# SAS Tutorial for Lab 7

**Author: Leonore Findsen, Cheng Li**

## 1. T Procedures for 2-sample pairs

The same code, proc ttest, is used for both the single sample inference, two independent samples and two-sample pairs. All diagnostic plots are generated from the procedure. 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 proc ttest's 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.

# SAS Tutorial for Lab 7

**Author: Leonore Findsen, Cheng Li**

**Solution:**

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

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;
```

Since only one code should be used for all parts, I choose to compute Driver – Computer because that is what is asked for in part e).

**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:**

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 greater.

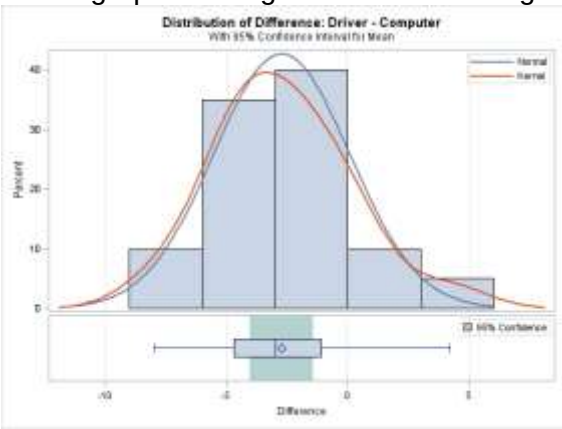
# SAS Tutorial for Lab 7

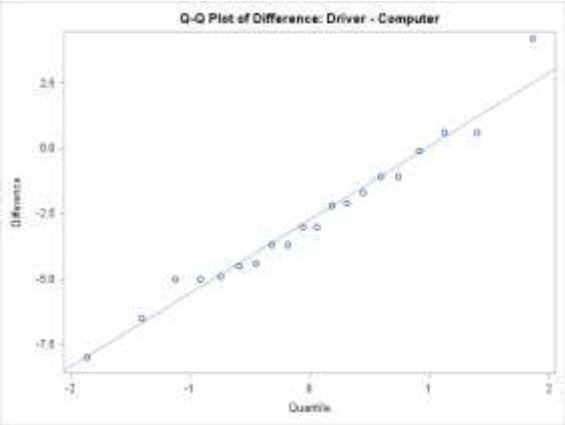
**Author: Leonore Findsen, Cheng Li**

**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:**

These graphs were generated in the diagnostics.




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:**

| 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% confident the that the population mean difference between fuel efficiency calculated between the driver and the computer is in the interval (-4.0412, -1.4188)

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

**Solution:**

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

# SAS Tutorial for Lab 7

**Author: Leonore Findsen, Cheng Li**

**Step 1: Definition of the terms**

$\mu_D$ 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

**Step 4: Conclusion:**

$\alpha$ = 0.05

Since 0.0003 ≤ 0.05 , we should reject $H_0$

The data provides strong evidence (P-value = 0.003) to the claim that the population mean difference between fuel efficiency calculated between the driver and the computer is different.

**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 $H_o$.

| N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|
| 20 | -2.7300 | 2.8015 | 0.6264 | -8.0000 | 4.2000 |

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 $\alpha$ (0.05).

# SAS Tutorial for Lab 7

**Author: Leonore Findsen, Cheng Li**

## 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.
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**

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

*Selecting data so only the categories needed are included in the data
  The data set 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;
data studynew;
   set study;
   if Sex = 'Men' or Sex = 'Women';
*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.
```

# SAS Tutorial for Lab 7

**Author: Leonore Findsen, Cheng Li**

```
*The data needs to be sorted with all of the data for each
   category are together;
proc sort data=studynew; *proc sort sorts the data;
  by Sex; *this is the variable that is used in the sort;
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;
```

**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 like in our airline data set. In SAS, we use an if statement to select the categories that we want. Since Sex is a categorical variable, be sure to put the categories that we want in single quotes. If a quantitative variable was selected, the single 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.

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

**Solution:**

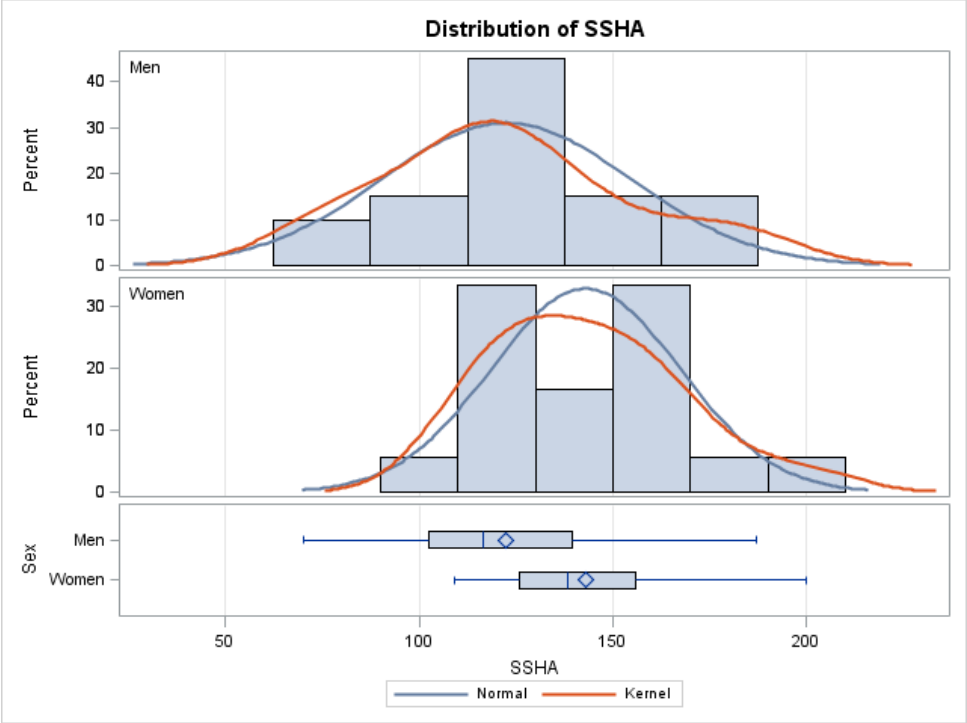
Since we have prior knowledge that mean score for the men is lower than the one for women, a one-sided alternative with $H_a$: $\mu_{men} - \mu_{women} < 0$.

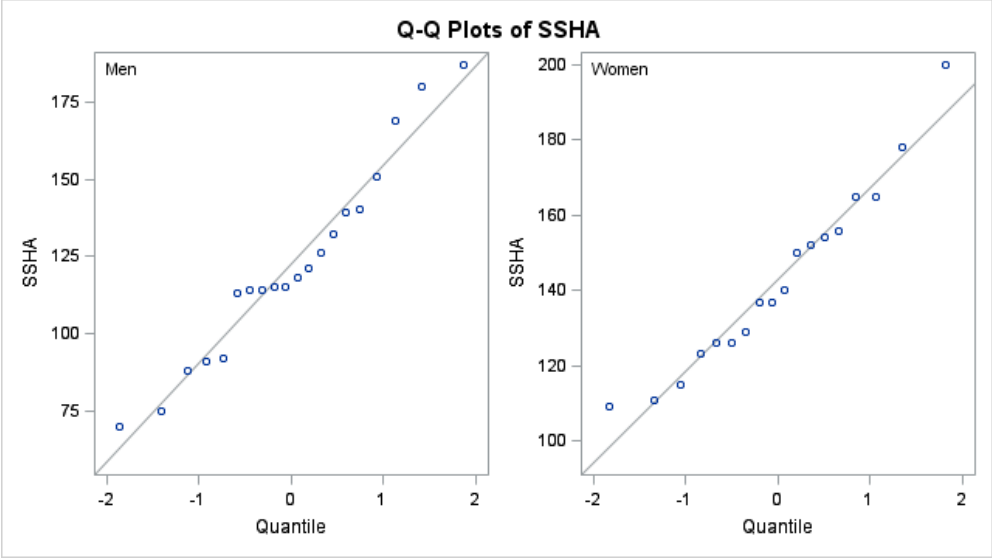
# SAS Tutorial for Lab 7

**Author: Leonore Findsen, Cheng Li**

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

These graphs were generated in the diagnostics.




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

**Author: Leonore Findsen, Cheng Li**

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

| 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 or Satterthwaite information.

**Step 1: Definition of the terms**

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

OR

$\mu_m$ is the population mean SSHA scores for men.
$\mu_w$ 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*.**

$t_t$ = - 2.22
DF = 35.039 (note, if we would look up the value in the table, this would be looked up as 35)
P-value = 0.0164

**Step 4: Conclusion:**

$\alpha = 0.01$

Since 0.0164 > 0.01 but it is close, we should maybe fail to reject $H_0$

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.

# SAS Tutorial for Lab 7

**Author: Leonore Findsen, Cheng Li**

**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.**

| 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? 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.

| Sex | N | Mean | Std Dev | Std Err | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Men | 20 | 122.5 | 32.1321 | 7.1850 | 70.0000 | 187.0 |
| Women | 18 | 142.9 | 24.3515 | 5.7397 | 109.0 | 200.0 |
| Diff (1-2) | | -20.4444 | 28.7218 | 9.3315 | | |

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

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 $\alpha$ (0.01).

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