R Tutorial for STAT 350 Lab 7

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

1. T Procedures for Matched Pairs

The same function, t.test(), is used for both the single sample inference, 2-sample independent, and 2-sample pairs. However, the diagnostic graphs are performed differently. For the 2-sample independent, the graphs are performed on each of the two populations. Sample code is provided to make the process easier in these cases using the library lattice which was introduced earlier. For the 2-sample pair, the graphs 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 so will not be provided. Remember, the same procedure should be used for both the confidence interval and significance test for a particular problem. Points will be taken off if there are two t.test()s functions for one problem unless multiple confidence intervals or hypothesis tests are requested.

Example 1: (Data Set: ex07-39mpgdiff.txt) 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) Should you use two sample independent or two-sample pairs t procedure to analyze the data? Please explain your answer without referring to the format of the data. If this is a pair situation, please provide a possible confounding variable.
b) Would you consider using a one-sided or two-sided alternative for this analysis? Explain your decision.
c) Do you think these data are Normally distributed? Use graphical methods to examine the appropriate distribution. Write a short summary of your findings.
d) Determine and interpret a 95% confidence interval of the difference between the car owner's calculation and the car's computer estimates.
e) Test the hypothesis that the two methods for calculating the fuel efficiency are the different at a significance level of 0.05.
f) Compare the answers of d) and e). Are they saying the same thing? In addition, answer the original question in one English sentence. You may need additional output to justify your answer.

R Tutorial for STAT 350 Lab 7

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

Solution:

```
mpg=read.table(file="ex07-39mpgdiff.txt",header=T)
attach (mpg)

# For the diagnoistics plots, you will need to create the one sample
# data which is the difference between the two sets. You will need to
# create the histogram, boxplot and QQPlot on this data set
# (code not included)
normaltest = Driver - Computer

#t.test (x,y,conf.level=C, mu = mu0, paired=TRUE, alternative="value")
#      is used for confidence intervals and hypothesis tests
#  conf.level = C = 1 - alpha
#  for the hypothesis test. mu is mu_0
#  paired = TRUE (2-sample paired)
#    The pairing will be x - y
#  alternative = "greater" or "less" or "two.sided" (this is the
#     appropriate alternative hypothesis)
t.test(Driver,Computer, mu=0, conf.level=0.95,paired = TRUE,
   alternative = "two.sided")
# Information required for f
std = sd(normaltest)
size = length(Driver)
se = std/sqrt(size)
```

I chose to computer Driver – Computer because that is the orientation that SAS uses. The order that you use does not matter as long as you are clear what you are doing.

a) Should you use two sample independent or two-sample pairs t procedure to analyze the data? Please explain your answer without referring to the format of the data. If this is a matched pair situation, please provide a possible confounding variable.

Solution:

This should be a matched pair situation because even though the driver and car are the same for each fill up, the conditions during the drive (confounding variable) are different. Stating that the answers are paired in the data set will result in 0 points.

b) Would you consider using a one-sided or two-sided alternative for this analysis? Explain your decision.

Solution:

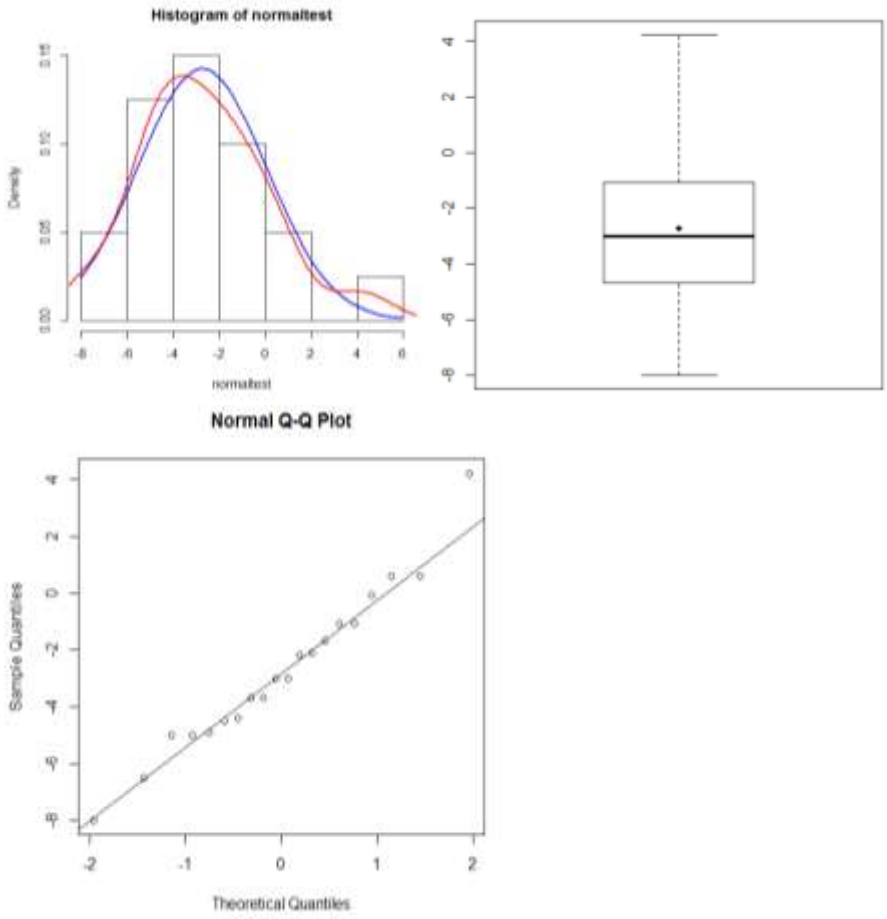
A two-sided alternative hypothesis is preferred here because there is no prior knowledge on whose computations (the car or the drivers) would be more.

R Tutorial for STAT 350 Lab 7

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

c) Do you think these data are Normally distributed? Use graphical methods to examine the appropriate distribution. Write a short summary of your findings.

Solution:



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

d) Determine and interpret a 95% confidence interval of the difference between the car owner's calculation and the car's computer estimates.

Solution:

```
        Paired t-test

data:  Driver and 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
```

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

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

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

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)

e) Test the hypothesis that the two methods for calculating the fuel efficiency are the different at a significance level of 0.05.

Solution:

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

Step 1: Definition of the terms

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

Step 2: State the hypotheses

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

Step 3: Find the *Test Statistic, p-value, report DF:*.

$t_{ts} = -4.358$
DF = 19
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.

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

f) Compare the answers of d) and e). Are they saying the same thing? In addition, answer the original question in one English sentence. You may need additional output to justify your answer.

Solution:

Parts d) and e) say the same thing because 0 is not in the 95% confidence interval therefore the we should reject H_o.

se	0.626439395144873
size	20L
std	2.80152214265557

$$standard\ error = \frac{standard\ deviation}{\sqrt{n}} = \frac{2.801}{\sqrt{20}} = 0.626$$

The calculated fuel efficiency is practically different between the driver and the computer with the computer calculating higher values because

1) The standard error is less than the upper bound (0.6264 < |-1.4188|).
2) It makes sense that it is possible to calculate mpg to the tenths place.
3) The P value (0.0003) is much less than α (0.05).

2. T Procedures for Two Independent Samples

Example 2: (Data Set: studyhabits.txt) 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. Most studies have found that the mean SSHA score for men is lower than the mean score in comparable groups of women. 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

The data for the men is:

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

a) Should you use two sample independent or two-sample pairs t procedure to analyze the data? Please explain your answer without referring to the format of the data. If this is a pair situation, please provide a possible confounding variable.
b) Would you consider using a one-sided or two-sided alternative for this analysis? Explain your decision.
c) Do you think these data are Normally distributed? Use graphical methods to examine the appropriate distribution. Write a short summary of your findings.
d) Test the hypothesis that men score lower than women at a significance level of 0.01.

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

e) Determine and interpret the appropriate 99% confidence bound (for consistency with part d and the question being asked) for the mean difference between the SSHA scores of male and female first-year students at this college.
f) Compare the answers of d) and e). Are they saying the same thing? In addition, answer the original question in one English sentence. You may need additional output to justify your answer.

Solution

```
study=read.table(file="studyhabits.txt",header=T)
# variables used are SSHA: quantitative, Sex: qualitative
#
# Selecting data so only the categories needed are included in the data
#  The table studynew only has the data where the sex is either Men
#     or Women, no other categories will be included (Note: in this
#     data set, the only possible categories are Men and Women
# It is important that you use a double =, '==', rather than any other
#    operator
# Since Sex is a categorical variable, you have to use single quotes to
$   specify the categories. If you are selecting data that is
$   quantitative, no quotes are required.

studynew <- subset(study, Sex=="Men" | Sex == "Women")

attach(studynew)

#  Note: it is required that you have two curves (red and blue) on the
#   histogram and the line on the normal quantile plot as done by the
#   code below.
library(lattice)
histogram(~SSHA | Sex, layout=c(1,2),type="density",
  panel=function(x,...)
   {panel.histogram(x,...)
    panel.mathdensity(dmath=dnorm,col="blue",lwd=2,
       args=list(mean=mean(x, na.rm=T), sd = sd(x,na.rm=T)),...)
    panel.densityplot(x,col="red",lwd=2,...)
    })
bwplot(~SSHA | Sex, layout = c(1, 2), pch = "|") #Boxplots side-by-side
qqmath(~SSHA| Sex, data = studynew, panel = function(x){
    panel.qqmath(x)
    panel.qqmathline(x)
  })
```

R Tutorial for STAT 350 Lab 7

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

```
# t test:
#t.test (qual ~ categories,conf.level=C, mu = mu0, paired=FALSE,
#        alternative="value", var.equal = FALSE)
#      is used for confidence intervals and hypothesis tests
#  the qualitative variable is first, the variable with the groups in
#     it is second.
#     The difference is in first alphabetically - second alphabetically
#  conf.level = C = 1 - alpha
#  for the hypothesis test. mu is mu_0
#  paired = FALSE (2 - sample independent)
#  alternative = "greater" or "less" or "two.sided" (this is the
#     appropriate alternative hypothesis)
#  var.equal = FALSE (the variances are not equal, R calls
#      the Satterthwaite approximation the Welch approximation)

t.test(SSHA ~ Sex, study, mu=0, conf.level=0.99,paired=FALSE,
   alternative = "less",var.equal=F)

# Information required for f
stdmen = sd(subset(study,Sex == "Men")$SSHA)
sizemen = length(subset(study,Sex == "Men")$SSHA)
stdwomen = sd(subset(study,Sex == "Women")$SSHA)
sizewomen = length(subset(study,Sex == "Women")$SSHA)
se = sqrt(stdmen^2/sizemen + stdwomen^2/sizewomen)
```

Selecting data (repeat of the comments above)

When performing a 2-sample procedure, we need to remove all categories except for the ones that we are interested in. Although selection is not required in the sample data set, it is required when there are a large number of categories. We use R subset () function. Note that or is indicated by a vertical line '|'. Note the use of a double = (==) instead of a single =. Since Sex is a categorical variable, be sure to put the categories that we want in double quotes. If a qualitative variable was selected, the double quotes should not be used.

a) Should you use two sample independent or two-sample pairs t procedure to analyze the data? Please explain your answer without referring to the format of the data. If this is a pair situation, please provide a possible confounding variable.

Solution:

The should be a two-sample t procedure because there are no conditions that are mentioned in the description that are not already accounted for. Both sets of students are freshman and we are comparing a difference between men and women. Stating that there are different numbers of scores in men and women will result in 0 points.

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

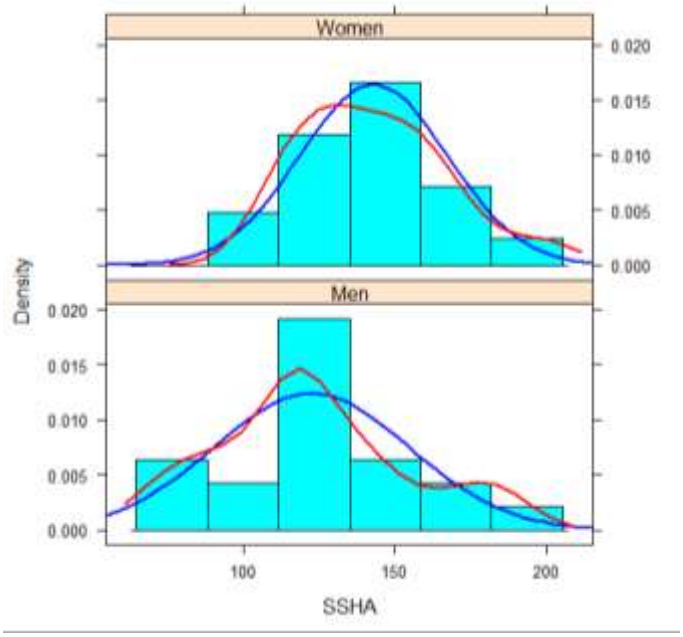
b) Would you consider using a one-sided or two-sided alternative for this analysis? Explain your decision.

Solution:

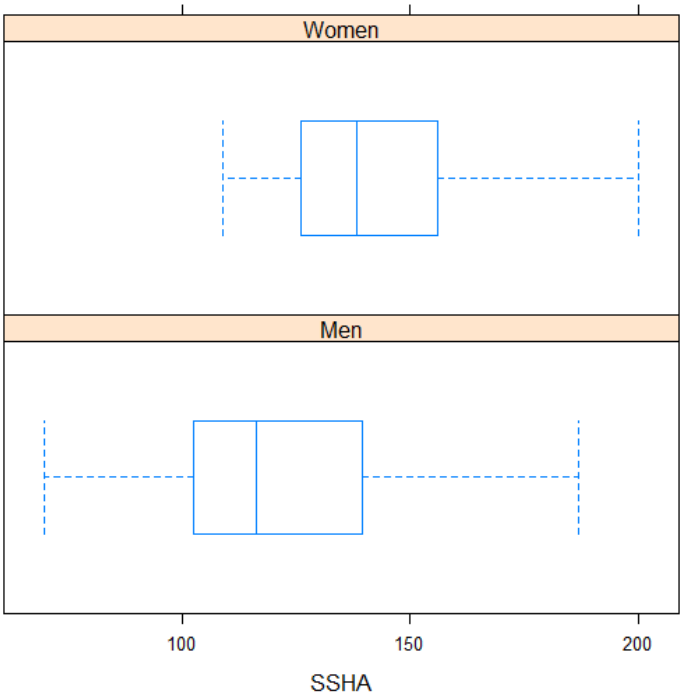
Since we have prior knowledge that mean score for the men is lower than the one for women, a one-sided alternative with H_a: $\mu_{men} - \mu_{women} < 0$.

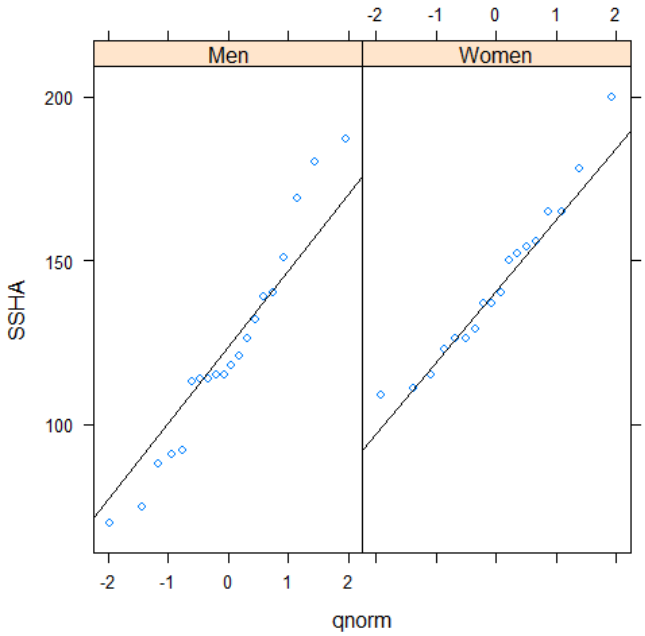
c) Do you think these data are Normally distributed? Use graphical methods to examine the appropriate distribution. Write a short summary of your findings.

Solution:



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.

R Tutorial for STAT 350 Lab 7

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

d) Test the hypothesis that men score lower than women at a significance level of 0.01.

Solution

```
        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 1: Definition of the terms

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

OR

μ_m is the population mean SSHA scores for men.
μ_w is the population mean SSHA scores for women.

Step 2: State the hypotheses

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

Step 3: Find the *Test Statistic, p-value, report DF*.

t_{ts} = - 2.2232
DF = 35.039 (note, if we would look up the value in the table, this would be looked up as 35)
P-value = 0.01638

Step 4: Conclusion:

$\alpha = 0.01$

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

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.

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

e) Determine and interpret the appropriate 99% confidence bound (for consistency with part d and the question being asked) for the mean difference between the SSHA scores of male and female first-year students at this college.

Solution

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

The upper bound is 1.97184.

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

f) Compare the answers of d) and e). Are they saying the same thing? In addition, answer the original question in one English sentence. You may need additional output to justify your answer.

Solution:

Parts d) and e) say the same thing because 0 is less than 1.97184 so the test scores could be the same.

se	9.19608324782174
sizemen	20L
sizewomen	18L
stdmen	32.1321285353231
stdwomen	24.3515242238785

$$Standard\ Error = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{32.1321^2}{20} + \frac{24.3515^2}{18}} = 9.196$$

It is very hard to tell what the practical answer at this significance level is because

1) The standard error, 9.196, is more than the bound, 1.97.
2) I am not sure how accurate these test scores are, that is, is the precision less than 2?
3) The P value (0.0164) is close to the value of α (0.01).

I would suggest another study (or sample) to obtain more convincing results in either direction.