Lab #7: Confidence Intervals-Hypothesis Testing (2)-T Test

A. Objectives:

1. Subsetting based on variable
2. Explore Normality
3. Explore Hypothesis testing using T-Tests

Confidence intervals and initial inferences

As you recall, the **PROC IMPORT** statement is the best way to enter external data sets. The CSV file we will be using is called “Bench.csv”. Download and save it in your favorite folder and mark the complete path to it. Then use the following code to import it, making sure you put the correct path on the **DATAFILE** argument.

```
PROC IMPORT OUT= Bench
            DATAFILE= "put the complete path here…\Bench.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

31 College students who signed up for weight training were asked to do one maximum bench press and the results were captured in the column benchpress.

Let’s try to print our dataset using the familiar **PROC PRINT** command as follows:

```
PROC PRINT data=Bench;
TITLE 'Maximum Benchpress';
RUN;
```

In order for us to better analyze the data, let’s sort it with respect to genders using the code:

```
PROC SORT data=Bench;
By Gender;
RUN;
PROC PRINT data=Bench;
TITLE 'Maximum Benchpress sorted by Gender';
RUN;
```

Let’s then explore our results and and try to find summary statistics for the total population by using proc univariate as follows:

```
ODS select BasicMeasures;
PROC UNIVARIATE DATA = Bench;
var Benchpress;
TITLE "Maximum Benchpress Statistics";
RUN;
```

This will only output the Basic statistical measures (BasicMeasures) in the dataset Bench for the variable Benchpress giving us the table:

Basic Statistical Measures			
Location		**Variability**	
Mean	82.61290	**Std Deviation**	50.87021
Median	68.00000	**Variance**	2588
Mode	54.00000	**Range**	197.00000
		Interquartile Range	60.00000

Let’s now create two smaller sets, by using the variable Gender. This is done by creating two new datasets from the set Bench named BenchF and BenchM respectively. We are using similar techniques as previous labs by now we use the **IF (Gender = ‘F’);** statement to pick only the entries with F in the gender column.

```
TITLE "Subsetting based on Gender";
Data BenchF;
SET Bench;
IF (Gender = ‘F’);
RUN;
PROC PRINT data= BenchF;
RUN;
```

Similarly we have:

```
Data BenchM;
SET Bench;
IF (Gender = ‘M’);
RUN;
PROC PRINT data= BenchM;
RUN;
```

This created the subset for males only by using the **IF (Gender = ‘M’);** statement.

Exploring Normality:

Let’s explore now how normal each of the two subsets are by using the codes which we saw in a previous lab:

```
TITLE “How Normal is the female Benchpress Histogram?”;
Ods select Histogram ParameterEstimates GoodnessOfFit FitQuantiles Bins;
PROC UNIVARIATE DATA = BenchF;
HISTOGRAM Benchpress/ normal(percents=20 40 60 80 midpercents);
INSET n normal(ksdpval) / pos = ne format =6.3;
RUN;
```

```
TITLE “How Normal is the male Benchpress Histogram?”;
Ods select Histogram ParameterEstimates GoodnessOfFit FitQuantiles Bins;
PROC UNIVARIATE DATA = BenchM;
HISTOGRAM Benchpress/ normal(percents=20 40 60 80 midpercents);
INSET n normal(ksdpval) / pos = ne format =6.3;
RUN;
```

Besides the histograms, we also got the goodness of fir tests for the normal distribution. As we learned in class, as long as the p values are greater than 0.05 we can assume that our data is “normal enough”, since the corresponding null hypothesis is “H_0: The data comes from a normal distribution”. Thus if p is large enough we fail to reject that hypothesis and we can safely proceed with our analysis.
The goodness of fit tests for females are:

Goodness-of-Fit Tests for Normal Distribution (Female)				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.13478581	**Pr > D**	>0.150
Cramer-von Mises	**W-Sq**	0.02915674	**Pr > W-Sq**	>0.250
Anderson-Darling	**A-Sq**	0.17245239	**Pr > A-Sq**	>0.250

It is ok to assume that we are working with a normal distribution when we are looking at the female data.

On the other hand, the male data wields the following goodness of fit table:

Goodness-of-Fit Tests for Normal Distribution(Male)				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.20756734	**Pr > D**	0.030
Cramer-von Mises	**W-Sq**	0.13797690	**Pr > W-Sq**	0.032
Anderson-Darling	**A-Sq**	0.85972286	**Pr > A-Sq**	0.023

Now the p values are lower than our agreed alpha of 0.05, so we must reject the null hypothesis and say that the male bencpress values do not follow a normal distribution.

Finally if we want to find how the totality of our dataset behaves we should use the whole dataset by utilizing the following code:

```
TITLE “How Normal is the Benchpress Histogram?”;
Ods select Histogram ParameterEstimates GoodnessOfFit FitQuantiles Bins;
PROC UNIVARIATE DATA = Bench;
HISTOGRAM Benchpress/ normal(percents=20 40 60 80 midpercents);
INSET n normal(ksdpval) / pos = ne format =6.3;
RUN;
```

The table we get then is:

Goodness-of-Fit Tests for Normal Distribution(All)				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.15087615	**Pr > D**	0.072
Cramer-von Mises	**W-Sq**	0.22817637	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	1.45276277	**Pr > A-Sq**	<0.005

Now we have a problem. If we believe the first test we fail to reject the null, so our data could come from a normal. If we believe the other two we need to reject the null, so our data cannot come from a normal! (Statistics is complicated like that). In this class we will believe the KS test and stick with that. But, even so, the value of 0.072 is very close to 0.05 so we will be diplomatic in our answer:

“The data could come from a normal distribution, but there is a high chance that it does not.”

Hypothesis Testing with T-test

The column BenchpressNew contains the maximum benchpress of the athletes after they completed their 3 week program. What we want to see is if the average benchpress has changed significantly before and after the program for female athletes only. Our hypotheses are thus:

H_0: The average maximum Benchpress before and after the program has not changed for female athletes.
H_A: The average maximum Benchpress before and after the program has increased for female athletes.

We are justified to use the one sided alternative, since if there is a change that will almost certainly be towards a bigger values. There are a couple of ways that we can check the hypothesis above. One is to use that fact that the distribution of maximum benchpress for females before the program was almost normal, with a mean on μ=46.91.

We can ask if it is possible to see the new average of the column Benchpressnew by chance.

To do that lets use the following code:

```
TITLE “Has the average benchpress of female athletes changed?”;
Ods graphics on;
PROC TTEST DATA = BenchF h0=46.91 sides=u alpha=0.05;
var BenchpressNew;
RUN;
```

The command **`h0=46.91`** sets the null value we are testing it against.
The command **`sides=u`** lets sas know that we want only upper values, that is the alternative hypothesis is “greater than” and the test is one sided. If we want smaller values we use **`sides=l`**. If we want it to be two-sided we use **`sides =2`**
The command **`alpha=0.05`** sets the threshold for alpha.

Apart from the summary statistics and a 95% confidence interval, SAS outputs the following table:

DF	t Value	Pr > t
11	1.38	0.0973

This contains the degrees of freedom of our test (12-1=11) and the corresponding t value as well as the p value. In our case p is equal to 0.0973 which is greater than alpha and thus we fail to reject the null hypothesis. But this is not exactly what we should be looking at.

It is better if we compare the benchpresses before the training program with the corresponding ones after the program. Or in other words, test if the difference in values is significantly different than zero.

To do that we simply need to compare the corresponding variables Benchpress and BenchpressNew with the code:

```
TITLE “Is the difference before and after trivial for females?”;
PROC TTEST DATA = BenchF;
paired Benchpress* BenchpressNew;
RUN;
```

The command `paired Benchpress* BenchpressNew;`

The summary statistics presented are for the difference Bencpress-BenchpressNew and the table:

DF	t Value	Pr > |t|
11	-4.42	0.0010

Tells us that the probability of seeing those differences is 0.001 if you are using a 2 sided test, or 0.005 if you are using a one sided test. This is a more appropriate statistics to use because it compares each individual before and after the training. In this case we can safely reject the null hypothesis and say that the difference in the maximum benchpress before and after training is statistically significant for female athletes.

Out of all the outputs that TTEST gives you the paired profiles is a very useful ne, since it shows how the values changed for each observation by a line. The red line is the mean trend.

You can ask a similar question for males by using the code

```
TITLE “Is the difference before and after trivial for males?”;
PROC TTEST DATA = BenchM;
paired Benchpress* BenchpressNew;
RUN;
```

The change here is even more pronounced, but we must be careful since our distribution is a bit more skewed than the one of females.

Finally, SAS has an excellent analysis on the t-test procedure that you can find here:

https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#ttest_toc.htm