# SAS 2:

# Getting comfortable with your data

**University of Guelph**

*Revised June 2011*

# Table of Contents

# SAS Availability

Faculty, staff and students at the University of Guelph may access SAS three different ways:

1. **Library computers**

   On the library computers, SAS is installed on all machines.

2. **Acquire a copy for your own computer**

   If you are faculty, staff or a student at the University of Guelph, you may obtain the site-licensed standalone copy of SAS at a cost. However, it may only be used while you are employed or a registered student at the University of Guelph. To obtain a copy, go to the CCS Software Distribution Site (**www.uoguelph.ca/ccs/download**).

# Goals of the workshop

This workshop builds on the skills and knowledge develop in "Getting your data into SAS". Participants are expected to have basic SAS skills and statistical knowledge. Specific goals of this workshop:

- To review reading data into SAS datasets
- To learn how to determine whether your data comes from a Normal distribution
- How do we transform data if it is needed
- Plotting your data using SAS/GRAPH procedures

# Data for SAS sessions

**Dataset: Canadian Tobacco Use Monitoring Survey 2010 – Person File**

This survey tracks changes in smoking status, especially for populations most at risk such as the 15- to 24-year-olds. It allows Health Canada to estimate smoking prevalence for the 15- to 24-year-old and the 25-and-older groups by province and by gender on a semi-annual basis.

The sample data used for this series of SAS workshops only includes respondents from the province of Quebec and only 14 of a possible 202 variables are being used.

To view the data, open the Excel spreadsheet entitled CTUMS_2010.xls

| Variable Name | Label for Variable |
|---|---|
| **PUMFID** | Individual identification number |
| **PROV** | Province of the respondent |
| **DVURBAN** | Characteristic of the community |
| **HHSIZE** | Number of people in the household |
| **HS_Q20** | Number of people that smoke inside the house |
| **DVAGE** | Age of respondent |
| **SEX** | Respondent's sex |
| **DVMARST** | Grouped marital status of respondent |
| **PS_Q30** | Age smoked first cigarette |
| **PS_Q40** | Age begin smoking cigarettes daily |
| **WP_Q10A** | Number of cigarettes smoked – Monday |
| **WP_Q10B** | Number of cigarettes smoked – Tuesday |
| **WP_Q10C** | Number of cigarettes smoked – Wednesday |
| **WP_Q10D** | Number of cigarettes smoked – Thursday |
| **WP_Q10E** | Number of cigarettes smoked – Friday |
| **WP_Q10F** | Number of cigarettes smoked – Saturday |
| **WP_Q10G** | Number of cigarettes smoked – Sunday |
| **SC_Q100** | What was the main reason you began to smoke again? |
| **WTPP** | Person weight (survey weight variable) |

# Review – Getting data into SAS

## Temporary SAS datasets

When you create/save a dataset in SAS, whether it by using an infile statement, cards statement or the import procedure in SAS, by default SAS places it in the Work library. You will not see a physical SAS dataset file on your computer when you use the Work library – in other words the SAS dataset that is created is a Temporary SAS dataset.

## Permanent SAS datasets

There may be situations when you may need to read from a permanent SAS dataset or you may need to create a physical SAS dataset file. This will require the use of the SET and LIBNAME statement. The SET statement refers to the filename of the SAS data set already created or referred to in this SAS session. LIBNAME refers to a location where you would like to save your SAS data set.

```
libname sasdata "C:\Users\edwardsm\Documents\Michelle_Docs\Workshops";


Data sasdata.ctums2;
  set ctums;
Run;
```

The dataset name will now need to use its full name which includes its library name – sasdata in this case.

The Set statement tells SAS to use the dataset we've already created and stored in the Work library of SAS

The libname statement lets SAS know WHERE you want to save the SAS dataset. In this example a directory called Workshop in a Michelle_Docs directory

Now we will have 2 SAS datasets – one in the Work Library called ctums and a second called ctums2 in the SASDATA library which is located in the C:\Users\edwardsm\Documents\Michelle_Docs\Workshops directory. If you look in the specified directory you should now see a file called ctums2.sas7bdat. This is referred to as a permanent SAS dataset – there is now a physical file that we can see and send to colleagues.

When someone sends you a *.sas7bdat file how do you read it?

1. You will need to create a Library – by using the libname statement. The name of the Library can be anything you'd like it to be. The name of the library is only used to create a location on your computer and to refer to it in your program.
2. By using a Data step – create either a local copy (Work library) or save it as a different name.

*Example:*
I have just received the file ctums2.sas7bdat in my email. I will save it in a new directory called C:\Users\edwardsm\Documents\Michelle_Docs\Research

```
libname newdata "C:\Users\edwardsm\Documents\Michelle_Docs\Research";

Data newctums;
  set newdata.ctums2;
Run;
```

My permanent SAS dataset ctums2.sas7bdat is located in the Research directory on my computer. I create the library called newdata which refers to that directory.

I save a new copy called newctums in my Work library.

# Statistical Refresher

## Types of Statistics

Two broad types of statistics exists which are descriptive and inferential. Descriptive statistics describe the basic characteristics of the data in a study. Usually generated through an Exploratory Data Analysis (EDA), they provide simple numerical and graphical summaries about the sample and the measures. Inferential statistics allows you to make conclusions regarding the data ie significant differences, relationships between variables, etc.

Here are some examples of descriptive and inferential statistics:

| **Descriptives** | **Inferential** |
|---|---|
| • Frequencies<br>• Means<br>• Standard Deviations<br>• Ranges<br>• Medians<br>• Modes | • T-tests<br>• Chi-squares<br>• ANOVA<br>• Friedman |

Which test to perform on your data largely depends on a number of factors including:

1. What type of data you are working with?
2. Are you samples related or independent?
3. How many samples are you comparing?

## Types of Variables

Variable types can be distinguished by various levels of measurement which are Nominal, Ordinal, Interval or Ratio.

### Nominal

Have data values that identify group membership. The only comparisons that can be made between variable values are equality and inequality. Examples of nominal measurement include gender, race, religious affiliation, telephone area codes or country of residence.

### Ordinal

Have data values arranged in a rank ordering with an unknown difference between adjacent values. Comparisons of greater and less can be made and in addition to equality and inequality. Examples include: results of a horse race, level of education or satisfaction/attitude questions.

### Interval

Are measured on a scale such that a one-unit change represents the same difference throughout the scale. These variables do not have true zero points. Examples include: temperature in the Celsius or Fahrenheit scale, year date in a calendar or IQ test results.

### Ratio

Have the same properties as interval variables plus the additional property of a true zero. Examples include: temperature measured in Kelvins, most physical quantities such as mass, length or energy, age, length of residence in a given place.

Interval and Ratio will be considered identical thus yielding three types of measurement scales.

## Appropriate Statistics

For each type of variable a particular measure of central tendency is most appropriate. By central tendency we mean one value that most effectively summarizes a variable's complete distribution.

| Measurement Scale | Measure of Central Tendency |
|---|---|
| Nominal | Mode – value that appears the most often in distribution. |
| Ordinal | Median – Value that divides the ordered distribution of responses into two equal size groups. (the value of the $50^{th}$ percentile) |
| Interval / Scale | Mean – The arithmetic average of a distribution. |

# Frequency

**How many males and females are in this dataset?**

**SAS Code:**

```
Proc freq data=newctums;
  tables sex;
Run;
```

**Tables** statement – list the variables you would like to see a frequency chart created.

Good SAS code writing etiquette – use a **data=** option in your **Proc** statement – this ensures that SAS uses the correct dataset in your analysis.

Ensure that the procedure is closed with a **Run;**

**SAS Output:**

The FREQ Procedure

SEX

| SEX | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 1 | 432 | 46.96 | 432 | 46.96 |
| 2 | 488 | 53.04 | 920 | 100.00 |

Based on the results we see that in our sample of CTUMS2010 47% of the sample are males and 53% are females.

Let's look at age_group – what is the frequency distribution for the age_group variable?

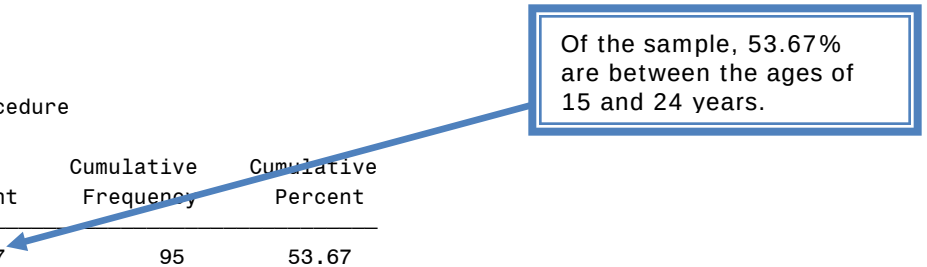
**SAS Code:**

```
Proc freq data=newctums;
  tables agegroup;
Run;
```

**SAS Output:**

The FREQ Procedure

| agegroup | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 15-24 years | 95 | 53.67 | 95 | 53.67 |
| 25-34 years | 15 | 8.47 | 110 | 62.15 |
| 35-44 years | 20 | 11.30 | 130 | 73.45 |
| 45-54 years | 18 | 10.17 | 148 | 83.62 |
| 55-64 years | 24 | 13.56 | 172 | 97.18 |
| 65-74 years | 4 | 2.26 | 176 | 99.44 |
| 75-84 years | 1 | 0.56 | 177 | 100.00 |



Now let's put the two tables together and create a Cross-tabulation to show us the age distribution of the 2 genders. To accomplish this task we will list the 2 variables of interest in the **Tables** statement and place an '*****' between the two to let SAS know that we want a crosstab.

**SAS Code:**

```
Proc freq data=newctums;
  tables agegroup*sex;
Run;
```

Note: The order of the variables in your **Tables** statement determine the structure of the table.

Row variable * column variable

**SAS Output:**

The FREQ Procedure
Table of agegroup by SEX

| agegroup | SEX(SEX) | | |
|---|---|---|---|
| Frequency<br>Percent<br>Row Pct<br>Col Pct | Male | Female | Total |
| 15-24 years | 47<br>26.55<br>49.47<br>52.22 | 48<br>27.12<br>50.53<br>55.17 | 95<br>53.67 |
| 25-34 years | 9<br>5.08<br>60.00<br>10.00 | 6<br>3.39<br>40.00<br>6.90 | 15<br>8.47 |
| 35-44 years | 7<br>3.95<br>35.00<br>7.78 | 13<br>7.34<br>65.00<br>14.94 | 20<br>11.30 |

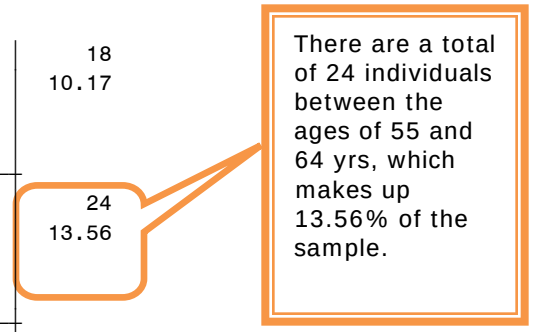
There are 6 females between the ages of 25 and 34 yrs.

3.39 % of this sample are females between the ages of 25 and 34 yrs.

Of all the individuals between 25 and 34 years, 40 % are female.

Of all the females, 6.90 % are between the ages of 25 and 34 yrs.

| | | | |
|---|---|---|---|
| 45-54 years | 10 | 8 | 18 |
| | 5.65 | 4.52 | 10.17 |
| | 55.56 | 44.44 | |
| | 11.11 | 9.20 | |
| 55-64 years | 13 | 11 | 24 |
| | 7.34 | 6.21 | 13.56 |
| | 54.17 | 45.83 | |
| | 14.44 | 12.64 | |
| 65-74 years | 3 | 1 | 4 |
| | 1.69 | 0.56 | 2.26 |
| | 75.00 | 25.00 | |
| | 3.33 | 1.15 | |
| 75-84 years | 1 | 0 | 1 |
| | 0.56 | 0.00 | 0.56 |
| | 100.00 | 0.00 | |
| | 1.11 | 0.00 | |
| Total | 90 | 87 | 177 |
| | 50.85 | 49.15 | 100.00 |




**Chi-square tests**

If we want to test whether a relationship exists between 2 categorical variables, a Chi-square test is one option. To conduct a Chi-square test in SAS we will add an option to the above coding.

```
Proc freq data=newctums;
  tables agegroup*sex /chisq;
Run;
```

The FREQ Procedure

Table of agegroup by SEX

agegroup SEX(SEX)

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Male | Female | Total |
|---|---|---|---|
| 55-64 years | 13 | 11 | 24 |
| | 7.34 | 6.21 | 13.56 |
| | 54.17 | 45.83 | |
| | 14.44 | 12.64 | |
| 65-74 years | 3 | 1 | 4 |
| | 1.69 | 0.56 | 2.26 |
| | 75.00 | 25.00 | |
| | 3.33 | 1.15 | |
| 75-84 years | 1 | 0 | 1 |
| | 0.56 | 0.00 | 0.56 |
| | 100.00 | 0.00 | |
| | 1.11 | 0.00 | |
| Total | 90 | 87 | 177 |
| | 50.85 | 49.15 | 100.00 |

Statistics for Table of agegroup by SEX

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 6 | 4.7499 | 0.5763 |
| Likelihood Ratio Chi-Square | 6 | 5.2141 | 0.5167 |
| Mantel-Haenszel Chi-Square | 1 | 0.6105 | 0.4346 |
| Phi Coefficient | | 0.1638 | |
| Contingency Coefficient | | 0.1617 | |
| Cramer's V | | 0.1638 | |

WARNING: 29% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Sample Size = 177

Note the Warning!!! With cells of 5 or less the Chi-square test may not be a valid test. Think about ways to recode your data ie. Create new groupings or re-examine your choice of statistical test

```
Proc freq data=newctums;
  tables dvmarst*sex /chisq;
Run;
```

Table of DVMARST by SEX

DVMARST(DVMARST) SEX(SEX)

| Frequency<br>Percent<br>Row Pct<br>Col Pct | Male | Female | Total |
|---|---|---|---|
| Common-law/Married | 31<br>17.61<br>51.67<br>34.83 | 29<br>16.48<br>48.33<br>33.33 | 60<br>34.09 |
| Widow/Divorced/Separated | 5<br>2.84<br>38.46<br>5.62 | 8<br>4.55<br>61.54<br>9.20 | 13<br>7.39 |
| Single | 53<br>30.11<br>51.46<br>59.55 | 50<br>28.41<br>48.54<br>57.47 | 103<br>58.52 |
| Total | 89<br>50.57 | 87<br>49.43 | 176<br>100.00 |

Statistics for Table of DVMARST by SEX

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 0.8237 | 0.6624 |
| Likelihood Ratio Chi-Square | 2 | 0.8299 | 0.6604 |
| Mantel-Haenszel Chi-Square | 1 | 0.0017 | 0.9671 |
| Phi Coefficient | | 0.0684 | |
| Contingency Coefficient | | 0.0683 | |
| Cramer's V | | 0.0684 | |

Sample Size = 176

The non-significant chi-square suggests that there is no association between marital status and sex in this sample.

# Exercise 1

**Exercise:** Is there an association between sex (***sex***) and whether the individual lived in an urban or rual area(***dvurban***)?

# Mean, mode and median

We've seen earlier that to determine the measures of central tendency – mean, median and mode are the best statistics. All three measures are available in **Proc UNIVARIATE** (see next section), but **Proc MEANS** also offers the mean and median.

## Mean

We will calculate the Mean for total number of cigarettes smoked in a week (***totcig***)

**SAS Code:**

```
Proc means data=newctums;
  var totcig;
Run;
```

**SAS Output:**

The MEANS Procedure

Analysis Variable : totcig

| N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|
| 176 | 108.1079545 | 137.3571143 | 0 | 691.0000000 |

**How do you get the standard error in this output?**

## Median

We will calculate the Median of SC_Q100 (What was the main reason you began to smoke again?)

**SAS Code:**

```
Proc means data=newctums median;
  var sc_q100;
Run;
```

**SAS Output:**

The MEANS Procedure

Analysis Variable : SC_Q100 SC_Q100

| Median |
|---|
| 4.0000000 |

## Mode

**Proc UNIVARIATE** is the only procedure that will calculate the Mode of a variable. The mode is defined as the value with the most observations – so you can accomplish with **Proc FREQ** as well.

**SAS Code (Proc UNIVARIATE):**

```
Proc univariate data=newctums;
  var sc_q100;
Run;
```

**SAS Output:**

The UNIVARIATE Procedure
Variable: SC_Q100 (SC_Q100)

Moments

| | | | |
|---|---|---|---|
| N | 72 | Sum Weights | 72 |
| Mean | 4.90277778 | Sum Observations | 353 |
| Std Deviation | 3.00778203 | Variance | 9.04675274 |
| Skewness | 0.53048708 | Kurtosis | -1.3134182 |
| Uncorrected SS | 2373 | Corrected SS | 642.319444 |
| Coeff Variation | 61.3485287 | Std Error Mean | 0.35447051 |

Basic Statistical Measures

| Location | | Variability | |
|---|---|---|---|
| Mean | 4.902778 | Std Deviation | 3.00778 |
| Median | 4.000000 | Variance | 9.04675 |
| Mode | 2.000000 | Range | 9.00000 |
| | | Interquartile Range | 6.00000 |

Tests for Location: Mu0=0

| Test | -Statistic- | | -----p Value------ | |
|---|---|---|---|---|
| Student's t | t | 13.83127 | Pr > \|t\| | <.0001 |
| Sign | M | 36 | Pr >= \|M\| | <.0001 |
| Signed Rank | S | 1314 | Pr >= \|S\| | <.0001 |

## Normality

One assumption we see many tests demanding is that the dependent variable (if it is a continuous variable) comes from a normal distribution. The "bell curve" is often used as an example of a normal distribution. How do we go about testing this assumption in SAS?

One powerful **Proc**edure to provide you with descriptive statistics along with normality tests in **UNIVARIATE.** To view all the syntax and options for this procedure please visit:

http://support.sas.com/onlinedoc/913/getDoc/en/procstat.hlp/univariate_index.htm#idxuni0002

Let's take a look at our "totcig" variable – this is the variable we created in SAS 1 that was the sum of cigarettes smoked during the week.

Let's look at the default output for **Proc UNIVARIATE** first:

**SAS Code:**

```
Proc univariate data=newctums;
  var totcig;
Run;
```

**SAS Output:**

The UNIVARIATE Procedure
Variable: totcig

Moments

| | | | |
|---|---|---|---|
| N | 176 | Sum Weights | 176 |
| Mean | 108.107955 | Sum Observations | 19027 |
| Std Deviation | 137.357114 | Variance | 18866.9769 |
| Skewness | 2.92438367 | Kurtosis | 9.10766212 |
| Uncorrected SS | 5358691 | Corrected SS | 3301720.95 |
| Coeff Variation | 127.055511 | Std Error Mean | 10.3536821 |

The moments provide you with a lot of descriptive statistics – from the Mean to Stderr to Kurtosis.

Basic Statistical Measures

| Location | | Variability | |
|---|---|---|---|
| Mean | 108.1080 | Std Deviation | 137.35711 |
| Median | 70.0000 | Variance | 18867 |
| Mode | 84.0000 | Range | 691.00000 |

```
                    Interquartile Range    105.00000

            Tests for Location: Mu0=0

Test           -Statistic-    -----p Value------

Student's t    t   10.4415    Pr > |t|    <.0001
Sign           M        85    Pr >= |M|   <.0001
Signed Rank    S    7267.5    Pr >= |S|   <.0001

             Quantiles (Definition 5)

              Quantile      Estimate

              100% Max           691
              99%                691
              95%                497
              90%                185
              75% Q3             140
              50% Median          70
              25% Q1              35
              10%                  3
              5%                   1
              1%                   0
              0% Min               0

               Extreme Observations

      ----Lowest----        ----Highest---

      Value      Obs        Value      Obs

          0      174          608      111
          0      155          691       66
          0      118          691       77
          0       85          691      107
          0       71          691      136
```

Three test for location test the hypothesis that the sample mean is equal to 0

A listing of extreme observations is also provided.

By reviewing this output, the researcher now has a very good picture of their data. However, let's add two options, one to provide plots (**plot**) and a second to provide Normality test statistics (**normal**).

**SAS Code:**

```
Proc univariate data=newctums normal plot;
  var totcig;
Run;
```

**SAS Output:**

The UNIVARIATE Procedure
Variable: totcig

Moments

| | | | |
|---|---|---|---|
| N | 176 | Sum Weights | 176 |
| Mean | 108.107955 | Sum Observations | 19027 |
| Std Deviation | 137.357114 | Variance | 18866.9769 |
| Skewness | 2.92438367 | Kurtosis | 9.10766212 |
| Uncorrected SS | 5358691 | Corrected SS | 3301720.95 |
| Coeff Variation | 127.055511 | Std Error Mean | 10.3536821 |

Basic Statistical Measures

| Location | | Variability | |
|---|---|---|---|
| Mean | 108.1080 | Std Deviation | 137.35711 |
| Median | 70.0000 | Variance | 18867 |
| Mode | 84.0000 | Range | 691.00000 |
| | | Interquartile Range | 105.00000 |

Tests for Location: Mu0=0

| Test | -Statistic- | | -----p Value------ | |
|---|---|---|---|---|
| Student's t | t | 10.4415 | Pr > \|t\| | <.0001 |
| Sign | M | 85 | Pr >= \|M\| | <.0001 |
| Signed Rank | S | 7267.5 | Pr >= \|S\| | <.0001 |

The first sections of the output match what we have seen above. With one exception:

Tests for Normality:

Tests for Normality

| Test | | --Statistic--- | | -----p Value------ |
|---|---|---|---|---|
| Shapiro-Wilk | W | 0.641166 | Pr < W | <0.0001 |
| Kolmogorov-Smirnov | D | 0.215624 | Pr > D | <0.0100 |
| Cramer-von Mises | W-Sq | 2.924821 | Pr > W-Sq | <0.0050 |
| Anderson-Darling | A-Sq | 17.57121 | Pr > A-Sq | <0.0050 |

Quantiles (Definition 5)

| Quantile | Estimate |
|---|---|
| 100% Max | 691 |
| 99% | 691 |
| 95% | 497 |
| 90% | 185 |
| 75% Q3 | 140 |
| 50% Median | 70 |
| 25% Q1 | 35 |

Extreme Observations

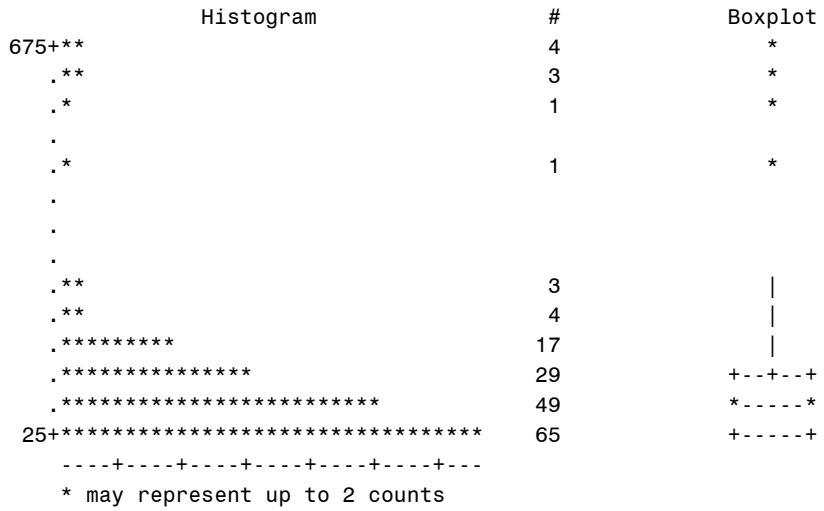
| ----Lowest---- | | ----Highest--- | |
|---|---|---|---|
| Value | Obs | Value | Obs |
| 0 | 174 | 608 | 111 |
| 0 | 155 | 691 | 66 |
| 0 | 118 | 691 | 77 |
| 0 | 85 | 691 | 107 |
| 0 | 71 | 691 | 136 |

Tests for Normality:

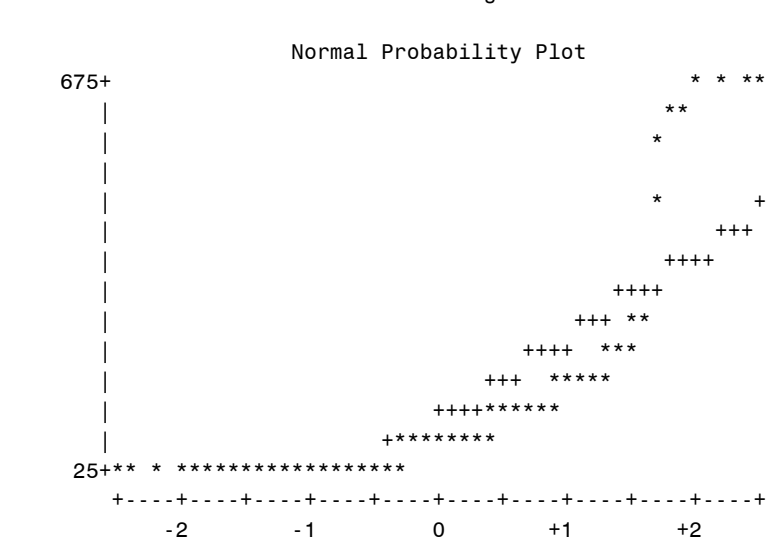
The Null hypothesis for this test is – your dependent variable comes from a Normal distribution. So if the p-value < 0.05 then you reject the Null hypothesis and your dependent variable is not normal.

Four tests are provided.

Shapiro-Wilk should only be used if the sample size is less than or equal to 2000.




Histogram and Boxplot are results of the **plot** option in **Proc UNIVARIATE.** A great visualization tool for the Normality test statistics.




Another plot produced by the **plot** option. The sample data points are represented by the '*'. If the data were normally distributed the '*' would cover the '+' of the normal probability plot.

Based on the graphs and the tests you have available to you - would you state that this variable is normal or not? How would you deal with this type of variable if you were doing any further statistical tests?

# Transformations

What happens when your data is not normally distributed as the above example? The first step is to try a transformation. Attempt to transform your data so it follows a normal distribution.

## Common Transformations:

1. Square root – commonly used for counts (**sqrt**)
2. Arcsin - commonly used for proportions (**arsin**)
3. Logarithmic transformation – log base 10 or natural log (**log10** or **log**)

For our data we will try the logarithmic transformation. Any data manipulations must be conducted in a **Data** step.

**SAS Code:**

```
Data t_ctums;
  set newctums;
  log_totcig = log(1+totcig);
Run;
```

**Data –** saving the new dataset under a new name "t_ctums" – think of this as the File-SaveAs – option in Excel

**Set -** telling SAS to use the dataset already in the program – same function as if we were reading in a Permanent SAS dataset.

**Log_totcig -** the name of a new variable we are going to create

**Log(totcig) –** totcig is the variable in our dataset that we want to transform – and log is the transformation function we want to use.

**Sample output:**

| Obs | totcig | log_ totcig |
|---|---|---|
| 1 | 2 | 1.09861 |
| 2 | 105 | 4.66344 |
| 3 | 140 | 4.94876 |
| 4 | 84 | 4.44265 |

# Exercise 2:

a) Rerun the Normality test to determine whether the new variable you created, log_totcig, is indeed normal.

b) What variable will you use to run your statistical tests? totcig or log_totcig?

c) How will you report your results?