Let’s explore SAS Proc T-Test

Ana Yankovsky
Research Statistical Analyst
Screening Programs, AHS
Ana.Yankovsky@albertahealthservices.ca
Goals of the presentation:

1. Look at the structure of Proc TTEST;
2. Talk about different types of tests;
3. Real data examples.
Background:

• The T-Test or Student’s T-Test is any statistical hypothesis test in which the test statistics (t-statistics) follows a Student’s t distribution if the null hypothesis is supported.

• William Sealy Gosset introduce the t-statistics in 1908 while he worked at the Guinness brewery in Dublin.

• He applied his findings to monitor the quality of stout in the production of dark beer.

• Because the company forbade its scientists to publish, Gosset published his work under the pseudonym “Student”.

\[ t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \]
Syntax

• The following statements are available in PROC TTEST:

```
PROC TTEST < options > ;
   CLASS variable ;
   PAIRED variables ;
   BY variables ;
   VAR variables ;
   FREQ variable ;
   WEIGHT variable ;
```

• No statement can be used more than once. There is no restriction on the order of the statements after the PROC statement.
The following options can appear in the **PROC TTEST** statement:

- **ALPHA=p** - specifies the confidence interval. By default ALPHA=0.05;
- **CI=EQUAL** - an equal tailed confidence interval
- **CI=UMPU** - an interval based on the uniformly most powerful unbiased test of $H_0: \sigma = \sigma_0$;
- **CI=NONE** - option requests that no confidence interval be displayed for $\sigma$;
- **COCHRAN** - requests the Cochran and Cox approximation of the probability level of the approximate $t$ statistic for the unequal variances situation;
- **DATA=SAS-data-set** - names the SAS data set for the procedure to use;
- **H0=m** - requests tests against $m$ instead of $0$. By default $H_0=0$. 
BY Statement:

BY variables;

• You can specify a BY statement with PROC TTEST to obtain separate analyses on observations in groups defined by the BY variables.

• When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

➢ Sort the data using the SORT procedure with a similar BY statement.
➢ Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the TTEST procedure.
➢ Create an index on the BY variables using the DATASETS procedure (in base SAS software).
CLASS Statement:

CLASS variable;

• Is giving the name of the classification (or grouping) variable and must accompany the PROC TTEST statement in the two independent sample cases.

• It should be omitted for the one sample or paired comparison situations.

• If it is used without the VAR statement, all numeric variables in the input data set (except those appearing in the CLASS, BY, FREQ, or WEIGHT statement) are included in the analysis.

• The class variable must have two, and only two, levels.

• You can use either a numeric or a character variable in the CLASS statement.
FREQ- Statement:

FREQ variable;

- Identifies a variable that contains the frequency of occurrence of each observation.

- PROC TTEST treats each observation as if it appears $n$ times, where $n$ is the value of the FREQ variable for the observation. If the value is not an integer, only the integer portion is used.

- If the frequency value is less than 1 or is missing, the observation is not used in the analysis.

- When the FREQ statement is not specified, each observation is assigned a frequency of 1.

- The FREQ statement cannot be used if the DATA= data set contains statistics instead of the original observations.
PAIRED Statement:

`PAIRED PairLists ;`

- `PairLists` identifies the variables to be compared in paired comparisons. You can use one or more `PairLists`.

- A pair formed by a variable with itself is ignored.

- Use the PAIRED statement only for paired comparisons.

- The CLASS and VAR statements cannot be used with the PAIRED statement.
PAIRED Statement (continued):

**PAIRED PairLists**;

- Variables or lists of variables are separated by an *asterisk* (*) or a *colon* (:).

- The (*) requests comparisons between each variable on the left with each variable on the right.

- The (:) requests comparisons between the first variable on the left and the first on the right, the second on the left and the second on the right, and so forth. The number of variables on the left must equal the number on the right when the colon is used.

<table>
<thead>
<tr>
<th>These PAIRED statements...</th>
<th>yield these comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAIRED A*B;</td>
<td>A-B</td>
</tr>
<tr>
<td>PAIRED A<em>B C</em>D</td>
<td>A-B and C-D</td>
</tr>
<tr>
<td>PAIRED (A B)*(C D);</td>
<td>A-C, A-D, B-C, and B-D</td>
</tr>
<tr>
<td>PAIRED (A B)*(C B);</td>
<td>A-C, A-B, and B-C</td>
</tr>
</tbody>
</table>
Syntax - Statement:

`VAR` variables;

• The VAR statement names the variables to be used in the analyses;

• One-sample comparisons are conducted when the VAR statement is used without the CLASS statement, while group comparisons are conducted when the VAR statement is used with a CLASS statement;

• If the VAR statement is omitted, all numeric variables in the input data set (except a numeric variable appearing in the BY, CLASS, FREQ, or WEIGHT statement) are included in the analysis.;

• The VAR statement can be used with one- and two-sample t tests and cannot be used with the PAIRED statement;
Syntax - Statement:

`WEIGHT` variable;

- The `WEIGHT` statement weights each observation in the input data set by the value of the `WEIGHT` variable.

- The values of the `WEIGHT` variable can be non integral, and they are not truncated.

- Observations with negative, zero, or missing values for the `WEIGHT` variable are not used in the analyses.

- Each observation is assigned a weight of 1 when the `WEIGHT` statement is not used.

- The `WEIGHT` statement cannot be used with an input data set of summary statistics.
PROC TTEST in SAS performs t-test for:

• One sample observations;
• Two sample observations;
• Paired observations;
The underlying **assumption** of the t test in all three cases is that the observations are:

- Random samples drawn from normally distributed populations. This assumption can be checked using the UNIVARIATE procedure;

  ```
  PROC UNIVARIATE DATA=CLINIC NORMAL PLOT; VAR WEIGHTLOSS; RUN;
  ```

- The two populations of a group comparison must also be independent.

For the **group comparison** PROC TTEST computes t statistic based on the assumption that the variances of the two groups are equal.
One sample t-test

• Tests the null hypothesis that the population mean is equal to the given number specified using the option $\mu_0 = a$

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

• The default value in SAS for H0 is $\mu_0 = 0$. 
One sample t-test: Example

The data that we will look at is:

• from a study that is focused on human weight loss;

• looking at the substitution of costly weight loss face-to-face clinics with phone clinics;

• the primarily goal of a clinic is to deliver personal counseling and educational tools to patients enrolled in a weight loss program;
One sample t-test: Example

For this example, we will compare the mean of the variable `weight_loss` in the clinic group for a pre-selected value of 4 and alpha value of 0.1:

```sas
DATA CLINIC;
   SET TTEST.MI_PHVSCL_CHDATA;
     IF GROUP='Clinic';
     WEIGHT_LOSS=WEIGHT_BASE_KG-WEIGHT_6MON_KG;
RUN;

PROC MI DATA=CLINIC SEED=54987 SIMPLE NIMPUTE=5 OUT=MI_CLINIC;
   MCMC CHAIN = MULTIPLE DISPLAYINIT INITIAL=EM(ITPRINT);
   VAR WEIGHT_LOSS Age RacialCategory;
RUN;

PROC TTEST DATA=MI_CLINIC H0=4 ALPHA=0.1;
   VAR WEIGHT_LOSS;
   TITLE 'TTEST OF H0: MEAN=4 FOR THE CLINIC GROUP';
RUN;
```
One sample t-test: Example Results

**Summary statistics**

**Variable** - This is the list of variables.

**N** - This is the number of valid (i.e., non-missing) observations used in calculating the t-test.

**Lower CL Mean** and **Upper CL Mean** - These are the lower and upper bounds of the confidence interval for the mean.

**Mean** - This is the mean of the variable.

**Lower CL Std Dev** and **Upper CL Std Dev** - Those are the lower and upper bound of the confidence interval for the standard deviation.

**Std Dev** - This is the standard deviation of the variable.

**Std Err** - This is the estimated standard deviation of the sample mean.
One sample t-test : Example

Test Statistics

| DF  | t Value | Pr > |t| |
|-----|---------|------|---|
| 4849| 89.39   | <.0001|

- **DF** - The degrees of freedom for the single sample t-test is simply the number of valid observations minus 1.

- **t Value** - This is the Student t-statistic. It is the ratio of the difference between the sample mean and the given number to the standard error of the mean. Since that the standard error of the mean measure the variability of the sample mean, the smaller the standard error of the mean, the more likely that our sample mean is close to the true population mean.

- **Pr > |t|** - The p-value is the two-tailed probability computed using t distribution. It is the probability of observing a greater absolute value of t under the null hypothesis. For a one-tailed test, halve this probability. If p-value is less than the pre-specified alpha level (usually .05 or .01) we will conclude that mean is statistically significantly different from zero. **In our example, the p-value for weight_loss is smaller than 0.1. So we conclude that the mean for weight_loss is significantly different from 4.**
One sample t-test: Example

Graphs

Distribution of WEIGHT_LOSS
With 90% Confidence Interval for Mean.

Q-Q Plot of WEIGHT_LOSS
Paired t-test

- Paired comparisons use the one sample process on the differences between the observations.

- Paired comparisons can be made between many pairs of variables with one call to PROC TTEST.

<table>
<thead>
<tr>
<th>These PAIRED statements...</th>
<th>yield these comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAIRED A*B;</td>
<td>A-B</td>
</tr>
<tr>
<td>PAIRED A<em>B C</em>D</td>
<td>A-B and C-D</td>
</tr>
<tr>
<td>PAIRED (A B)*(C D);</td>
<td>A-C, A-D, B-C, and B-D</td>
</tr>
<tr>
<td>PAIRED (A B)*(C B);</td>
<td>A-C, A-B, and B-C</td>
</tr>
</tbody>
</table>
Paired t-test: Example

For this example, we will compare the mean of the variable weight loss at the base to the weight loss at 6 months in the clinic group.

```sas
PROC TTEST DATA=MI_CLINIC;
PAIRED WEIGHT_BASE_KG*WEIGHT_6MON_KG;
TITLE 'PAIRED TTEST';
RUN;
```
Paired t-test: Example Results

**Summary statistics**

The TTEST Procedure

**Difference: Weight_Base_Kg - Weight_6mon_Kg**

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>4175</td>
<td>13.5230</td>
<td>7.2927</td>
<td>0.1129</td>
<td>-2.8123</td>
<td>45.0878</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mean</th>
<th>95% CL Mean</th>
<th>Std Dev</th>
<th>95% CL Std Dev</th>
</tr>
</thead>
</table>

| DF | t Value | Pr > |t| |
|----|---------|------|-----|
| 4174 | 119.82 | <.0001 |
Paired t-test: Example Results

Graphs

Distribution of Difference: Weight_Base_Kg - Weight_Smon_Kg
With 95% Confidence Interval for Mean

Agreement of Weight_Smon_Kg and Weight_Base_Kg
Paired t-test: Example Results

Graphs

Paired Profiles for (Weight_Base_Kg, Weight_6mon_Kg)

O-Q Plot of Difference: Weight_Base_Kg - Weight_6mon_Kg
Two sample t-test

• The SAS PROC TTEST procedure is used to test for the equality of means for a two-sample (independent group) t-test.

• The typical hypotheses for a two-sample t-test are:

\[ H_0: \mu_1 = \mu_2 \]
\[ H_a: \mu_1 \neq \mu_2 \]

• Key assumptions: underlying the two-sample t-test are that the random samples are independent and that the populations are normally distributed with equal variances.
Two sample t-test: Example

We would like to know if the mean weight loss in the phone group is equal to the mean weight loss in the clinic group

```
DATA TWOSAMPLE;
  SET TTEST.MI_PHVSCL_CHDATA;
  WEIGHT_LOSS=WEIGHT_BASE_KG-WEIGHT_6MON_KG;
  PROC SORT; BY GROUP;
  RUN;

PROC MI DATA=TWOSAMPLE SEED=54987 SIMPLE NIMPUTE=5 OUT=MI_TWOSAMPLE;
MCMC CHAIN = MULTIPLE DISPLAYINIT INITIAL=EM(ITPRINT);
VAR WEIGHT_LOSS Age RacialCategory;
RUN;

PROC TTEST DATA=TWOSAMPLE;
CLASS GROUP;
VAR WEIGHT_LOSS ;
TITLE 'TTEST OF EQUALITY OF MEANS';
RUN;
```
Two sample t-test: Example Results

**Summary statistics**

- Same as for one sample t-test, but for each group and for the difference of means;
Two sample t-test: Example Results

The **pooled** estimator of variance is a weighted average of the two sample variances, with more weight given to the larger sample;

**Satterthwaite** is an alternative to the pooled-variance t test and is used when the assumption that the two populations have equal variances seems unreasonable.

In our example, there is no equality of variance (Pr>F is less than 0.05), we consider Satterthwaite t statistics, which concludes that the means are different (Pr>|t| is less than 0.05).
Two sample t-test: Example Results

Graphs
References:

1. SAS Online Doc. Chapter 67: The Ttest Procedure

2. Institute For Digital Research And Education – UCLA. SAS Annotated Output. Proc TTest.
   http://www.ats.ucla.edu/stat/sas/output/ttest.htm

3. SAS Tutorials. Independed Group t-test using SAS.
   http://www.stat tutorials.com/SAS/TUTORIAL-PROC-TTEST-INDEPENDENT.htm