Two Sample t-tests

In this exercise, we will learn how to perform a two sample t-test using software. We will use the dataset `Pulmonary.xls` posted on Blackboard. This dataset describes a multi-center study that enrolled patients and evaluated their lung function by measuring the forced expiratory volume in 1 second (FEV_1).

Often in multi-center studies, there are systematic differences in data from different centers. This is referred to as "between-center variability" and may stem from a variety of reasons, such as the use of different equipment or slightly different procedures from one site to another. It's also possible that the patients attending one center are sicker than patients attending another even though they are theoretically supposed to be the same. For instance, patients of a center in Chicago may have exacerbated symptoms compared to patients of a center at UNC.

The SAS code in this exercise can be found in the file Lab8.sas, which is posted on Blackboard.

1 SAS

1.1 Read In The Data

We will read in the dataset `Pulmonary.xls` using the SAS Import Wizard. Follow the steps below to import the dataset into SAS.

1. From the File menu in SAS, select 'Import Data ...'.
2. Check 'Standard Data Source' and select 'Microsoft Excel 97/2000/2002/2003 Workbook' from the pulldown menu.
3. When the window pops up, click Browse and locate your file ('H:\bios600\Data\pulmonary.xls'). Click Open and then OK.
4. Select the table 'pulmonary function$' (Sheet title) from the pulldown menu and hit Next.
5. Select 'WORK' to put the file in the Work library. Name your SAS dataset in the 'Member:' box by typing 'Pulmonary'.
6. Click Finish.

After completing these steps, click on the Log to view it. You should see the following statement in blue.

```
NOTE: WORK.PULMONARY data set was successfully created.
```

1.2 Descriptive Statistics

Our first step is to explore the data using descriptive statistics on the variable `fev1`. We are interested in describing `fev1` separately for each of the two centers, as well as combined across center. We use the `SORT` and `MEANS` procedures for this.

```
PROC SORT DATA=pulmonary;
  BY center ;
RUN ;

*fev1 stats separately by center;
PROC MEANS DATA=pulmonary N NMISS MEAN STD MIN MAX MAXDEC=4;
  CLASS center ;
  VAR fev1;
RUN ;

*fev1 stats for both centers combined;
PROC MEANS DATA=pulmonary N NMISS MEAN STD MIN MAX MAXDEC=4;
  VAR fev1;
RUN ;
```

1.3 Two Sample T-test

1.3.1 Null Value: 0

Suppose we would like to perform a 2-sided t-test on the variable `fev1`, compared between the two levels of the `center` variable. We will use $\mu_0 = 0$ and significance level $\alpha = 0.05$. That is, we would like to test whether the average FEV_1 from center 1 is significantly different from the average FEV_1 from center 2. We can write these hypotheses in one of the following two ways.

$$H_0 : \mu_1 = \mu_2 \qquad \text{-OR-} \qquad H_0 : \mu_1 - \mu_2 = 0$$
$$H_A : \mu_1 \neq \mu_2 \qquad \text{-OR-} \qquad H_A : \mu_1 - \mu_2 \neq 0$$

To carry out a t-test, we use the `TTEST` procedure. We implement the above hypothesis test in SAS by submitting the following four lines of code.

```
PROC TTEST DATA=pulmonary ;
  CLASS center ;
  VAR fev1 ;
RUN ;
```

Assuming everything runs correctly, SAS will perform the desired two-sample 2-sided hypothesis test and generate the results. The output is shown on the next page.

The TTEST Procedure

Variable: fev1 (fev1)

center	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	21	2.6262	0.4962	0.1083	1.6900	3.4700
2	16	3.0325	0.5232	0.1308	1.7100	3.8600
Diff (1-2)		-0.4063	0.5079	0.1686		

center	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		2.6262	2.4003	2.8520	0.4962	0.3796	0.7165
2		3.0325	2.7537	3.3113	0.5232	0.3865	0.8098
Diff (1-2)	Pooled	-0.4063	-0.7485	-0.0641	0.5079	0.4120	0.6626
Diff (1-2)	Satterthwaite	-0.4063	-0.7524	-0.0602			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	35	-2.41	0.0213
Satterthwaite	Unequal	31.504	-2.39	0.0229

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	15	20	1.11	0.8100

SAS output often includes lots of things you don't need. The most important things that you'll need from this output are

1. First table
   - Descriptive statistics on variable `fev1` for both levels of the variable `center`. ($n_i,\ \bar{x}_i,\ s_i,\ SE_{\bar{x}_i}$ where $i = 1, 2$ denotes centers 1 or 2)
   - Descriptive statistics on the difference between the means of variable `fev1` by variable `center`. ($\bar{x}_1 - \bar{x}_2,\ s_{\text{pooled}},\ SE_{\text{pooled}}$)
2. Second table
   - Two-sided 95% CIs for μ_1 and μ_2, where $i = 1, 2$ denotes centers 1 or 2 (95% CL Mean).
   - Two-sided 95% CI for $\mu_1 - \mu_2$ using pooled measures.
   - Two-sided 95% CI for $\mu_1 - \mu_2$ using non-pooled (Satterthwaite) measures.
3. Third table
   - Pooled Variance t Test: DF, test statistic t, p-value (textbook §12.3).
   - Non-pooled Variance t Test: DF, test statistic t, p-value (textbook §12.4).

1.3.2 Null Value: Not 0

In §1.3.1, we performed a typical two-sample t-test with the hypothesized mean difference $\mu_0 = 0$. While this is the typical setting for a 2-sample t-test, occasionally we want to use some other null value μ_0. That is, we are interested in the more general hypothesis test

$$H_0 : \mu_1 - \mu_2 = \mu_0$$
$$H_0 : \mu_1 - \mu_2 \neq \mu_0$$

If we want to use another value of mu_0, the hypothesized difference between the means, then we modify our previous code very slightly. In the `PROC` statement, we add the SAS keyword `H0=`, along with whatever we have chosen to be our value of μ_0. For instance, if we prefer $\mu_0 = 4$, then we would submit the following code.

```
PROC TTEST DATA=pulmonary H0=4;
  CLASS center ;
  VAR fev1 ;
RUN ;
```

1.3.3 Varying Significance Level

Both of the t-tests above were conducted at the significance level $\alpha = 0.05$, which is the default level. We can change this in the `PROC` statement by adding the SAS Keyword option `alpha=`, along with whatever value we prefer. For instance, to conduct the previous test at the $\alpha = 0.01$ significance level, we would use the following code.

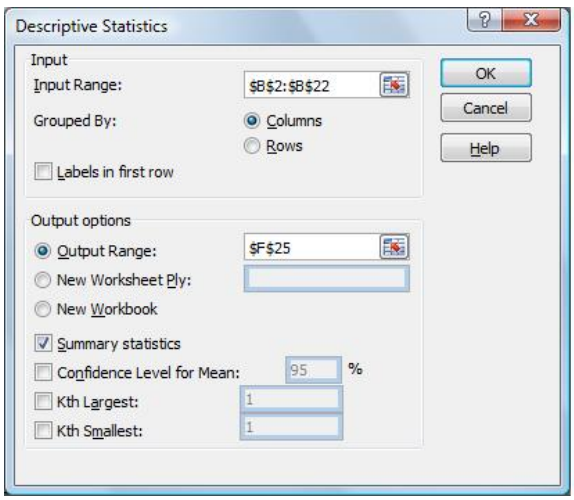
```
PROC TTEST DATA=pulmonary H0=4 alpha=0.01;
  VAR fev1 ;
  CLASS center ;
RUN ;
```

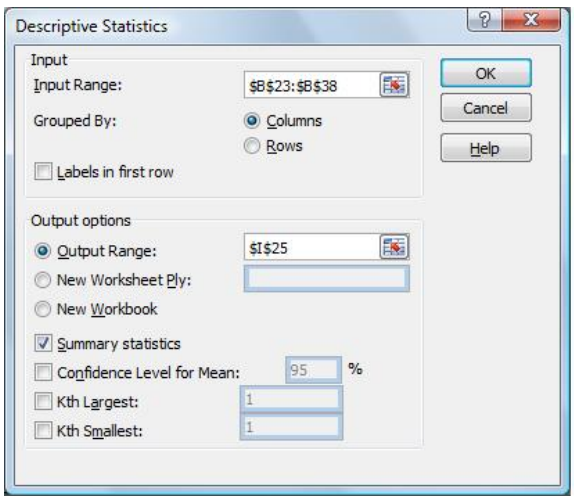
2 Excel

2.1 Descriptive Statistics

The first step of any data analysis is to explore the data using summary statistics. Before we do this however, we will have to **manually delete the rows with missing values** – Rows 13, 31 and 38. To delete a row, select the two cells corresponding to the missing value. With your cursor over the selected area, right-click and then left-click 'Delete...'. Then select 'Shift cells up' and press OK.

We use the Descriptive Statistics tool from the Analysis ToolPak in order to compute summary statistics on the fev1 variable. We must compute these statistics separately for data where center=1 and for data where center=2. Below is a picture of the parameters that we will use with this tool. In order to follow the rest of the exercise instructions exactly, please use these parameters.



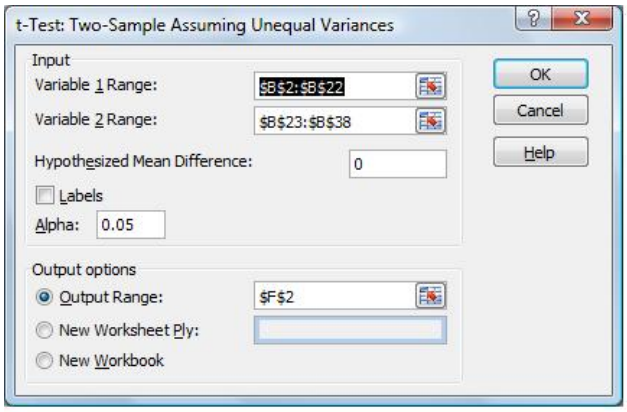


2.2 Two Sample T Test - Unequal Variances

This is equivalent to the first two-sample t-test procedure discussed in Chapter 12 of the textbook (non-pooled method). That is, Excel uses the non-pooled estimate of the variance in order to compute tests and confidence intervals.

We can use the Two Sample t-test with Unequal Variances tool from the Analysis ToolPak in order to conduct the desired t-test. These are the steps we shall use to carry out the two-sample t-test on fev1 between the two centers. The parameters used in this exercise are shown in the picture below.

1. Open the Data Analysis dialog box from the Data ribbon.
2. Highlight "t-Test: Two-Sample Assuming Unequal Variances" and press OK.
3. Set Variable 1: Select all fev1 values with center=1.
4. Set Variable 2: Select all fev1 values with center=2.
5. Set Hypothesized Mean Difference to be 0 (or whatever value you want).
6. Set Alpha to 0.05 (or whatever value you want).
7. Decide where you want the output to print. Press OK.

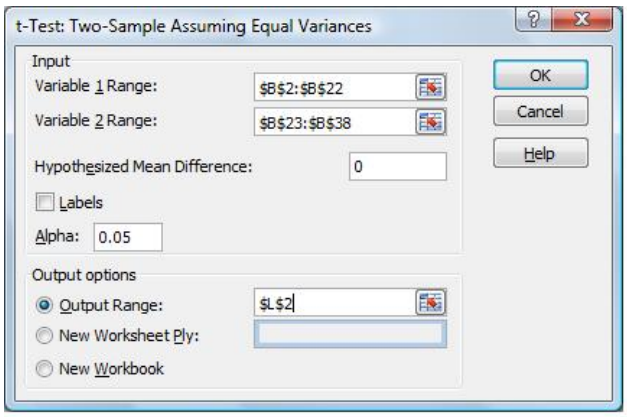


2.3 Two Sample T Test - Equal Variances

This is equivalent to the pooled variance t-test procedure discussed in Chapter 12.4 of the textbook. That is, Excel uses the pooled estimate of the variance in order to compute tests and confidence intervals.

We can use the Two Sample t-test with Equal Variances tool from the Analysis ToolPak in order to conduct the desired t-test. These are the steps we shall use to carry out the two-sample t-test on fev1 between the two centers. The parameters used in this exercise are shown in the picture below.

1. Open the Data Analysis dialog box from the Data ribbon.
2. Highlight "t-Test: Two-Sample Assuming Equal Variances" and press OK.
3. Set Variable 1: Select all fev1 values with center=1.
4. Set Variable 2: Select all fev1 values with center=2.
5. Set Hypothesized Mean Difference to be 0 (or whatever value you want).
6. Set Alpha to 0.05 (or whatever value you want).
7. Decide where you want the output to print. Press OK.



3 References

1. UCLA ATS- SAS Annotated Output: Proc ttest
   http://www.ats.ucla.edu/stat/sas/output/ttest.htm