SAS Tutorial for Lab 7

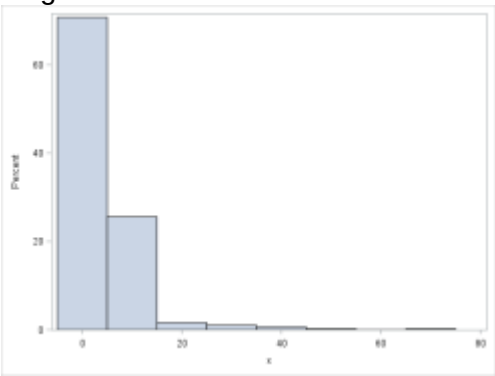
Author: Leonore Findsen, Cheng Li

1. Data Transformations

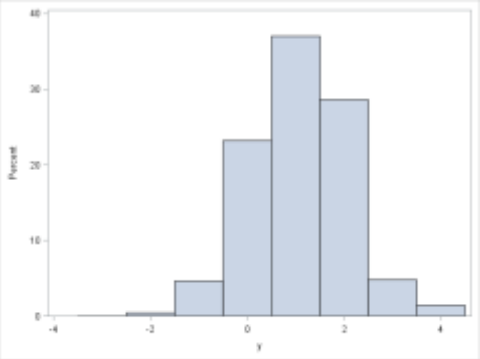
The statistical procedures introduced in this course depend on the assumption of normality (or the conditions of the Central Limit Theorem to be in place). Sometimes, if the data are not well approximated by a normal distribution, we apply a transformation of the data. One such transformation is taking 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.

```
** Create a skewed dataset;
data skew;
  do i = 1 to 500;
    x = rand('lognormal', 1, 1);
    output;
  end;
run;
** Histogram of x;
proc sgplot data = skew;
  histogram x / binwidth = 10;
run;
** Log transform;
data symm;
  set skew;
  y = log(x);
run;
** Histogram of y;
proc sgplot data = symm;
  histogram y / binwidth = 1;
run;
```

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.

SAS Tutorial for Lab 7

Author: Leonore Findsen, Cheng Li

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 an "if" statement. We illustrate this in Example 2 below (see Section 5). I am also including the code here.

```
** Selecting/Subsetting data
   Assume that the data are in data set 'study' and we want to restrict
     the gender of the person to either men or women.;
data studynew;
  set study;
  if Sex = 'Men' or Sex = 'Women';
** The data set studynew only has the data rows where the sex is either
Men or Women;
** The single quotes are used because Sex is a categorical variable, if
   you are selecting a quantitative variable (number), no quotes are
   needed.;
run;
```

Remember that in SAS, all text values like "Women" and "Men" have to be put in quotes. No quotes are required for numbers. The possible operators are shown in the following table.

Operator name	symbol in SAS
and	`and`
or	`or`
equal to	`eq` or =
not equal to	`ne` or ^=
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 procedure, `proc ttest`, is used for all the three types of t-tests. However, the format of the data that we use to analyze depends on the inference procedure.

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 even though the specific numbers will be different. *However, you are usually given an order, as in "estimate the mean difference of $a - b$*

SAS Tutorial for Lab 7

Author: Leonore Findsen, Cheng Li

Sections 4 and 5 below provide details about the two types of two-sample t-tests via two example questions. Just like for one-sample problems, all diagnostic plots are generated from the procedure, and the same procedure should be used for both the confidence interval and hypothesis test for a particular inference. Points will be taken off if there are two `proc ttest`'s 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:

```
data mpg;
  infile "W:\STAT350\ex07-39mpgdiff.txt" delimiter = '09'x
         firstobs = 2;
  input Fillup Computer Driver;
run;
```

Author: Leonore Findsen, Cheng Li

You may also read in the data using the GUI interface:
File --> Import Data --> Tab Delimited File (.txt) --> Next --> browse for the file --> Open --> Next --> Library: Work, Member: wine --> Finish

```
proc ttest data = mpg H0 = 0 sides = 2 alpha = 0.05;
  paired Driver*Computer;
** generates the hypothesis test/CI for matched pair test for
   Driver - Computer;
run;
```

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..

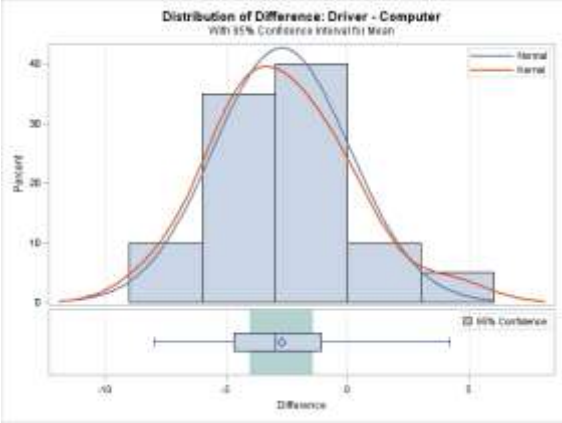
SAS Tutorial for Lab 7

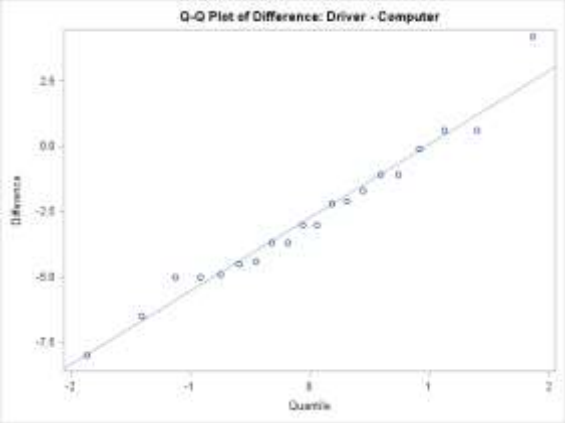
Author: Leonore Findsen, Cheng Li

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:

These graphs were generated by the procedure for the diagnostics.





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.

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:

DF	t Value	Pr > |t|
19	-4.36	0.0003

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:.

$t_t = -4.36$
DF = 19
p-value = 0.0003

SAS Tutorial for Lab 7

Author: Leonore Findsen, Cheng Li

Step 4: Conclusion:

$\alpha = 0.05$

Since $0.0003 \leq 0.05$, 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:

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-2.7300	-4.0412	-1.4188	2.8015	2.1305	4.0918

The 95% confidence interval is (-4.0412, -1.4188).

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 (-4.0412, -1.4188)

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:

SAS Tutorial for Lab 7

Author: Leonore Findsen, Cheng Li

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.
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

```
data study;
  infile "W:\STAT350\studyhabits.txt" delimiter = '09'x firstobs = 2;
  input Student Sex$ Group SSHA;
run;
```

You may also read in the data using the GUI interface:

```
File --> Import Data --> Tab Delimited File (.txt) --> Next --> browse
for the file --> Open --> Next --> Library: Work, Member: wine -->
Finish

** Subsetting/Selecting data;
** This is described in detail on Section 2 above;
** While this step 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
   in a data set.;
```

Author: Leonore Findsen, Cheng Li

```
data studynew;
  set study;
  if Sex = 'Men' or Sex = 'Women';
** The data set studynew only has the data rows where the sex is either
Men or Women;
** The single quotes are used because Sex is a categorical variable, if
   you are selecting a quantitative variable, no quotes are needed.;
** Please see part 2 for the listing of the rest of the possible
   operators.;
run;
proc ttest data = studynew H0 = 0 sides = L alpha = 0.01;
** SAS will always generate the pair alphabetically, therefore it will
be Men - Women in this case;
  class Sex; * categorical variable;
  var SSHA;  * numeric variable;
run;
```

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 "$\mu_{women} - \mu_{men} > 0$."

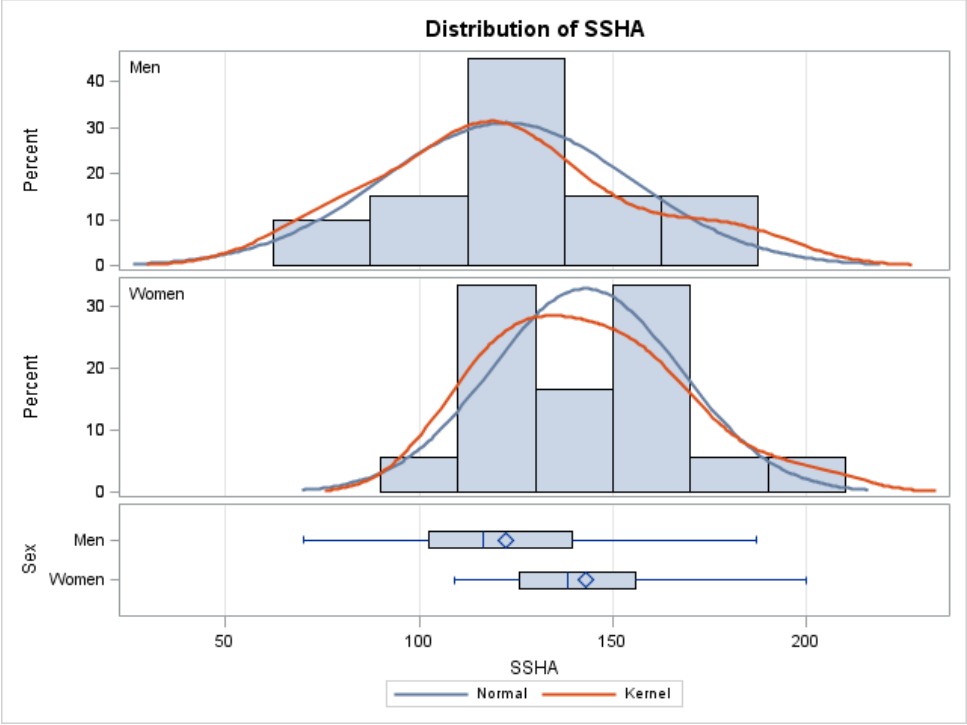
SAS Tutorial for Lab 7

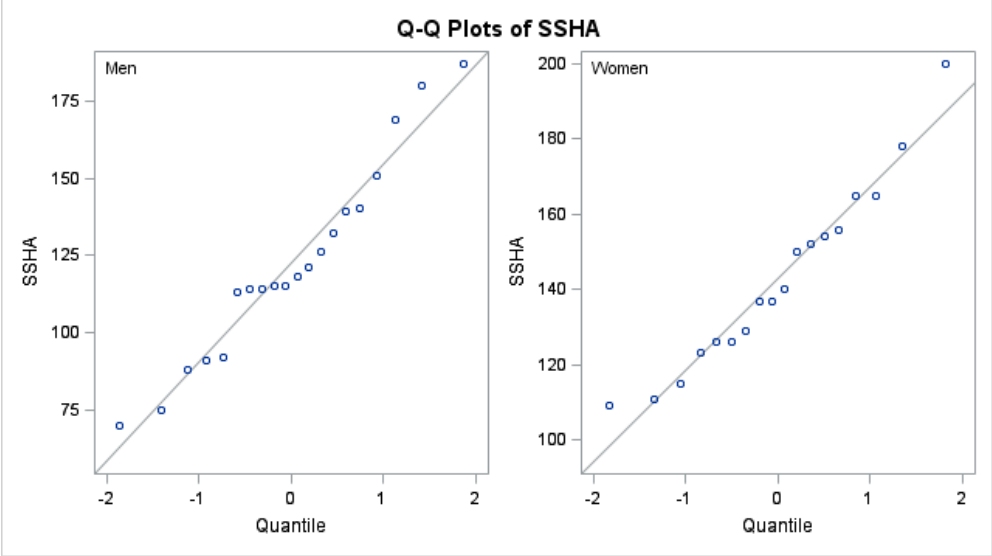
Author: Leonore Findsen, Cheng Li

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:

These graphs were generated by the procedure for the diagnostics.





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.

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:

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	36	-2.19	0.0175
Satterthwaite	Unequal	35.039	-2.22	0.0164

We always use the unpooled (also called Satterthwaite) information.

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.

$t_t = -2.22$
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 = 0.0164

SAS Tutorial for Lab 7

Author: Leonore Findsen, Cheng Li

Step 4: Conclusion:

$\alpha = 0.01$

Since 0.0164 > 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 = 0.0164) 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:

Sex	Method	Mean	99% CL Mean		Std Dev	99% CL Std Dev	
Men		122.5	101.9	143.1	32.1321	22.5488	53.5380
Women		142.9	126.3	159.6	24.3515	16.7998	42.0648
Diff (1-2)	Pooled	-20.4444	-Infty	2.2731	28.7218	21.9603	40.7472
Diff (1-2)	Satterthwaite	-20.4444	-Infty	1.9719			

The upper bound is 1.9719.

We are 99% confident that the difference between the population mean SSHA scores for men versus women is less than 1.9719.

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.