Data Set: fuelcomp.txt – website) Fuel efficiency comparison *t* test. One of the authors of this book records the mpg of his car each time he fills the tank. He does this by dividing the miles driven since the last fill-up by the amount of gallons at fill-up. He wants to determine if these calculations differ from what his car's computer estimates.

Fill-up:	1	2	3	4	5	6	7	8	9	10
Computer:	41.5	50.7	36.6	37.3	34.2	45.0	48.0	43.2	47.7	42.2
Driver:	36.5	44.2	37.2	35.6	30.5	40.5	40.0	41.0	42.8	39.2
Fill-up:	11	12	13	14	15	16	17	18	19	20
Computer:	43.2	44.6	48.4	46.4	46.8	39.2	37.3	43.5	44.3	43.3
Driver:	38.8	44.5	45.4	45.3	45.7	34.2	35.2	39.8	44.9	47.5

(a) Make a graphical check for outliers or strong skewness in the data that you will use in your statistical test, and report your conclusions on the validity of the test.
(b) Carry out the significance test to determine if the two methods for calculating the fuel efficiency are the same at a significance level of 0.05.
(c) Give a 95% confidence interval of the difference between the car owner's calculation and the car's computer estimates and interpret the result.

Solution:

```
> mpg=read.table(file="fuelcomp.txt",header=T)
> mpg
```

(a) Make a graphical check for outliers or strong skewness in the data that you will use in your statistical test, and report your conclusions on the validity of the test.

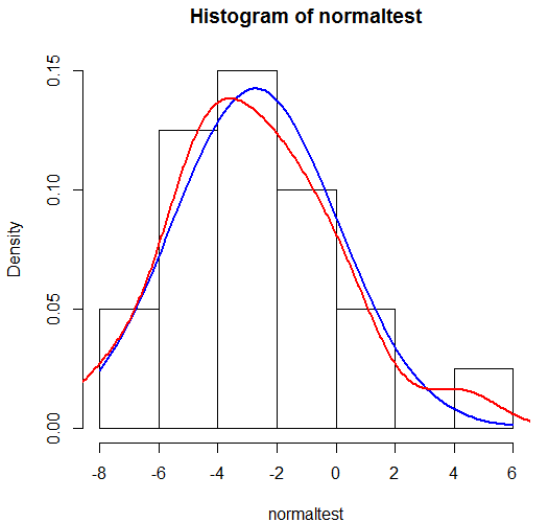
Solution:

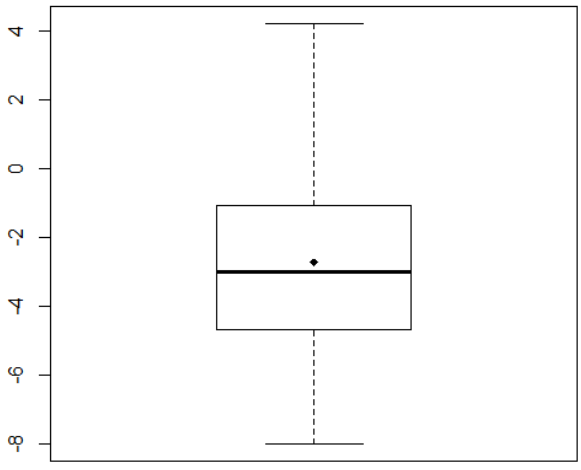
R Tutorial for STAT 350 Lab 6

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

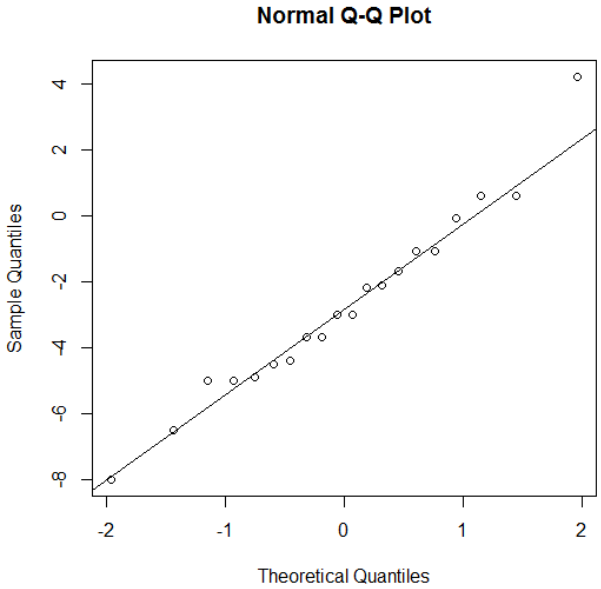
In this case, we need to determine if the difference is normal. The following code shows you how to do that. The rest of the code is not provided.

```
> # the following creates the one sample data. You will need to
> # create the histogram, boxplot and QQPlot on this data set
> # (code not included)
> normaltest = mpg$Driver - mpg$Computer
```





Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi



I do not see any strong skewness or outliers. The data looks reasonably normal. Therefore, the t test should be appropriate.

(b) Carry out the significance test to determine if the two methods for calculating the fuel efficiency are the same at a significance level of 0.05.

Solution:

```
#t.test (x,y,...) is used for confidence intervals and hypothesis tests
#  conf.level = C = 1 - alpha
#  for the hypothesis test. mu is mu_0
#  var.equal = FALSE (the variances are not equal, R calls
#    the Satterthwaite approximation the Welch approximation)
#  alternative = "greater" or "less" or "two.sided" (this is the
#     appropriate alternative hypothesis)
#  paired = True (2 sample paired)
#    The pairing will be x - y

t.test(mpg$Driver,mpg$Computer,conf.level=0.95,paired = TRUE, alternative
   = "two.sided")

        Paired t-test

data:  mpg$Driver and mpg$Computer
t = -4.358, df = 19, p-value = 0.0003386
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.041153 -1.418847
```

```
sample estimates:
mean of the differences
                  -2.73
```

The output for this part is highlighted in yellow.

Step 0: Definition of the terms

μ_D is the population mean difference between fuel efficiency calculated between the driver and the computer.

Step 1: State the hypotheses

H_0: $\mu_D = 0$
H_a: $\mu_D \neq 0$

Step 2: Find the *Test Statistic, report DF*.

$t_t = -4.358$
DF = 19

Step 3: Find the *p-value:*

P-value = 0.0003386

Step 4: Conclusion:

$\alpha = 0.05$

Since $0.0003386 \leq 0.05$, we should reject H_0

The data provides strong evidence (P-value = 0.003386) to the claim that the population mean difference between fuel efficiency calculated between the driver and the computer is different.

(c) Give a 95% confidence interval of the difference between the car owner's calculation and the car's computer estimates and interpret the result.

Solution:

The output for this part is highlighted in green in the previous output.

The 95% confidence interval is (-4.041253, -1.418847).

R Tutorial for STAT 350 Lab 6

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

We are 95% confidence the that the population mean difference between fuel efficiency calculated between the driver and the computer is in the interval (-4.041253, -1.418847)

Not Required: Parts 2 and 3 are the same because 0 is not in the 95% confidence interval.

2. T Procedures for Two Independent Samples

Example 2: **(Data Set: studyhabits.txt – website)** The Survey of Study Habits and Attitudes (SSHA) is a psychological test designed to measure the motivation, study habits, and attitudes toward learning of college students. These factors, along with ability, are important in explaining success in school. Scores on the SSHA range from 0 to 200. A selective private college gives the SSHA to an SRS of both male and female first-year students. The data for the women are as follows:

156	109	137	115	152	140	154	178	111
123	126	126	137	165	165	129	200	150

Here are the scores of the men:

118	140	114	91	180	115	126	92	169	139
121	132	75	88	113	151	70	115	187	114

(a) Examine each sample graphically, with special attention to outliers and skewness. Is the use of a *t* procedure acceptable for these data?
(b) Most studies have found that the mean SSHA score for men is lower than the mean score in a comparable group of women. Carry out this significance test at a 0.01 significance level. That is, state hypotheses, carry out the test and obtain a *P*-value, and give your conclusions.
(c) Give the appropriate 99% confidence bound for the mean difference between the SSHA scores of male and female first-year students at this college. Please interpret the result.

Solution

```
> study=read.table(file="studyhabits.txt",header=T)
> study
```

(a) Examine each sample graphically, with special attention to outliers and

skewness. Is the use of a *t* procedure acceptable for these data?

Solution

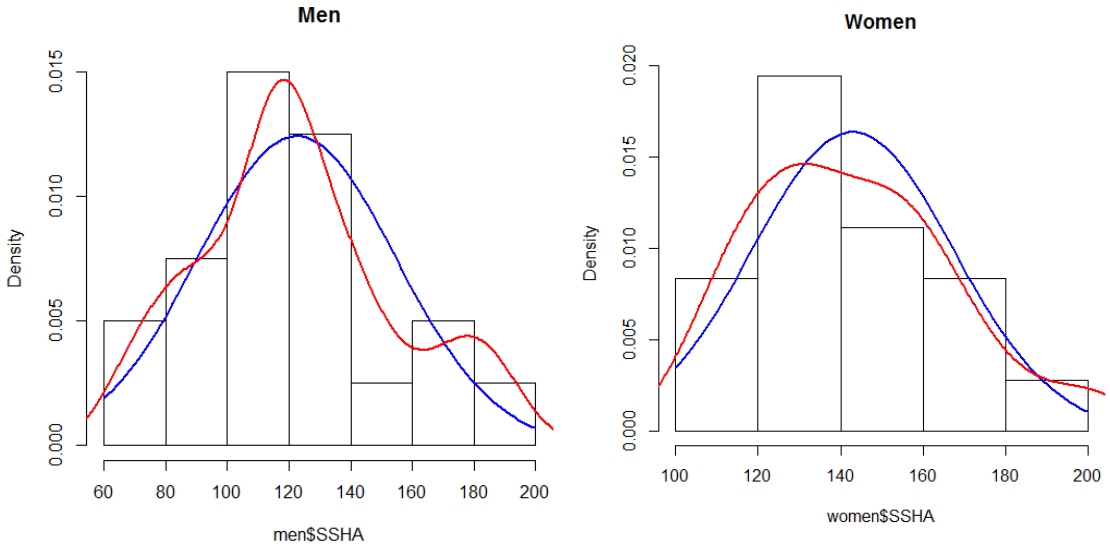
In this case, we want to plot the two populations separately, but they are in the same column. The Lattice package handles this quite easily:

```
> library(lattice)
> attach(study)
> histogram(~SSHA | Sex) #Histograms side-by-side in the same file
> histogram(~SSHA, data = study, subset = Sex == "Men") #Individually
> bwplot(~SSHA | Sex, layout = c(1, 2)) #Boxplots side-by-side
> bwplot(~SSHA, data = study, subset = Sex == "Men") #Individually
> qqmath(~SSHA | Sex) #qqplots side-by-side
> qqmath(~SSHA, data = study, subset = Sex == "Men") #Individually
```

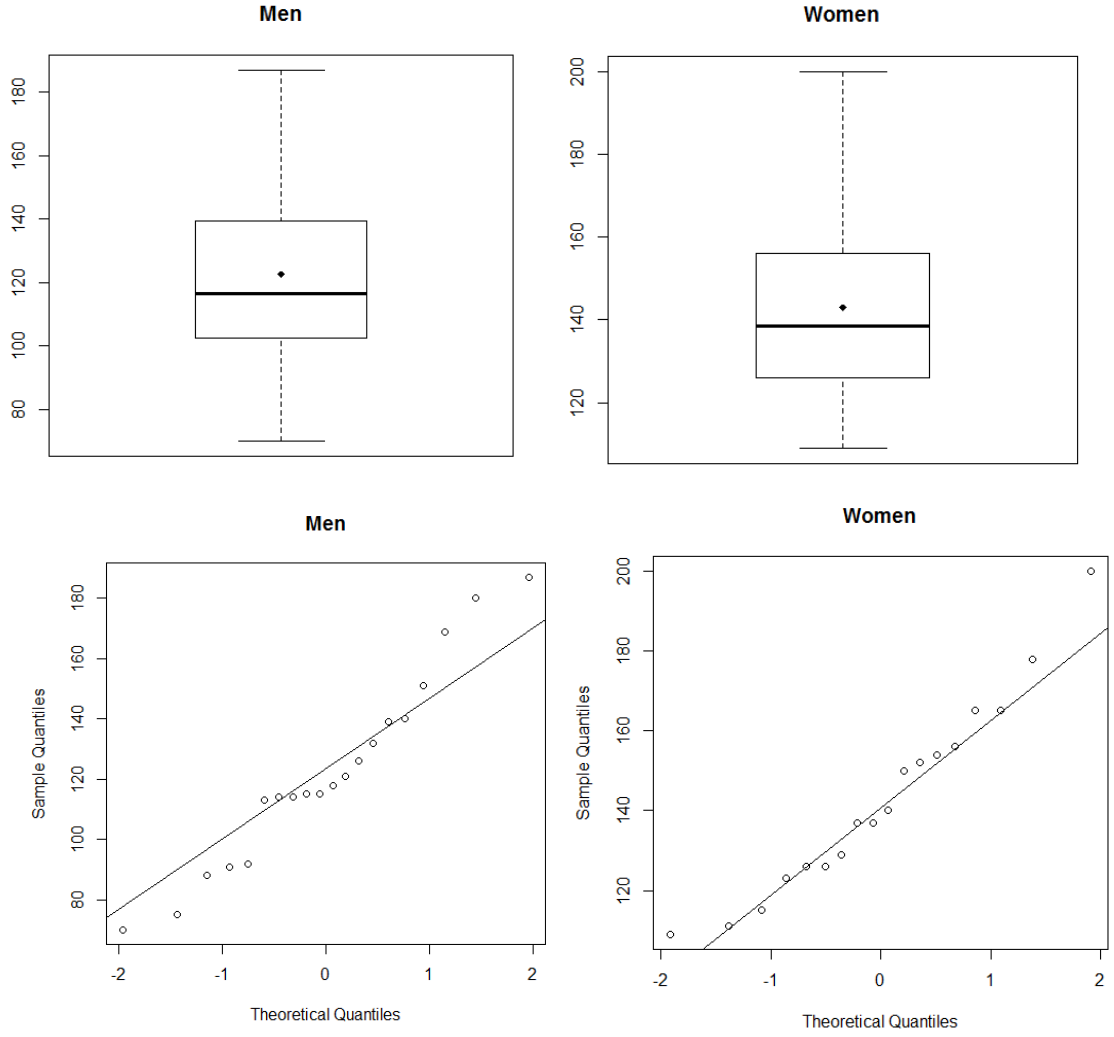
However, should you wish to separate the two populations, the R command is subset().

```
> attach(study)
> men <- subset(study, Gender == "Men")
> women <- subset(study, Gender == "Women")
```

The rest of the code for the graphs utilizing the base R graphics package is not included. We already did this in previous labs, but we will demonstrate how the plots look here:



Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi



Both of these distributions look close to normal with no outliers. Therefore the t procedure is appropriate.

(b) Most studies have found that the mean SSHA score for men is lower than the mean score in a comparable group of women. Carry out this significance test at a 0.01 significance level. That is, state hypotheses, carry out the test and obtain a *P*-value, and give your conclusions.

Solution

R Tutorial for STAT 350 Lab 6

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

```
#t.test (x,y,...) is used for confidence intervals and hypothesis tests
#  conf.level = C = 1 - alpha
#  for the hypothesis test. mu is mu_0
#  the first column is quantitative values ~ categorical column
#  the second column is the name of the table
#  var.equal = FALSE (the variances are not equal, R calls the
#     Satterthwaite approximation the Welch approximation)
#  paired = FALSE (2 sample independent)
#     this will be for x - y.

t.test(men$SSHA, women$SSHA,conf.level=0.99,paired=F,
   alternative = "less",var.equal=F)
```

An alternative code that you can use if you have not separated the two populations is shown next. If you use this code, the two groups will be first alphabetically – second alphabetically.

```
t.test(SSHA ~ Sex, study, conf.level=0.99,paired=F,
   alternative = "less",var.equal=F)
```

`t.test(men$SSHA, women$SSHA, …`	`Welch Two Sample t-test`
	
	`data:  men$SSHA and women$SSHA`
	`t = -2.2232, df = 35.039, p-value = 0.01638`
	`alternative hypothesis: true difference in means is less than 0`
	`99 percent confidence interval:`
	`-Inf 1.971854`
	`sample estimates:`
	`mean of x mean of y`
	`122.5000  142.9444`

`t.test(SSHA ~ Sex, study, …`	`Welch Two Sample t-test`
	
	`data:  SSHA by Sex`
	`t = -2.2232, df = 35.039, p-value = 0.01638`
	`alternative hypothesis: true difference in means is less than 0`
	`99 percent confidence interval:`
	`-Inf 1.971854`
	`sample estimates:`
	`mean in group Men mean in group Women`
	`122.5000  142.9444`

The output for this part is highlighted in yellow.

Step 0: Definition of the terms

Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

$\mu_m - \mu_w$ is the population mean difference between the SSHA scores for men versus women.

Step 1: State the hypotheses

H_0: $\mu_m - \mu_W = 0$
H_a: $\mu_m - \mu_W < 0$

Step 2: Find the *Test Statistic, report DF*.

$t_t = -2.2232$
DF = 35.039 (note, if we would look up the value in the table, this would be looked up as 35)

Step 3: Find the *p-value:*

P-value = 0.01638

Step 4: Conclusion:

$\alpha = 0.01$

Since 0.01638 > 0.01 but it is close, we should fail to reject H_0 maybe

The data might not provide evidence (P-value = 0.01638) to the claim that population mean SSHA scores for men is less than that for women.

(c) Give the appropriate 99% confidence bound for the mean difference between the SSHA scores of male and female first-year students at this college. Please interpret the result.

Solution

The output for this part is highlighted in green in the previous output.

The upper bound is 1.97184.

We are 99% confidence that the difference between the population mean SSHA scores for men versus women is less than 1.97184.

Not Required: The significance test and confidence bound are the same because 0 is less than 1.97184 so the test scores could be the same. However, this is a very small number so if another sample was taken, it could be negative.