SAS Tutorial for Lab 6

Author: Leonore Findsen, Cheng Li

1. T Procedures for Matched Pairs

The same code, proc ttest, is used for both the single sample inference, two independent samples and matched pairs.

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.

(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 and interpret the answer. Compare the answers of b) and c). Are they saying the same thing?

Solution:

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

proc print data = mpg;
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;
SAS Tutorial for Lab 6

Author: Leonore Findsen, Cheng Li

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 (c).

(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:

These graphs were generated in the diagnostics.

![Graphs showing normal distribution and Q-Q plot](image)

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:

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

Step 0: Definition of the terms

μ₀ is the population mean difference between fuel efficiency calculated between the driver and the computer.

Step 1: State the hypotheses

H₀: μ₀ = 0
SAS Tutorial for Lab 6

Author: Leonore Findsen, Cheng Li

H₀: μ₀ ≠ 0

Step 2: Find the Test Statistic.

\[ t_i = -4.36 \]

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

DF = 19
P-value = 0.0003

Step 4: Conclusion:

\( \alpha = 0.05 \)

Since 0.0003 ≤ 0.05, we should reject H₀

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.

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

solution:

<table>
<thead>
<tr>
<th>Mean</th>
<th>95% CL Mean</th>
<th>Std Dev</th>
<th>95% CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.7300</td>
<td>-4.0412</td>
<td>-1.4188</td>
<td>2.8015</td>
</tr>
</tbody>
</table>

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

We are 95% confidence the that the population mean difference between fuel efficiency calculated between the driver and the computer is in the interval (-4.0412, -1.4188)

Parts 2 and 3 say the same thing 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:

Here are the scores of the men:

(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. Compare the answers of 3 and 4. Are they saying the same thing?

Solution

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

proc print data = study;
run;

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

proc ttest data = study 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) Examine each sample graphically, with special attention to outliers and skewness. Is the use of a $t$ procedure acceptable for these data?

These graphs were generated in the diagnostics:

- Distribution of SSHA
- Q-Q Plots of SSHA
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Variances</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &lt; t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled</td>
<td>Equal</td>
<td>36</td>
<td>-2.19</td>
<td>0.0175</td>
</tr>
<tr>
<td>Satterthwaite</td>
<td>Unequal</td>
<td>35.039</td>
<td>-2.22</td>
<td>0.0164</td>
</tr>
</tbody>
</table>

We always use the unpooled or Satterthwaite information.

Step 0: Definition of the terms

\( \mu_m - \mu_w \) 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

\( t_i = -2.22 \)

Step 3: Find the \( p \)-value, report DF:

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 fail to reject \( H_0 \) maybe

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.
(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. Compare the answers of 3 and 4. Are they saying the same thing?

<table>
<thead>
<tr>
<th>Sex</th>
<th>Method</th>
<th>Mean</th>
<th>99% CL Mean</th>
<th>Std Dev</th>
<th>99% CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td></td>
<td>122.5</td>
<td>101.9</td>
<td>143.1</td>
<td>32.1321</td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td>142.9</td>
<td>126.3</td>
<td>159.6</td>
<td>24.3515</td>
</tr>
<tr>
<td>Diff (1-2) Pooled</td>
<td></td>
<td>-20.444</td>
<td>-Inf</td>
<td>2.731</td>
<td>28.7218</td>
</tr>
<tr>
<td>Diff (1-2) Satterthwaite</td>
<td></td>
<td>-20.444</td>
<td>-Inf</td>
<td>1.9719</td>
<td></td>
</tr>
</tbody>
</table>

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.

The significance test and confidence bound are the same because 0 is less than 1.9719 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.