R Tutorial for STAT 350 Lab 7

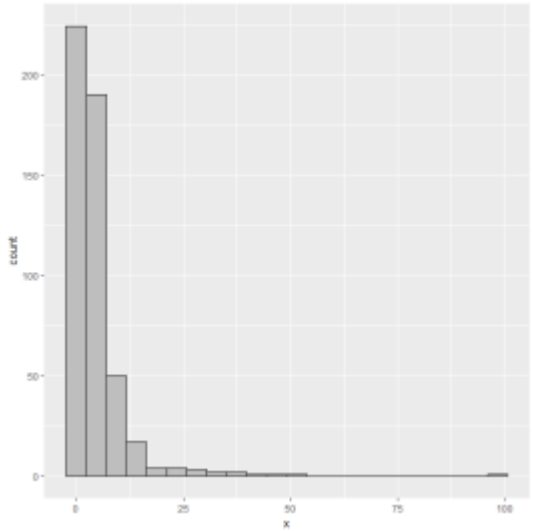
Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

1. Data Transformations

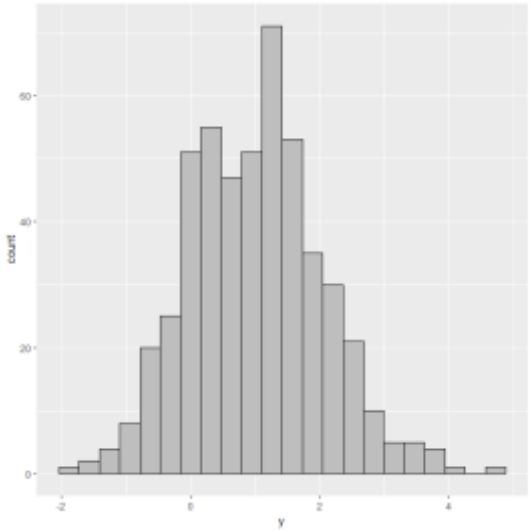
The statistical procedures introduced in this course depend on the assumption of normality (or the conditions for the Central Limit Theorem to be in place). Sometimes, if the data are not well approximated by a normal distribution, we apply a transformation of the data. One such transformation is taking the natural logarithm (or 'ln' which is called "log" in both R and SAS) to make the data more normally distributed. Here is an idealized demonstration of this effect.

```
library(ggplot2)
# Generate random numbers from a skewed distribution
x <- rlnorm(500, 1, 1)
# Histogram (data.frame() can be used when the variable is not
#   stored in a dataset
windows()
ggplot(data.frame(x=x), aes(x)) +
  geom_histogram(bins=sqrt(length(x)), fill="grey", col="black") +
  ggtitle("The original data")
# Take the logarithm
y <- log(x)
windows()
ggplot(data.frame(y=y), aes(y)) +
  geom_histogram(bins=sqrt(length(y)), fill="grey", col="black") +
  ggtitle("The log-transformed data")
```

Original Data



Log-Transformed Data



We did not add the two density curves to avoid coding clutter. However, they should be included when assessing normality. Please refer to Lab 2 tutorial for the code.

R Tutorial for STAT 350 Lab 7

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

2. Selecting Data

When performing a two-sample procedure, we may need to remove categories other than the two that we are interested in. Although selection is not required in the tutorial datasets, it is required when there are a large number of categories. In this case, we use R's `subset()` function. We illustrate this in Example 2 below (see Section 5). I am also including the code here.

```
# Assume that the data are in the data set 'study' and we want
#   restrict the gender of the person to either men or women
study.sub <- subset(study, Sex == "Women" | Sex == "Men")
# Double quotes are used because "Women" and "Men" are text. If the
#   value that you are selecting is a number, no quotes are needed.
```

Remember that in R, all text values like "Women" and "Men" have to be put in double quotes. No quotes are required for numbers. Note that the logical relation "or" is indicated by a vertical line "`|`." The symbols for common operators are in the following table:

Operator name	symbol in R
and	`&`
or	`|`
equal to	`==`
not equal to	`!=`
greater than	`>`
greater than or equal to	`>=`
less than	`<`
less than or equal to	`<=`

3. t-Test of Different Types

In this course we learn three types of t-tests: one sample, two independent samples, and two paired samples. The same function, `t.test()`, is used for all the three types of t-tests. However, the format of the data that we use to generate diagnostic plots and analyze depends on the inference procedure.

We discussed the one sample case in Lab 6. In this lab, we are discussing the two different two-sample cases. Each of these will be described in each of their own section. The procedure is similar, but not identical, in each of these two situations.

For the two-sample paired case, you must generate a new variable which is the difference of the two variables under study. Then you will generate the diagnostic plots and the inference for this new variable. Note that, provided you specify the alternate hypothesis appropriately, the differences $a - b$ and $b - a$ will give consistent hypothesis test results; the order does not matter. *However, you are usually given an order, as in "estimate the mean difference of $a - b$*." We have provided the code to generate the difference variable but not the code for the plots since it is the same as for the one-sample case in Lab 6.

R Tutorial for STAT 350 Lab 7

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

In the two-sample independent case, you simply use the two relevant variables given in the dataset.

Sections 4 and 5 below provide details about the two types of two-sample t-tests via two example questions. Just like in the one-sample problems, the same call to `t.test()` should be used for both the confidence interval and the hypothesis test for a particular inference. Points will be taken off if there are two `t.test()` functions for one inference.

4. t Procedures for Two-Sample Matched Pairs

Example 1: (Data Set: ex07-39mpgdiff.txt) Fuel efficiency comparison. A researcher records the mpg (miles per gallon, a measurement of the fuel economy) of his car each time he filled the tank. He did this by dividing the miles driven since the last fill-up by the amount of gallons pumped 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 a two-sample independent or two-sample paired t procedure to analyze the data? Please explain your answer without referring to the format of the data. If this is a paired situation, please state the common characteristic that makes these data paired.
b) What alternative hypothesis should be used? Please explain your answer.
c) Make a graphical check for outliers or strong skewness in the data that you will use in your statistical inference and report your conclusions on the validity of the procedure.
d) Carry out the hypothesis test to determine if the two methods for calculating the fuel efficiency are the same at a significance level of 0.05.
e) Give a 95% confidence interval of the difference between the car owner's calculation and the car's computer estimates and interpret the result.
f) Compare the answers of d) and e). Are they saying the same thing? What is the final answer to the question. Please also comment on whether there is a practical difference.

Solution:

To read in the data set: Import Dataset → From CSV → Browse to find file → Delimiter: Tab, Name: "mpg" → Import

```
mpg <- read.table(file = "ex07-39mpgdiff.txt", header = TRUE)
library(ggplot2)
```

R Tutorial for STAT 350 Lab 7

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

```
# For the diagnostics 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 QQ plot on this data set.
#   The code is not provided for the graphs. The following is the
#   the variable to use in the plots.
#
# The code you used in Lab 6 assumes the variable to plot is in a
#   data frame. So, we force the difference variable into a data frame.
#
difdata <- data.frame(normaltest = mpg$Driver - mpg$Computer)
#
# INCLUDE CODE FOR DIAGNOSTIC GRAPHS HERE.
#
# In t.test(), you may either use the difference variable and treat it
#   like a one-sample t test. Or, you may specify both variables in the
#   beginning and use paired=TRUE. We will show the latter.
# Parameters for t.test():
#   conf.level = C = 1 - alpha
#   mu: the null mean, mu_0. The default is 0, but we include it for
#     completeness.
#   alternative: form of the alternative hypothesis and confidence
#     interval/bounds, possible options including
#     - "two.sided" (not equal to, confidence interval)
#     - "less" (<, upper confidence bound)
#     - "greater" (>, lower confidence bound)
#   paired: whether the two-sample t test is paired, possible options
#     being TRUE or FALSE. The pairing will be based on the order of
#     values in the two variables.
t.test(mpg$Driver, mpg$Computer, mu = 0, conf.level = 0.95,
       alternative = "two.sided", paired = TRUE)
```

I chose to compute Driver – Computer because of the wording in the confidence interval question, part e). However, Computer – Driver would be acceptable.

a) Should you use a two-sample independent or two-sample paired t procedure to analyze the data? Please explain your answer without referring to the format of the data. If this is a paired situation, please state the common characteristic that makes these data paired.

Solution:

This should be a matched pair situation because even though the driver and car are the same at each fill-up, the conditions during the drive (common characteristic) are different. We want to "subtract out" this confounding factor. Stating that the values are paired in the data set will result in **0** points.

b) What alternative hypothesis should be used? Please explain your answer.

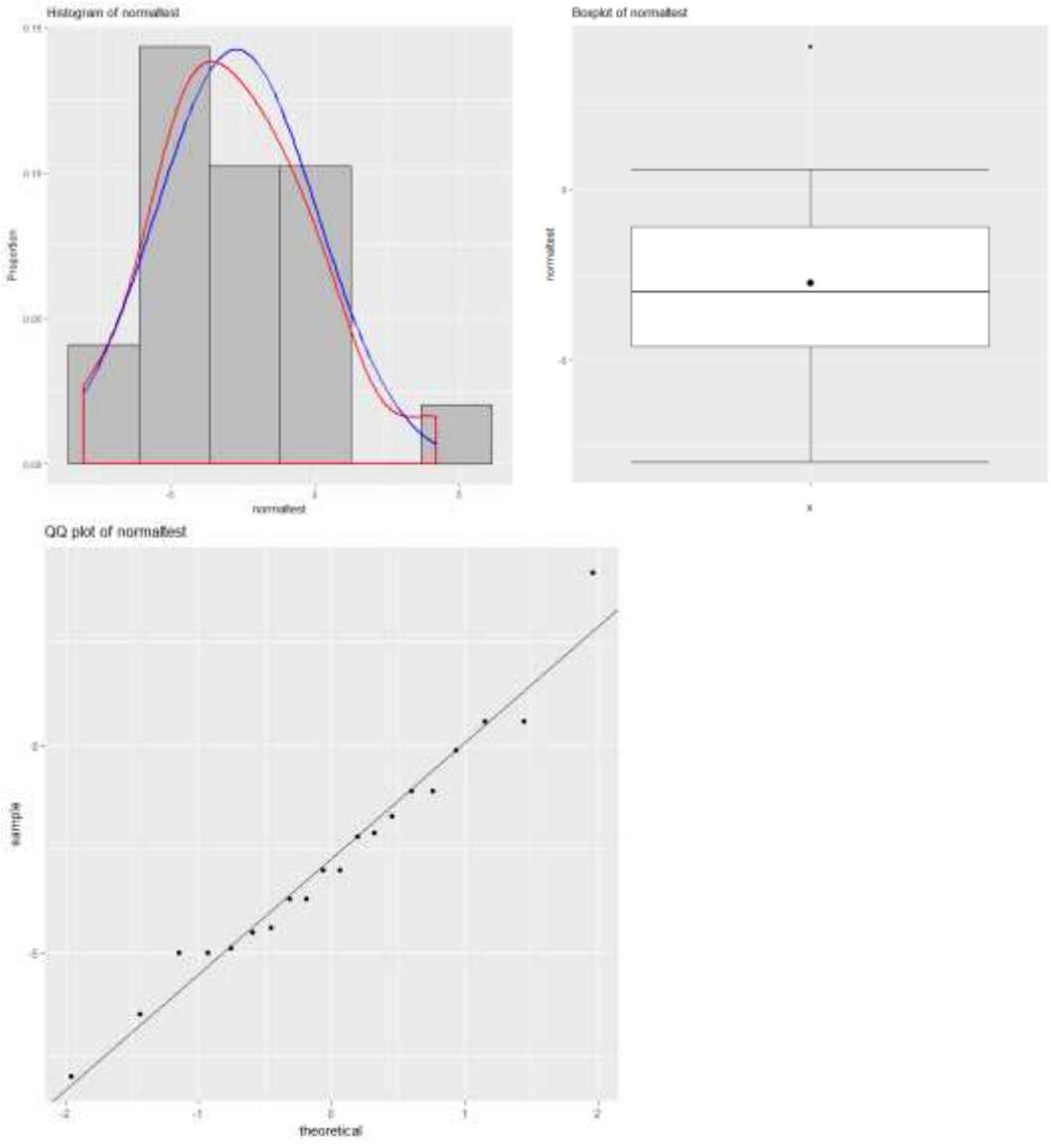
Solution:

A two-sided alternative hypothesis is preferred here because the researcher only wanted to know if the computations (the car's and the driver's) were different..

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

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

Solution:



I do not see any strong skewness or outliers. The data look reasonably normal. Therefore, assuming that the gas mileage calculations are from an SRS, the t procedure should be appropriate.

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

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

Solution:

```
        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 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:

α = 0.05

Since 0.0003386 ≤ 0.05, we should reject H_0.

The data provide strong evidence (p-value = 0.0003386) to the claim that the population mean difference between fuel efficiencies calculated by the driver and by the computer is different.

e) 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).

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

f) Compare the answers of d) and e). Are they saying the same thing? What is the final answer to the question. Please also comment on whether there is a practical difference.

Solution:

Parts d) and e) say the same thing. Note that 0 is not in the 95% confidence interval, indicating it is highly likely that the difference is not 0. Therefore, we should reject Ho,

For the practical analysis, most of the "driver – car" difference values are negative, which implies that computer produces higher numbers than the driver. Note that the upper limit is only -1.4: If you consider 1.4 a significant number, then the numbers are different.

5. 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 a comparable group 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 are:

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

a) Should you use a two-sample independent or two-sample paired t procedure to analyze the data? Please explain your answer without referring to the format of the data. If this is a paired situation, please state the common characteristic that makes these data paired.

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

b) What alternative hypothesis should be used? Please explain your answer.
c) Make a graphical check for outliers or strong skewness in the data that you will use in your statistical inference and report your conclusions on the validity of the procedure.
d) Carry out the hypothesis test at a 0.01 significance level to see if scores for men are lower than that for women.
e) 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.
f) Compare the answers of d) and e). Are they saying the same thing? What is the final answer to the question. Please also comment on whether there is a practical difference.

Solution

The code for diagnostic graphs is provided for this example to demonstrate how to make comparative plots by incorporating the categorical variable.

To read in the data set: Import Dataset → From CSV → Browse to find file → Delimiter: Tab, Name: "study" → Import

```
library(ggplot2)
study <- read.table(file = "studyhabits.txt", header = TRUE)
#
# SUBSETTING/SELECTING DATA
# While subsetting is not needed for this tutorial, we illustrate the
#   command should it be needed in future labs and the project. This
#   will be useful if you want to select only two of multiple groups
#   from a data set.
#   See Section 2 above for more information.
#
# Assume that the data is in the data set 'study' and we want
#   restrict the gender of the person to either men or women
# - Double quotes are used because Women and Men are text. If the
#   value that you are selecting is a number, not quotes are needed.
# - "|" means the logical operator "or"
study.sub <- subset(study, Sex == "Women" | Sex == "Men")
#
# HISTOGRAM FOR EACH GROUP
# (1) Obtain sample mean and standard deviation for each group. Now
#     xbar and s are vectors.
#
xbar  <- tapply(study.sub$SSHA, study.sub$Sex, mean)
s     <- tapply(study.sub$SSHA, study.sub$Sex, sd)
#
# (2) Create the estimated normal density curve by group, based on
#     xbar and s.
#     You need to specify names of the categories in xbar[] and s[].
#
study.sub$normal.density <- ifelse(study.sub$Sex== "Women",
        dnorm(study.sub$SSHA, xbar["Women"] , s["Women"]),
        dnorm(study.sub$SSHA, xbar["Men"], s["Men"]))
```

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

```
#
# (3) ggplot with facet_grid(),
#     The number of bins has to be the same for each of the histograms.
#     facet_grid() tells R what variable contains the categories.
#
windows()
ggplot(study.sub, aes(x = SSHA)) +
  geom_histogram(aes(y = ..density..),
                 bins = sqrt(length(study.sub$SSHA)),
                 fill = "grey", col = "black") +
  facet_grid(Sex ~ .) +
  geom_density(col = "red", lwd = 1) +
  geom_line(aes(y = normal.density), col = "blue", lwd = 1) +
  ggtitle("Histograms of SSHA by Gender")
#
# BOXPLOT
# Specify an x variable (categorical) to graph multiple boxplots
#
windows()
ggplot(study.sub, aes(x = Sex, y = SSHA)) +
  geom_boxplot() +
  stat_boxplot(geom = "errorbar") +
  stat_summary(fun.y = mean, col = "black", geom = "point", size = 3) +
  ggtitle("Boxplots of SSHA by Sex")
#
# QQPLOT FOR EACH GROUP
# (1) Calculate slope and intercept of the reference line in
#     the QQ plot for each group.  These need to be vectors too.
#
study.sub$intercept <- ifelse(study.sub$Sex == "Women",
                              xbar["Women"], xbar["Men"])
study.sub$slope <- ifelse(study.sub$Sex == "Women",
                          s["Women"], s["Men"])
#
# (2) Make QQ plots using facet_grid()
#
windows()
ggplot(study.sub, aes(sample = SSHA)) +
  stat_qq() +
  facet_grid(Sex~ .) +
  geom_abline(data= study.sub, aes(intercept = intercept,
              slope = slope)) +
  ggtitle("QQ Plots of SSHA by sex")
```

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

```
#
# t TEST
# In t.test(), the first argument is
#     quantitativeVariable ~ categoricalVariable,
#   telling R that we want to compare the groups specified by
#   the categoricalVariable in terms of the quantitativeVariable.
# The comparison will be based on the alphabetical order of the group
#   names ("Men" being the first and "Women" being the second).
# Other parameters for t.test():
#   conf.level = C = 1 - alpha
#   mu: the null mean, mu_0. The default is 0, but we include it for
#     completeness.
#   alternative: form of the alternative hypothesis and confidence
#     interval/bounds, possible options including
#     - "two.sided" (not equal to, confidence interval)
#     - "less" (<, upper confidence bound)
#     - "greater" (>, lower confidence bound)
#   paired: whether the two-sample t-test is paired, possible options
#     being TRUE or FALSE. Use FALSE for the two-sample independent
#     case.
#   var.equal: FALSE means the variances of different groups are
#     not assumed equal. In the output, you will see that R calls
#     the Welch approximation (i.e., the Satterthwaite approximation).
t.test(study.sub$SSHA ~ study.sub$Sex, mu = 0, conf.level = 0.99,
       paired = FALSE, alternative = "less", var.equal = FALSE)
```

a) Should you use a two-sample independent or two-sample paired t procedure to analyze the data? Please explain your answer without referring to the format of the data. If this is a paired situation, please state the common characteristic that makes these data paired.

Solution:

This should be a two-sample independent t procedure because there are no conditions mentioned in the description that could be used for matching the male and female students. Both sets of students are freshman and we are comparing a difference between men and women. Stating that "there are different numbers of scores for the men and the women" will result in **0** points.

b) What alternative hypothesis should be used? Please explain your answer.

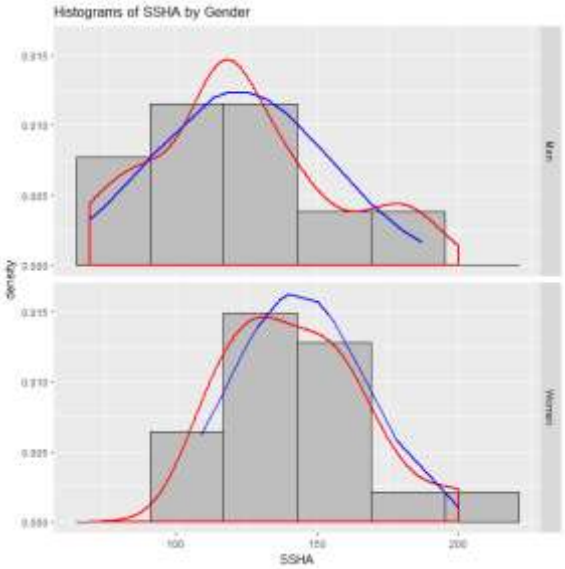
Solution:

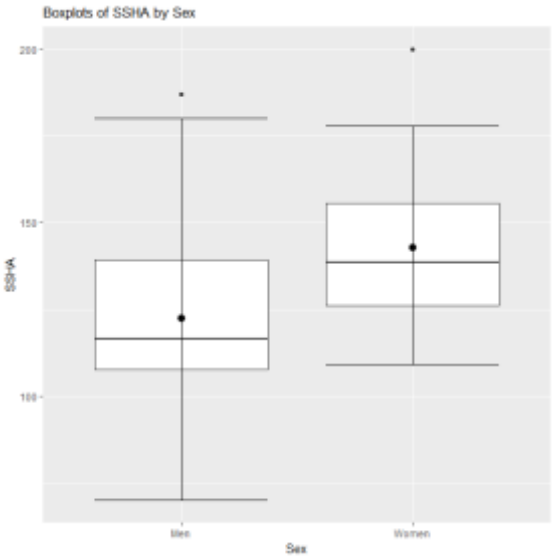
Since we would like to investigate whether the mean score for men is lower than that for women, I would use H_a: $\mu_{men} - \mu_{women} < 0$. Remember in R, the order is alphabetical, and thus the output would not be consistent if you use "$\mu_{women} - \mu_{men} > 0$."

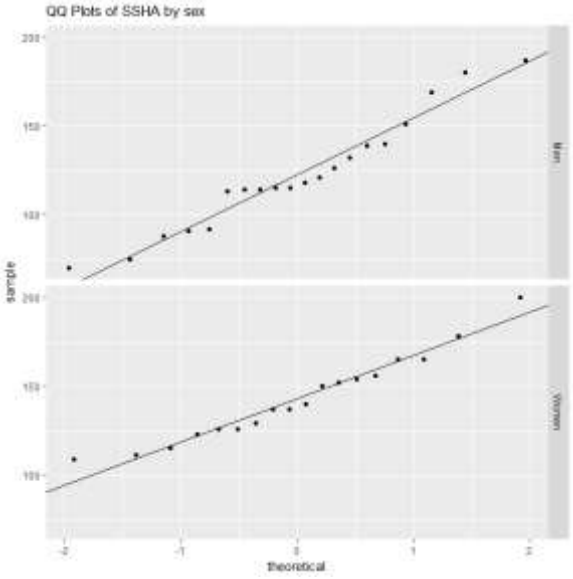
Author: Leonore Findsen, Chunyan Sun, Sarah H. Sellke, Jeremy Troisi

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

Solution:







Neither of these distributions show major outliers or strong skewness in either of the groups. It was already mentioned in the problem description that the data were from an SRS. Therefore, the t procedure is appropriate.

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

d) Carry out the hypothesis test at a 0.01 significance level to see if scores for men are lower than that for women.

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:

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

OR

$\mu_m - \mu_w$ is the population mean difference between the SSHA scores for men versus 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. We always round the degrees of freedom down, to get a more conservative estimate. See your class notes for more details.)
p-value = 0.01638

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

Step 4: Conclusion:

$\alpha = 0.01$

Since 0.01638 > 0.01 we fail to reject H_0 but we recognize that the p value is close to the cutoff.

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.

e) 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 1.971854.

We are 99% confident that the difference between the population mean SSHA scores for men versus women is less than 1.971854.

f) Compare the answers of d) and e). Are they saying the same thing? What is the final answer to the question. Please also comment on whether there is a practical difference.

Solution:

Parts d) and e) say the same thing because 1.971854 is greater than 0 so the difference between the SSHA scores for male and for female students could be non-negative. Similarly, we failed to reject the null hypothesis, that is, the SSHA scores for male students is not less than that for the female students.

Note that the difference of mean SSHA scores for men and for women is small compared to the variability of the scores in either group, which is reflected in the boxplot, which is consistent with the statistical inference results. Practically, we do not think the male scores are significantly less than the female scores.