Fall 2020 BIOS 500L (Sections 1, 6, 12)
Lab 7: Two-sample t-tests in SAS using PROC TTEST

OBJECTIVE: After completing this lab, students will be able to perform pooled and unpooled two-sampled t-tests in SAS using PROC TTEST.

Introduction

After you have reviewed descriptive information for a particular outcome variable (both graphically and numerically), you are ready to enter the world of inferential statistics. If you recall, the methods of inferential statistics allow you to make a statement about the sampled population based on sample data. While reporting descriptive statistics is an important step in every data analysis, inferential statistics provides you with the tools to address the research question of interest.

Inferential statistics involve: (1) constructing confidence intervals and (2) performing tests of hypothesis. On this lab, we will focus on comparing the means of two independent samples (two-sample t-tests).

PROC TTEST can be used to perform two sample t-tests. Here is the strategy we should follow:

1. Perform descriptive statistical analyses for the continuous variable for which the means are being compared, and the categorical variable that defines the two groups. We can use PROC UNIVARIATE and PROC FREQ; PROC TTEST also produces some descriptive statistics.

2. Assess whether the required conditions for the two-sample t-tests are met or not:
   - Are both sample sizes "large" (i.e., >=30); if not, then are both samples at least approximately normal (if they are, that suggests that the populations from which the samples were drawn are at least approximately normal).

3. Choose between the pooled (equal variance) and unpooled (unequal variance) two-sample t-tests. Do not simply look at the values of the sample variances when making this choice! Instead, if, *a priori*,
   - it can be justified that the population variances are equal, use the pooled test
   - it can be justified that the population variances are unequal, use the unpooled test
   - information about equality of population variances is not available, then either
     - perform the F-test for equality of variances to help you choose between the pooled and unpooled t-tests (information about this test appears on page 3), or
     - choose the t-test with the larger p-value, to be conservative.

4. Perform the appropriate t-test!

Part I: In-lab Demonstration

Your lab instructor will demonstrate SAS code and work with you on the following example during lab:

The Research Question

In a clinical trial, various attributes of stroke patients were recorded (SAS dataset *strokestudy*, posted on Canvas*)*. Suppose our research question is this: are the mean ages of male and female stroke patients the same, or not? A two-sample t-test can be used to make this comparison. One key point in choosing this test is that the two groups are composed of unrelated individuals—the measurements (samples) are independent. In SAS, PROC TTEST can be used to perform a two-sample t-test.

Follow along with the lab instructors as they type and work through the code below in SAS, and discuss the output that is generated.

SAS Syntax Code:

```
/*lab guide*/
*********************************************
BIOS 500 Lab
Weekly Exercise for Lab 7: Two-sample t-tests
*********************************************;

*Make sure you have downloaded the strokestudy data set from Canvas
 Before proceeding with the rest of the code;

*Step 1. Define libname;

LIBNAME u 'U:\'; *Replace with the LIBNAME statement
                  that is appropriate for your session;

*Step 2. Examine the contents of the data set and produce basic
descriptive statistics for the variables gender and age. Gender=1 means
male,=2 means female.

PROC CONTENTS DATA= u.strokestudy;
RUN;

*Freqency table for gender;
PROC FREQ DATA= u.strokestudy;
  TABLES gender;
RUN;

*Descriptive statistics for age;
PROC MEANS DATA= u.strokestudy;
  VAR age;
RUN;
```

```
*Basic descriptive statistics for age, stratified by gender. Also
 Histograms and normal probability plots to help assess whether the
 Sample data for age are at least approximately normal for males and
 Females;
PROC UNIVARIATE DATA= u.strokestudy;
  CLASS gender; *Identify the grouping variable using CLASS statement;
  VAR age;      *Identify the continuous variable using the VAR statement;
  HISTOGRAM age / NORMAL KERNEL;
  PROBPLOT age / NORMAL;
RUN;

/*Step 3
PROC TTEST code to compare the true average ages of females and males
in the stroke study population. Type it into your SAS session exactly as
Shown below, and run it. Your output should look like the output on page
5. Wait for your instructor so that you can work through the strategy on
page 1 together!*/

PROC TTEST DATA = u.strokestudy;
  CLASS gender; *Identify the grouping variable using CLASS statement;
  VAR age;      *Identify the analysis variable using the VAR statement;
RUN;
```

Let's fill in the results from our analysis strategy below, as we go along; the SAS output on pages 5 and 6 will help us:

1. Perform descriptive statistical analyses for the continuous variable for which the means are being compared, and the categorical variable that defines the two groups. Write a sentence or two summarizing the results for each variable:

2. Assess whether the required conditions for the two-sample t-tests are met or not:
   - Are both sample sizes "large" (i.e., >=30); if not, then are both samples at least approximately normally distributed (if they are, that suggests that the populations from which the samples were drawn are at least approximately normal). Write in the result of your assessment below.

3. Choose between the pooled (equal variance) and unpooled (unequal variance) two-sample t-tests. Do not simply look at the values of the sample variances when making this choice! Instead, if, a priori,
   - it can be justified that the population variances are equal, use the pooled test
   - it can be justified that the population variances are unequal, use the unpooled test
   - information about equality of population variances is not available, then either

- perform the F-test for equality of variances to help you choose between the pooled and unpooled t-tests (information about this appears on page 3), or
- choose the t-test with the larger p-value, to be conservative.

Which test should we choose, and why?

4. State the Null and Alternative Hypothesis and significance level

H_0:
H_A:
$\alpha =$

Name of the test:

Justification for being able to conduct the test:

Test statistic value (read directly from SAS output):

P-value (read directly from SAS output):

Decision:

Conclusion:

SAS Output:

Variable: age (age)

gender	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		194	60.5979	11.2523	0.8079	32.0000	81.0000
2		129	61.4806	11.0043	0.9689	34.0000	81.0000
Diff (1-2)	Pooled		-0.8827	11.1541	1.2672		
Diff (1-2)	Satterthwaite		-0.8827		1.2615		

gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		60.5979	59.0046	62.1913	11.2523	10.2330	12.4990
2		61.4806	59.5635	63.3977	11.0043	9.8056	12.5396
Diff (1-2)	Pooled	-0.8827	-3.3757	1.6104	11.1541	10.3540	12.0892
Diff (1-2)	Satterthwaite	-0.8827	-3.3660	1.6006			

PROC TTEST produces sample statistics for age for both sexes including the number of observations, mean standard deviation, standard error of the mean, min, max.

It also produces and 95% confidence limits for the mean age for each sex, and for the standard deviation of age for each sex.

It also produces 95% confidence intervals for the difference in mean ages using both the pooled and unpooled ('Satterthwaite') approach.

Degrees of freedom, t test statistic value, and p-value for the **unpooled** two-sample t-test.

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	321	-0.70	0.4866
Satterthwaite	Unequal	278.56	-0.70	0.4847

Degrees of freedom, t test statistic value, and p-value for the **pooled** two-sample t-test.

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	193	128	1.05	0.7912

"F-test for equality of variances": IF both populations are normally distributed, then we can use the result of this test to choose between the pooled and unpooled two-sample t-tests. The hypotheses for the F-test are:

H_0: $\sigma^2_1 = \sigma^2_2$ the population variances are equal
H_A: $\sigma^2_1 \neq \sigma^2_2$ the population variances are unequal

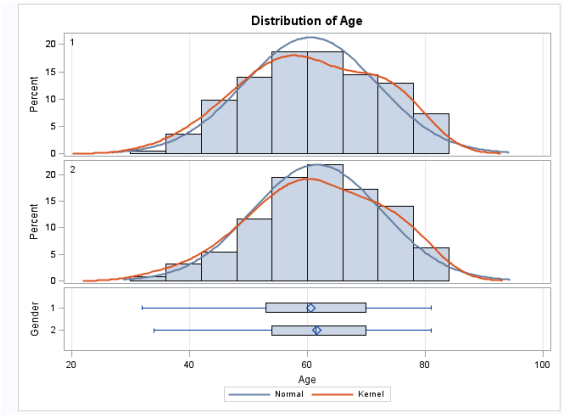
If we fail to reject the null hypothesis, the population variances are assumed to be equal, and we can go on to perform the pooled two-sample t-test.

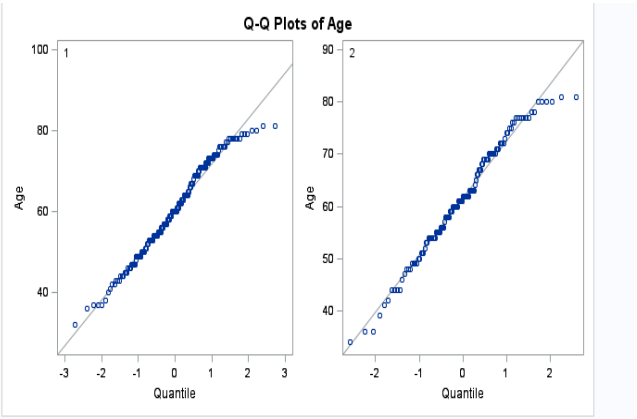
If we reject the null hypothesis, we conclude that the populations variances are unequal, and we can go on to perform the unpooled two-sample t-test.

According to statistical theory, the F-test requires **both** populations to be strictly normally distributed ('approximately normal' is not good enough, and large sample sizes do not help!) For this reason, this F-test is usually not ok to use in real-world studies. Instead, we will need to make the choice of unpooled or pooled t-test using another approach…see the notes on the page 1.

[We will not go over the mathematical approach for determining the test statistic value and p-value for the F-test. Just know that *if* both populations had been known to be normally distributed, the p-value for the F-test (0.7912) would have resulted in failing to reject the null; in other words, we would assume that the population variances were equal.]

In addition to the numerical output above, PROC TTEST produces histograms of age and quantile-quantile plots of age, stratified by gender. If we had not had large enough sample sizes to warrant invocation of the Central Limit Theorem, we could have determined whether the sample data were at least approximately normal for each group using these plots:





In this example, both samples are *approximately* normal, because:

- The histograms appear to be bell-shaped. The red outlines (the red 'Kernel' or best fitting outline for each histogram) are quite close to the blue outlines -which represent true normal distributions.
- The Q-Q plots appear to be approximately linear, suggesting that the quantiles for age for each gender approximately match the quantiles for a true normal distribution.

If the sample data are approximately normal, that suggests that the populations from which the samples were drawn are also approximately normal.

Exercise (Complete and submit on Canvas by 11:59pm on Friday, November 20):

In the *strokestudy* dataset, the variable 'arm' identifies the treatment arm that individuals were randomly assigned to in the clinical trial. We are interested in determining if baseline low-density lipoprotein (LDL) is different between the two arms.

Perform the appropriate analysis, at the 5% significance level.

Submit your well-documented and well-formatted SAS program. Include the following comment at the end of your program making sure you fill in all of the items:

****
H0:
HA:

Alpha:

Name of the test:
Justification for being able to conduct the test:

Test statistic value:

P-value:

Decision:

Conclusion:
**********;