



SPSS DATA SCREENING

TUTORIAL PREPARED BY:
Ashley Brookshier and Rielly Boyd
College of Education Research Center

PURPOSE OF TUTORIAL

The purpose of this tutorial is to familiarize researchers with data screening and to apply methods to prepare data for univariate analyses. As Tabachnick and Fidell (2007) point out, data screening is critical to protect the integrity of inferential statistics.

TUTORIAL TOPICS

1. Introductory Comments
2. Detecting Erroneous Data Entry Errors
3. Identifying and Dealing with Missing Data
4. Detecting and Making Decisions about Univariate Outliers
5. Screening for and Making Decisions about Univariate Outliers
6. Transformation of the Dependent Variable

This tutorial is an update from the *SPSS Data Screening Workshop* presented by Robert A. Horn, Ph.D. (Associate Professor, Educational Psychology) and Amy Prosser (Faculty Research Center's Research Assistant).

1. INTRODUCTORY COMMENTS

This tutorial deals with the issues that are resolved after the data have been collected – but before the main data analyses are performed. Careful consideration of data screening and assumption testing can in fact be very time consuming and sometimes very tedious. It is not uncommon to spend several days making careful examinations of the data prior to actually running the main statistical analyses, which can themselves only take a few minutes. Careful consideration and necessary resolutions of any issues before the main analyses is fundamental to an honest analysis of the data, which in turn protects the integrity of the inferential statistic (Tabachnick and Fidell, 2007).

The data set (FRC-SCREEN) we will be using is part of a larger data set from Osterlind and Tabachnick (2001). The study involved conducting structured interviews that focused on “a variety of health, demographic, and attitudinal measures, given to a randomly selected group of 465 female, 20 to 59-year old, English-speaking residents of the San Fernando Valley, a suburb of Los Angeles, in February 1975” (Tabachnick & Fidell, 2007, p. 934).

2. DETECTING ERRONEOUS DATA ENTRY ERRORS

Obviously, the integrity of your data analyses can be significantly compromised by entering wrong data. It is most ideal to enter and check your own data comparing the original data to the data in the Data View.

When someone else is entering your data, you need to train and trust them as well as monitor their data entry. If you cannot check all of their data entry then, at a minimum, check a random subset of the data set.

Additionally, use the procedure *Analyze > Descriptive Statistics > Frequencies*. From the output you will be answering questions such as:

- Are all the variable scores in expected range?
- Are means and standard deviations plausible?

2.1. Let's look at an example using the directions below.

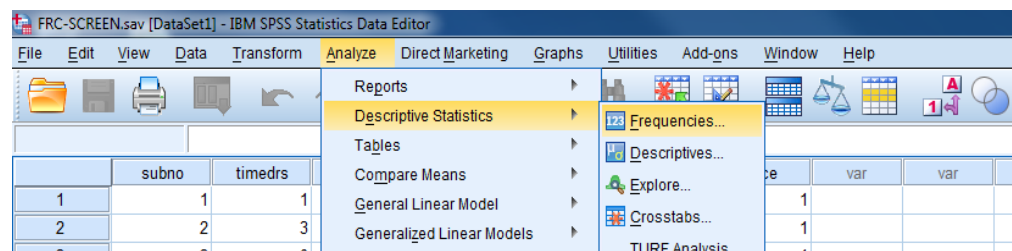
2.1.1. Open (double click) the data set called **FRC-SCREEN**. This data set is modified from the one used by Osterlind and Tabachnick (2001).

2.1.2. While in Data View go to the top of the screen and initiate

> *Analyze*

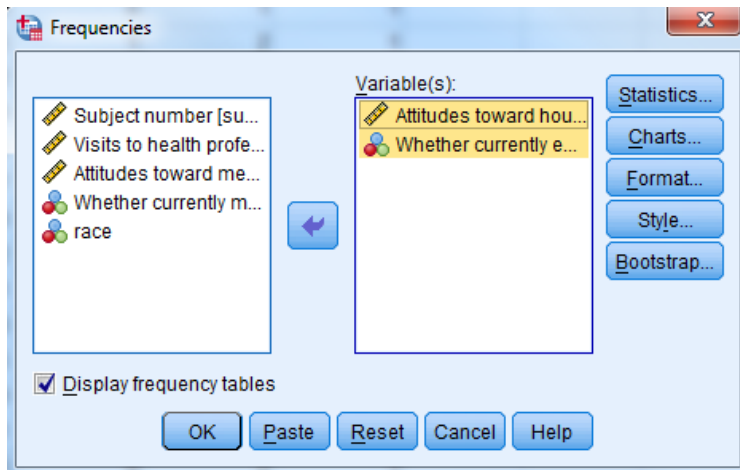
> *Descriptive Statistics*

> *Frequencies*



2.1.3. Highlight “Attitudes toward housework (**atthhouse**)” and “Whether currently employed (**emplmnt**)”

2.1.4. Click over both variables to Variable(s): list



2.1.5. Click **OK**

2.1.6. Review the frequency tables for each variable. It should be evident that the value **331** in the frequency table for “Attitudes toward housework (**atthhouse**)” and the value of **11** in the frequency table for “Whether currently employed (**emplmnt**)” have dubious accuracy. Both values were entered incorrectly.

32	4	.9	.9	98.1
33	4	.9	.9	98.9
34	2	.4	.4	99.4
35	2	.4	.4	99.8
331	1	.2	.2	100.0
Total	464	99.8	100.0	
Missing System	1	.2		
Total	465	100.0		

Whether currently employed				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 0 PAIDWORK	246	52.9	52.9	52.9
1 HOUSEWFE	218	46.9	46.9	99.8
11	1	.2	.2	100.0
Total	465	100.0	100.0	

2.1.7. Click out of the frequency table output (don’t save the output unless you want to) and go to **Data View** and correct the two incorrect values, accordingly:

<u>SubNo</u>	<u>Variable</u>	<u>Incorrect Value</u>	<u>New Value</u>
26	“Attitudes toward housework (atthhouse)”	331	31
28	“Whether currently employed (emplmnt)”	11	1

1 : subno

	subno	timedrs	attdrug	athouse	emplmt	mstatus	race	var
1	1	1	8	27	1	2	1	
2	2	3	7	20	0	2	1	
3	3	0	8	23	0	2	1	
4	4	13	9	28	1	2	1	
5	5	15	7	24	1	2	1	
6	6	3	8	25	0	2	1	
7	7	2	7	30	1	2	1	
8	8	0	7	24	1	2	1	
9	9	7	7	20	1	2	1	
10	10	4	8	30	0	1	1	
11	11	15	9	15	1	2	1	
12	12	0	6	22	1	2	1	
13	13	2	6	19	1	2	1	
14	14	13	8	25	1	2	1	
15	15	2	5	17	1	2	1	
16	16	2	8	19	0	2	2	
17	21	1	8	22	1	2	1	
18	22	2	6	21	0	1	1	
19	23	5	8	28	1	2	1	
20	24	5	10	25	0	2	1	
21	25	3	6	19	0	2	1	
22	26	4	5	331	0	2	1	
23	27	2	8	25	0	2	1	
24	28	0	8	26	11	2	1	
25	29	13	9	26	0	2	1	
26	30	7	9	33	0	2	1	
27	31	2	8	20	1	2	1	
28	32	12	9	26	1	2	1	

3. IDENTIFYING AND DEALING WITH MISSING DATA

Missing values occur when participants in a study do not respond to some items or sections of items, participant attrition, and data management mistakes, etc. Stevens (2002) states, “Probably the “best” solution is to make every attempt to minimize the problem before and during the study, rather than having to manufacture data” (p. 33).

According to Tabachnick and Fidell (2007), **concern** is with:

- *Pattern of missing data* – if data is scattered randomly through a data set then it is less of a serious problem. However, nonrandom missing values make it difficult to generalize findings.
- *How much data is missing* – if 5% or less follows a random pattern and is missing from a large data set then the problem is less serious and most strategies for dealing with missing values will work.

Tabachnick and Fidell (2007) identified the following alternatives to handling missing data:

- *Delete cases or variables*
 - *Less of a problem* if only a few cases have missing values.
 - *Less of a problem* if missing values are concentrated to a few variables not critical to analysis or highly correlated with other variables.
 - *More of a problem* if lose a great deal of participants.
 - *More of a problem* if missing values are not randomly distributed, case or variable deletion can distort findings and generalizability.

- **Estimating missing values** – estimate (impute) missing values and then use estimates in the data analysis. Estimation methods include:
 - **Prior knowledge** – replace missing value with a value reflecting researcher judgment. This might include estimating the value that may have been a median or downgrade a continuous variable to a dichotomous variable (high, low) and estimate which category the missing value would fall into. This results in a loss of information regarding the variable.
 - **Mean substitution** – the mean is a *good estimate* about the value of a variable. It is a conservative option and it results in a loss of variance (since it becomes a constant). A less conservative approach, but not as liberal as prior knowledge, is to use a group mean as the estimate instead of a total sample mean.
 - **Regression** – cases with complete data generate a regression equation that is used to predict missing values.
 - Two newer and more sophisticated methods being used more often are expectation maximization (EM) and multiple imputations. These two methods are available as an add-on program to SPSS.
- **Repeating analyses with and without missing data** is highly recommended. If results are similar then you have confidence in the missing data procedure. If results are different then conduct further investigation, evaluate which results most closely reflect “reality” or report both sets.

3.1. The following example will illustrate how to detect a missing value in a variable and replace it with the mean of the variable (before the missing variable is replaced).

3.1.1. Initiate

> **Analyze**

> **Descriptive Statistics**

> **Frequencies**

3.1.2. Click **Reset**

3.1.3. Click over “Attitudes toward medication (**attdrug**),” “Attitudes toward housework (**atthouse**),” and “Whether currently married (**mstatus**)” under the Variable(s): list

3.1.4. Click **OK**

3.2. From the “Statistics” table, you will see that Attitudes toward housework (**atthouse**)” has one missing value.

Frequencies

		Statistics		
		Attitudes toward medication	Attitudes toward housework	Whether currently married
N	Valid	465	464	465
	Missing	0	1	0

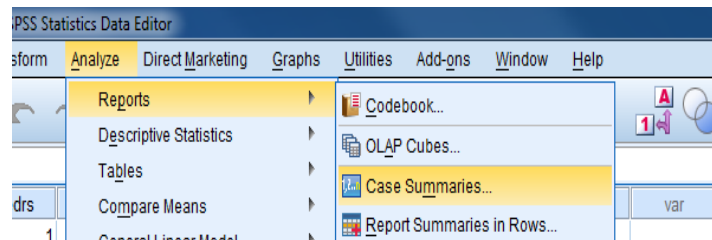
To find the case number (row) with the missing value, click out of the frequencies output

3.2.1. Initiate

> *Analyze*

> *Reports*

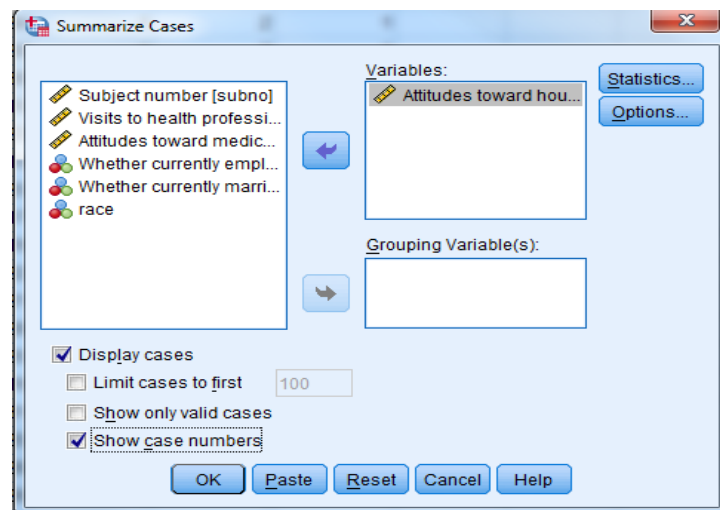
> *Case Summaries*



3.2.2 Click over the variable “Attitudes toward housework (**atthouse**)” under Variables: list.

3.2.3. Click off *Limit cases for first 100*, and *Show only valid cases*

3.2.4. Click on *Show case numbers*



3.2.5. Click *OK*

3.3. Scroll down the **Case Summaries** table until you find the missing value identification and make note of the case number (row). The row number is 253 for the missing value.

Active Dataset	250	250	23
Case Processing Summary	251	251	26
Case Summaries	252	252	22
	(missing) 253	253	
	253	254	25
	254	255	19

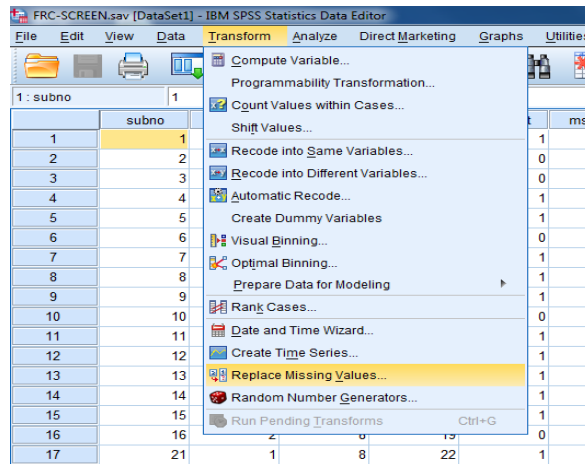
- 3.4. Click out of the output and go to Data View to find the “**subno**” for row 253 under the variable “**atthouse**”. The “**subno**” for the missing value is 338.

To replace the missing value for “**subno**” 338 on the variable “**atthouse**”

3.4.1. Initiate

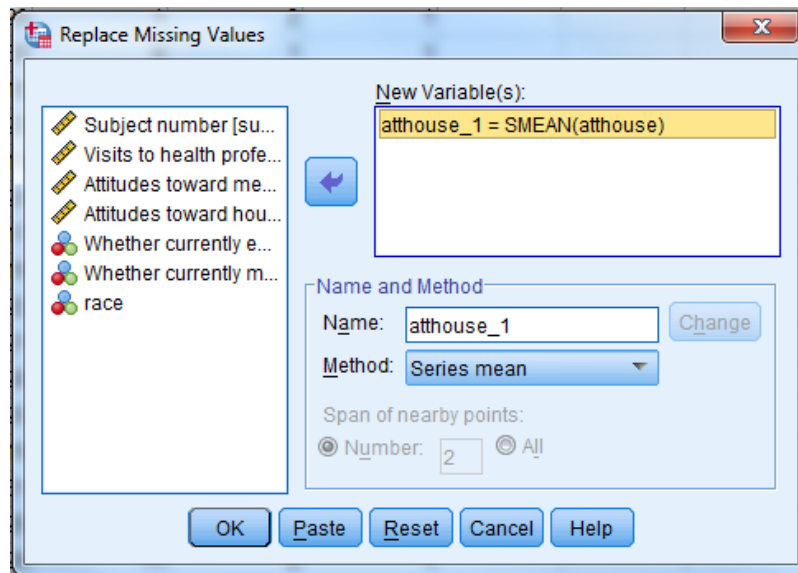
> *Transform*

> *Replace Missing Values*



- 3.4.2. Click over “Attitudes toward housework (**atthouse**)” under the New Variable(s): list

- 3.4.3. Choose *Series Mean* beside Method: (this should be the default)



- 3.4.4. Click *OK*

- 3.5. Click out of the output produced and see that a new column (variable) was created called “**atthouse_1**” as the last column on Data View

1: subno 1

	subno	timedrs	attdrug	atthouse	emplmnt	mstatus	race	atthouse_1	var
1	1	1	8	27	1	2	1	22.0	
2	2	3	7	20	0	2	1	20.0	
3	3	0	8	23	0	2	1	23.0	

3.6. Scroll down the Data View spreadsheet until you find row 253 (“subno” 338). Notice that the cell under “atthouse_1” is 23.5 where the original “atthouse” is a missing value. This new variable “atthouse_1” with the mean replacement can be used for subsequent analyses rather than the original “atthouse” with the missing value.

251	336	5	1	20	0	2	1	20.0
252	337	2	9	22	0	2	1	22.0
253	338	2	6	.	0	1	2	23.5
254	339	4	7	25	1	2	2	25.0
255	340	2	7	19	1	2	2	19.0

Make note that you had choices other than the mean to replace the missing value. For example, one of the choices is *Linear trend at point* in which regression is used to predict the missing value

Next, to demonstrate how to screen for univariate outliers and assumptions and correct for problems, we are going to prepare our data to answer one question: ***Is there a significant difference between married and not married women in the number of visits they make to health professionals.*** We will be testing the null hypothesis that there will be no differences between married and not married women in the number of visits they make to health professionals ($H_0: \mu_1 = \mu_2$). We will eventually run a one way ANOVA (F -test) to see if there are significant differences (we could have also run an independent t -test). The variables used for this analysis have been screened and corrected for data entry problems and missing data. We will continue the data screening process to prepare our data for analysis to protect the integrity of the inferential statistical test (F -test).

4. DETECTING AND MAKING DECISIONS ABOUT UNIVARIATE OUTLIERS

Many statistical methods are sensitive to outliers so it is important to identify outliers and make decisions about what to do with them. The reason according to Stevens (2002) is, “Because we want the results of our statistical analysis to reflect most of the data, and not to be highly influenced by just one or two errant data points” (p. 13). Results do not generalize except to another sample with similar outliers.

Reasons for outliers (Tabachnick & Fidell, 2007):

- incorrect data entry
- failure to specify missing values in computer syntax so missing values are read as real data
- outlier is not a member of population that you intended to sample
- outlier is representative of population you intended to sample but population has more extreme scores than a normal distribution

Detecting univariate and multivariate outliers.

- Univariate outlier for dichotomous variables *90-10 split between categories*.
- Univariate outlier for continuous variables in excess of $z = \pm 3.29$ ($p < .001$, two-tailed test) (Tabachnick & Fidell, 2007, p. 73). Although not as precise, one can also look at histograms, box plots, and normal probability plots.

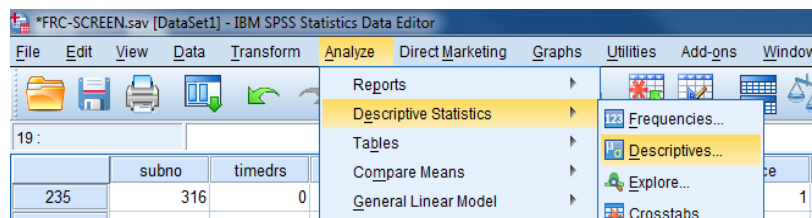
4.1. To obtain z scores (standard scores) to detect univariate outliers for cases under a variable

4.1.1. Initiate

> *Analyze*

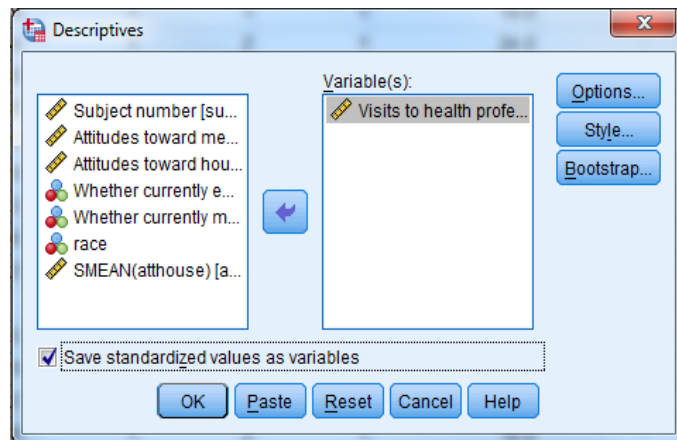
> *Descriptive Statistics*

> *Descriptives*



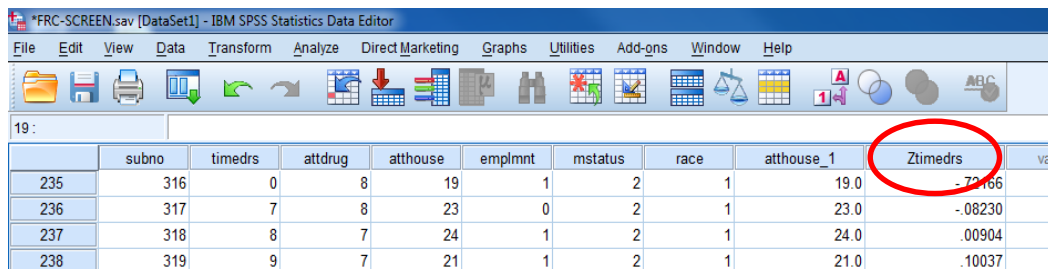
4.1.2. Click over “Visits to health professional (**timedsr**)” under Variable(s): list

4.1.3. Click on the box at the bottom of the screen that says *Save standardized values as variables*



4.1.4. Click *OK*

- 4.2. Click out of the “Descriptives” output and see the new variable at the end of the Data View spreadsheet called “**Ztimedrs**”



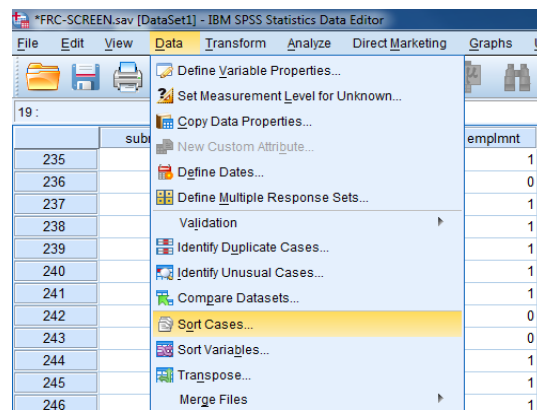
	subno	timedrs	attdrug	atthouse	emplmnt	mstatus	race	atthouse_1	Ztimedrs	var
235	316	0	8	19	1	2	1	19.0	-72.66	
236	317	7	8	23	0	2	1	23.0	-.08230	
237	318	8	7	24	1	2	1	24.0	.00904	
238	319	9	7	21	1	2	1	21.0	.10037	

- 4.3. Our task is to identify any value that is greater than ± 3.29 in each cell under “**Ztimedrs**”
- 4.4. Since there are so many cases, let’s sort to put the most extreme scores at the top and bottom of the “**Ztimedrs**” column

4.4.1. Initiate

> *Data*

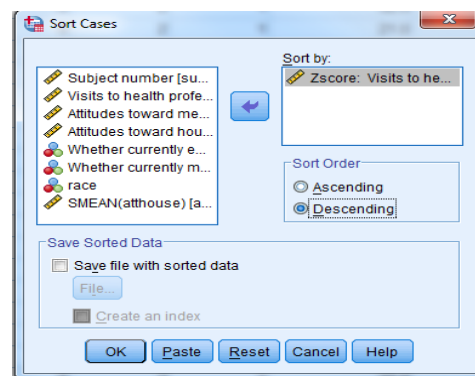
> *Sort Cases*



- 4.4.2. Highlight “Zscore: Visits to health professionals (**Ztimedrs**)”

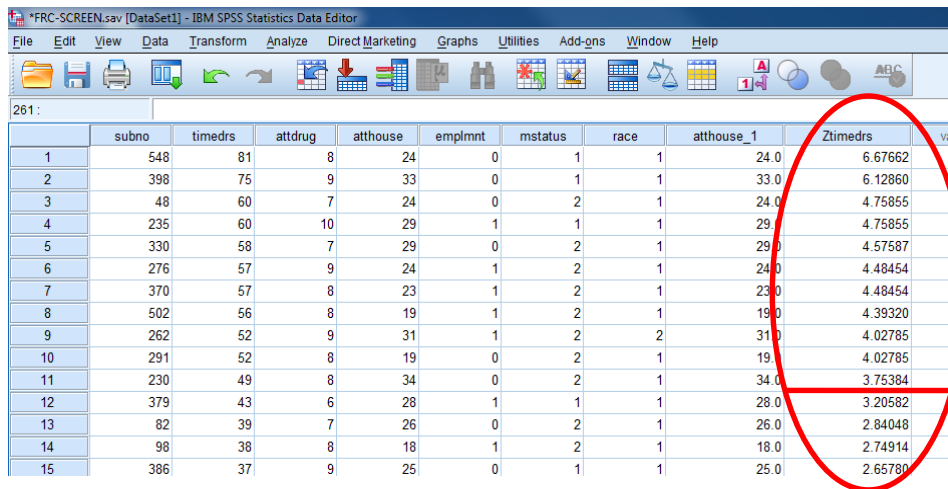
- 4.4.3. Click it over under Sort by:

- 4.4.4. Click on *Descending* under Sort Order



- 4.4.5. Click *OK*

- 4.5. We can also sort by moving the cursor over the variable of interest (e.g., **Ztimedrs**), right clicking on the mouse and click on *Sort Descending*
- 4.6. See sorted “**Ztimedrs**” and notice the first eleven cases are univariate outliers (> 3.29). Also, go to the bottom of the column and notice there are no negative univariate outliers.



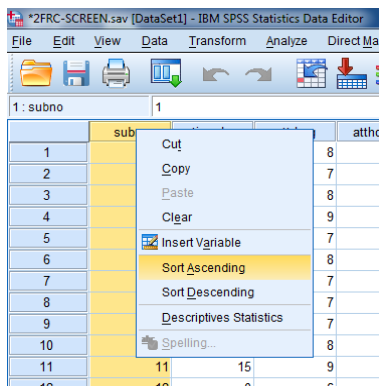
	subno	timedrs	attdrug	athouse	emplmnt	mstatus	race	athouse_1	Ztimedrs	var
1	548	81	8	24	0	1	1	24.0	6.67662	
2	398	75	9	33	0	1	1	33.0	6.12860	
3	48	60	7	24	0	2	1	24.0	4.75855	
4	235	60	10	29	1	1	1	29.0	4.75855	
5	330	58	7	29	0	2	1	29.0	4.57587	
6	276	57	9	24	1	2	1	24.0	4.48454	
7	370	57	8	23	1	2	1	23.0	4.48454	
8	502	56	8	19	1	2	1	19.0	4.39320	
9	262	52	9	31	1	2	2	31.0	4.02785	
10	291	52	8	19	0	2	1	19.0	4.02785	
11	230	49	8	34	0	2	1	34.0	3.75384	
12	379	43	6	28	1	1	1	28.0	3.20582	
13	82	39	7	26	0	2	1	26.0	2.84048	
14	98	38	8	18	1	2	1	18.0	2.74914	
15	386	37	9	25	0	1	1	25.0	2.65780	

- 4.7. If there were no univariate outliers for this ungrouped variable and you plan to split the variable into groups to run an analysis such as ANOVA, then you would analyze *z*-scores for each group’s scores on the variable to see if there are univariate outliers existing within each group.
- 4.8. Save this revised data set, **Save as “2FRC-SCREEN”**.

Tabachnick and Fidell (2007) have identified the following ways to reduce the influence of univariate outliers:

- Delete the variable(s) that may be responsible for many outliers especially if it highly correlated with other variables in the analysis.
- If you decide that cases with extreme scores are not part of the population you sampled then delete them.
- If cases with extreme scores are considered part of the population you sampled then a way to reduce the influence of a univariate outlier is to *transform the variable* to change the shape of the distribution to be more normal. Tukey said you are merely re-expressing what the data have to say in other terms (Howell, 2007).
- Another strategy for dealing with a univariate outlier is to “assign the outlying case(s) a raw score on the offending variable that is one unit larger (or smaller) than the next most extreme score in the distribution” (Tabachnick & Fidell, 2007, p. 77).
- Univariate transformations and score alterations often help reduce the impact of multivariate outliers but they can still be problems. These cases are usually deleted (Tabachnick & Fidell, 2007). **All transformations, changes of scores, and deletions are reported in the results section with the rationale and with citations.**

- 4.9. Our best choice is probably to transform the variable but since the outliers are probably affecting the normality of the variable “Visits to health professionals (**timedrs**)”, however, let’s go ahead and assess the univariate assumptions before we make a decision.
- 4.10. Before we do proceed, let’s return the data set to its original order.
 - 4.10.1. Initiate
 - > *Data*
 - > *Sort Cases*
 - 4.10.2. Click *Reset*
 - 4.10.3. Highlight “**subno**”
 - 4.10.4. Click it over under **Sort by:** click on *Ascending* under **Sort by:**
 - 4.10.5. Click *OK*
 - 4.10.6. Click *Save*
- 4.11. We can also sort by moving the cursor over the variable of interest (e.g., **subno**), right clicking on the mouse and click on *Sort Ascending*



5. SCREENING FOR AND MAKING DECISIONS ABOUT UNIVARIATE ASSUMPTIONS

Many statistical analyses, including ANOVA, “require that all groups come from normal populations with the same variance” (Norusis, 1994a, p. 89).

- Remember that the Central Limit Theorem tells us that, regardless of the shape of the population distribution, the sampling distribution of means, drawn from a population with variance σ^2 and mean μ , will approach a normal distribution with σ^2/N as sample size N increases (Howell, 2007).
- So, the larger the size of each sample that generated each sample mean of the sampling distribution of the mean, the more likely the sampling distribution of the mean will be normally distributed.
- The further the population raw scores depart from normality, the larger the sample size must be for the sampling distribution of the mean to be normally shaped.

Therefore, we need to assess whether all the group variances are equal and that samples come from normal populations.

- If these assumptions are violated then we want to identify appropriate transformations.
- More specifically, for normality we would assess histograms, normal Q-Q plots, skewness, kurtosis, and the Shapiro-Wilks' statistic. We would examine the variance ratio (F_{\max}) and the Levene's Test to make a decision about the assumption of homogeneity of variance.

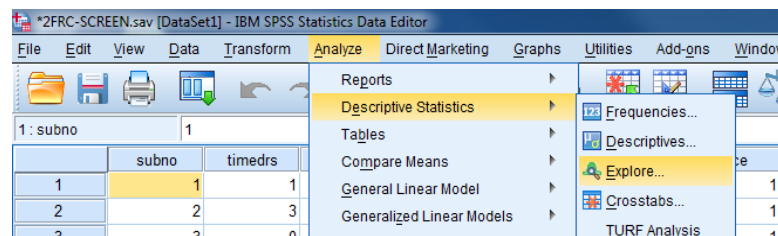
5.1. To illustrate screening for univariate assumptions let's begin by analyzing the dependent variable ("Visits to health professionals [**timedrs**"] from our original research question, for normality. We discovered in the univariate outlier screening that this variable had eleven univariate outliers. We postponed a decision on what to do with the univariate outliers until after our screening for normality and homogeneity of variance. Conduct the following data analysis.

5.1.1. Initiate

> *Analyze*

> *Descriptive Statistics*

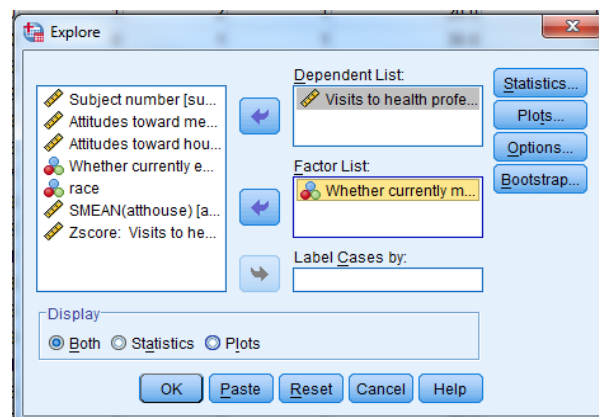
> *Explore*



5.1.2. Click over the dependent variable ("Visits to health professionals [**timedrs**"] to Dependent List:

5.1.3. Click over the independent variable ("Whether currently married [**mstatus**"] to Factor List:

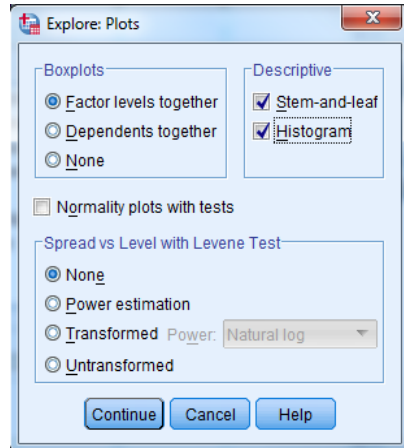
5.1.4. Do not change Display choices – leave on *Both*



5.1.5. In the upper right corner of the dialog box are three buttons. Click on *Plots*

5.1.6. Then, select *Histogram* and *Normality plots with tests*

5.1.7. Click *Continue*



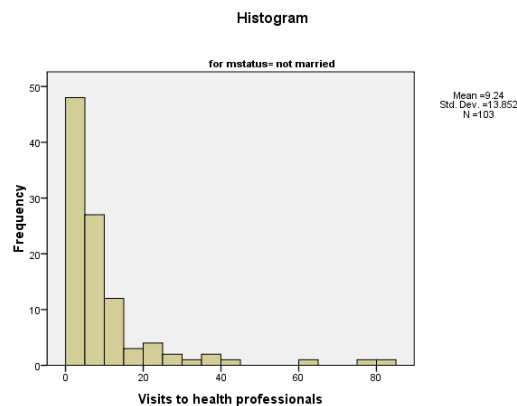
5.1.8. Click *OK*

5.2. Screening for Normality

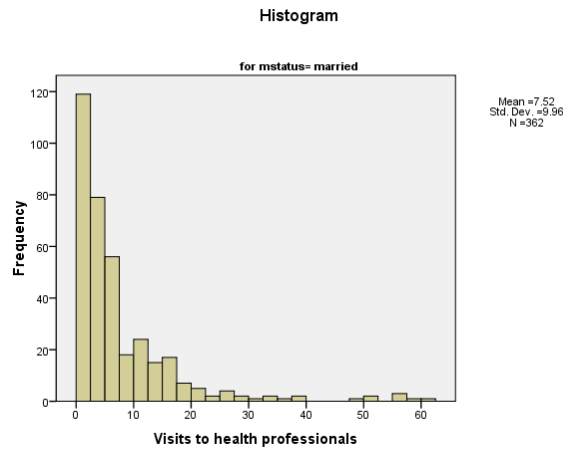
5.2.1. *Histograms*

Histograms give us a general visual description of the distribution of data values. Histograms show to what extent a distribution of values is symmetrical (mesokurtic) and whether cases cluster around a central value. You can see if the shape of the distribution is more peaked or narrow (high in the middle-leptokurtic) or more flat (dispersed-platykurtic). You can also tell if there are values far removed from the other values such far removed values to the right of the distribution (positive skew) or far removed values to the left of the distribution (negative skew).

Refer to the computer output and scroll down until you find “Visits to health professionals, Histograms, for **mstatus** = not married and for **mstatus** = married.” Briefly write down some observations of the two distributions.



“for **mstatus** = not married”

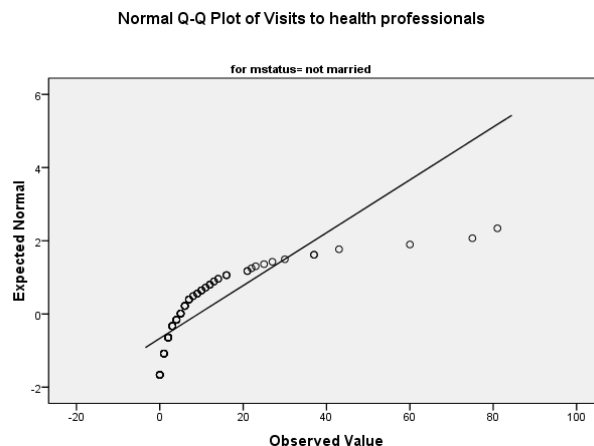


“for **mstatus** = married”

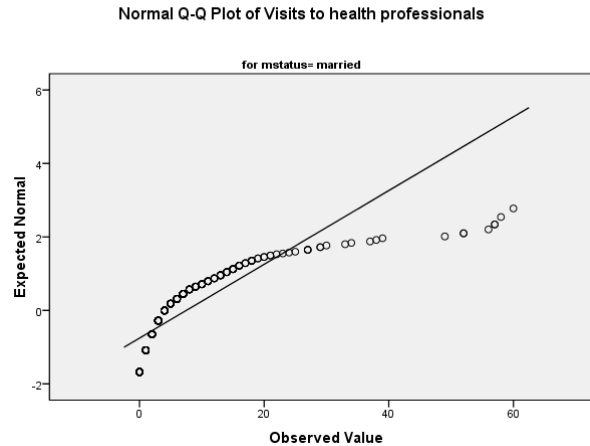
5.2.2. Normal Q-Q Plots

In a **normal probability plot**, each observed value is paired with its expected value from the normal distribution. The expected value from the normal distribution is based on the number of cases in the sample and the rank order of the case in the sample. If the sample is from a normal distribution, we expect that the points will fall more or less on a straight line. A **detrended normal plot** is the actual deviations of the points from a straight line. If the sample is from a normal population, the points should cluster around a horizontal line through 0, and there should be no pattern. A striking pattern suggests departure from normality (Norusis, 1994b).

Scroll down the output and find the normal Q-Q plots and briefly interpret them below.



“for **mstatus** = not married”



“for **mstatus** = married”

5.2.3. Skewness

A distribution that is not symmetric but has more cases (more of a “tail”) toward one end of the distribution than the other is said to be **skewed** (Norusis, 1994a).

- Value of 0 = normal
- Positive Value = positive skew (tail going out to right)
- Negative Value = negative skew (tail going out to left)

Divide the skewness statistic by its standard error. We want to know if this standard score value significantly departs from normality. Concern arises when the skewness statistic divided by its standard error is greater than $z = \pm 3.29$ ($p < .001$, two-tailed test) (Tabachnick & Fidell, 2003, 2007).

Scroll up to the top part of the output to “Descriptives.” You will see the skewness values and their standard error values for “**mstatus** = not married” and for “**mstatus** = married.” Interpret the skewness of the distributions providing the information asked for below.

Descriptives						
Whether currently married				Statistic	Std. Error	
Visits to health professionals	1 not married	Mean		9.24	1.365	
		95% Confidence Interval for Mean	Lower Bound	6.54		
			Upper Bound	11.95		
		5% Trimmed Mean		7.02		
		Median		5.00		
		Variance		191.891		
		Std. Deviation		13.852		
		Minimum		0		
		Maximum		81		
		Range		81		
		Interquartile Range		8		
		Skewness		3.277	.238	
		Kurtosis		12.430	.472	
	2 married	Mean		7.52	.523	
		95% Confidence Interval for Mean	Lower Bound	6.49		
			Upper Bound	8.55		
		5% Trimmed Mean		6.00		
		Median		4.00		
		Variance		99.192		
		Std. Deviation		9.960		
		Minimum		0		
		Maximum		60		
		Range		60		
		Interquartile Range		7		
		Skewness		2.989	.128	
		Kurtosis		10.667	.256	

	Skewness Standard Score	Direction of the Skewness	Significant Departure? (yes, no)
“not married” $\frac{\text{Skewness Value}}{\text{Std. Error}} =$	_____	_____	_____
“married” $\frac{\text{Skewness Value}}{\text{Std. Error}} =$	_____	_____	_____

5.2.4. Kurtosis

Kurtosis is the relative concentration of scores in the center, the upper and lower ends (tails) and the shoulders (between the center and the tails) of a distribution (Norusis, 1994a).

- Value of 0 = mesokurtic (normal, symmetric)
- Positive Value = leptokurtic (shape is more narrow, peaked)
- Negative Value = platykurtic (shape is more broad, widely dispersed, flat)

Divide the kurtosis statistic by its standard error. We want to know if this standard score value significantly departs from normality. Concern arises when the kurtosis statistic divided by its standard error is greater than $z = \pm 3.29$ ($p < .001$, two-tailed test) (Tabachnick & Fidell, 2003, 2007).

Descriptives						
Whether currently married				Statistic	Std. Error	
Visits to health professionals	1 not married	Mean		9.24	1.365	
		95% Confidence Interval for Mean	Lower Bound	6.54		
			Upper Bound	11.95		
		5% Trimmed Mean		7.02		
		Median		5.00		
		Variance		191.891		
		Std. Deviation		13.852		
		Minimum		0		
		Maximum		81		
		Range		81		
		Interquartile Range		8		
		Skewness		3.277	.238	
		Kurtosis		12.430	.472	
	2 married	Mean		7.52	.523	
		95% Confidence Interval for Mean	Lower Bound	6.49		
			Upper Bound	8.55		
		5% Trimmed Mean		6.00		
		Median		4.00		
		Variance		99.192		
		Std. Deviation		9.960		
		Minimum		0		
		Maximum		60		
		Range		60		
		Interquartile Range		7		
		Skewness		2.989	.128	
		Kurtosis		10.667	.256	

Interpret the kurtosis of the distributions providing the information asked for below.

	Kurtosis Standard Score	Direction of the Kurtosis	Significant Departure? (yes, no)
“not married” $\frac{\text{Kurtosis Value}}{\text{Std. Error}} =$	_____	_____	_____
“married” $\frac{\text{Kurtosis Value}}{\text{Std. Error}} =$	_____	_____	_____

5.2.5. Shapiro-Wilks' Test

The **Shapiro-Wilks' test** and the **Kolmogorov-Smirnov test** with **Lilliefors correction** are statistical tests that test the hypothesis that the data are from a normal distribution. If either test is significant then the data is not normally distributed. It is important to remember that whenever the sample size is large, almost any goodness-of-fit test will result in rejection of the null hypothesis since it is almost impossible to find data that are exactly normally distributed. For most statistical tests, it is sufficient that the data are approximately normally distributed (Norusis, 1994a). According to Stevens (2002), the Kolmogorov-Smirnov test was not shown as powerful as the Shapiro-Wilk. Also, with sampling sizes from 10-50, the combination of skewness and kurtosis coefficients and the Shapiro-Wilk test were the most powerful in detecting departures from normality (Stevens, 2002).

Scroll down the output to the box that has the heading “Tests of Normality” and interpret the Shapiro-Wilks’ test providing the information asked for below. Use an alpha level of .001.

Tests of Normality						
	Whether currently married	Kolmogorov-Smirnov ^a			Shapiro-Wilk	
		Statistic	df	Sig.	Statistic	Sig.
→ Visits to health professionals	1 not married	.252	103	.000	.607	.000
	2 married	.225	362	.000	.660	.000

a. Lilliefors Significance Correction

We are testing H_0 : sampling distribution = normal.

	S-W Statistic	Sig. Probabilities	Reject Null? (yes or no)	Normal? (yes or no)
“not married”	_____	_____	_____	_____
“married”	_____	_____	_____	_____

5.3. Screening for Homogeneity of Variance

5.3.1. Variance Ratio Analysis

A **variance ratio analysis** can be obtained by dividing the lowest variance of a group for two groups into the highest group variance of the two group variances. Concern arises if the resulting ratio is 4-5 + which indicates that the largest variance is 4 to 5 times the smallest variance. Tabachnick and Fidell (2007) refer to this ratio as F_{\max} and state “If sample sizes are relatively equal (within a ratio of 4 to 1 or less for largest to smallest cell size), an F_{\max} as great as 10 is acceptable. As the cell-size discrepancy increases (say, goes to 9 to 1 instead of 4 to 1), an F_{\max} as small as 3 is associated with inflated Type I error if the larger variance is associated with the smaller cells size” (p. 86).

Scroll up the output to the “Descriptives” section and calculate the variance ratio by dividing the smallest group variance into the largest group variance. Then, interpret the variance ratio analysis.

Descriptives				Statistic	Std. Error
Whether currently married					
Visits to health professionals	1 not married	Mean		9.24	1.365
		95% Confidence Interval for Mean	Lower Bound	6.54	
			Upper Bound	11.95	
		5% Trimmed Mean		7.02	
		Median		5.00	
		Variance		191.891	
		Std. Deviation		13.852	
		Minimum		0	
		Maximum		81	
		Range		81	
		Interquartile Range		8	
		Skewness		3.277	.238
		Kurtosis		12.430	.472
	2 married	Mean		7.52	.523
		95% Confidence Interval for Mean	Lower Bound	6.49	
			Upper Bound	8.55	
		5% Trimmed mean		6.00	
		Median		4.00	
		Variance		99.192	
		Std. Deviation		9.960	
		Minimum		0	
		Maximum		60	
		Range		60	
		Interquartile Range		7	
		Skewness		2.989	.128
		Kurtosis		10.667	.256

$$\text{Variance Ratio} = \frac{\text{Largest Group Variance}}{\text{Smallest Group Variance}} = \underline{\hspace{2cm}}$$

5.3.2. Levene Test

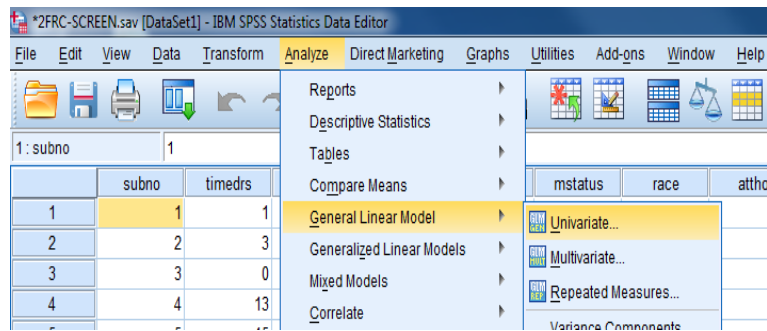
The **Levene test** is a homogeneity-of-variance test that is less dependent on the assumption of normality than most tests and thus is particularly useful with analysis of variance. It is obtained by computing, for each case, the absolute differences from its cell mean and performing a one-way analysis of variance on these differences. If the Levene test statistic is significant then the groups are not homogeneous and we may need to consider transforming the original data or using a non-parametric statistic (Norusis, 1994a).

To obtain results for the Levene test, we will need to conduct another analysis. Follow the directions below. First, click out of the output we have been using.

Commands to obtain the Levene Statistic for homogeneity of variance.

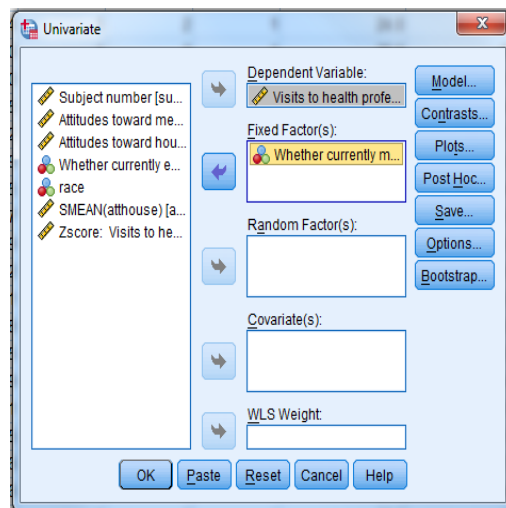
5.3.2.1. Initiate

- > *Analyze*
- > *General Linear Model*
- > *Univariate*



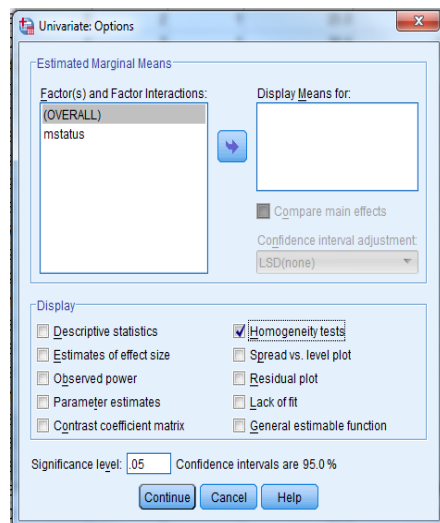
5.3.2.2. Click over the dependent variable (“Visits to health professionals [timedrs]”) to **Dependent Variable:** list.

5.3.2.3. Click over the independent variable (“Whether currently married [mstatus]”) to **Fixed Factor(s):** list.



5.3.2.4. Click on *Options*

5.3.2.5. Click on *Homogeneity tests* in the Display box.



5.3.2.56 Click *Continue*

5.3.2.7. Click *OK*

Briefly interpret the Levene's Test for Homogeneity of Variance. Use an alpha of .01.
We are testing the hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

Levene's Test of Equality of Error Variances^a

Dependent Variable: Visits to health professionals			
F	df1	df2	Sig.
4.421	1	463	.036

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
a. Design: Intercept + mstatus

Levene Statistic	Sig. Probability	Reject Null? (yes or no)	Homogeneous? (yes or no)
_____	_____	_____	_____

5.4. Summary of Our Screening Results for Univariate Outliers and Assumptions

- We detected 11 univariate outliers.
- We found the measures of homogeneity of variance (variance ratio and Levene test) were acceptable. However, the measures of normality (histograms, Q-Q plots, skewness, kurtosis, and Shapiro-Wilk test) were unacceptable.
- One option would be to delete the univariate outliers if you decide that cases with extreme scores are not part of the population that you sampled then delete them and see if normality improves.
- Another strategy for dealing with univariate outliers is to “assign the outlying case(s) a raw score on the offending variable that is one unit larger (or smaller) than the next most extreme score in the distribution” (p. 77, Tabachnick & Fidell, 2007).
- The option we will choose to employ is transformation of the dependent variable (“Visits to health professionals [**timedrs**]”).

6. TRANSFORMATION OF THE DEPENDENT VARIABLE

If cases with extreme scores are considered part of the population you sampled then a way to reduce the influence of a univariate outlier is to *transform the variable* to change the shape of the distribution to be more normal. Tukey said you are merely **re-expressing** what the data have to say in other terms. (Howell, 1987).

Both Tabachnick and Fidell (2007) and Stevens (2002) provide guides on what type of transformation to use depending upon the shape of the distribution you are planning to transform. For example, a square root or a log10 transformation can be used for positively skewed distributions to normalize them. For negatively skewed distributions, reflecting the negative distribution and then use a square root or a log transformation may normalize the distribution.

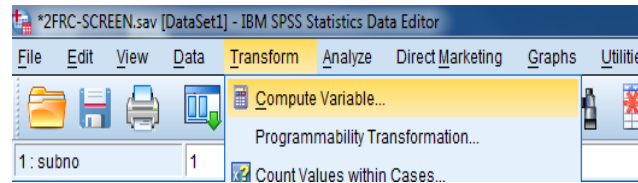
Since we have a seriously positively skewed dependent variable (“Visits to health professionals [**timedrs**]”), we are going to employ a log10 transformation.

6.1. Run the following analysis after clicking out of the output we have been using.

6.1.1. Initiate

> *Transform*

> *Compute Variable*



6.1.2. Under Target Variable type “**Itimedrs**”

6.1.3. Under Function Group: click ALL (or Arithmetic)

6.1.4. Then in Functions and Special Variables: scroll down until you find Lg10

6.1.5. Click on it (Lg10)

6.1.6. Click on the arrow at the top (to the left) of the Function Group: box.

6.1.7. Next go to variables under Type and Label

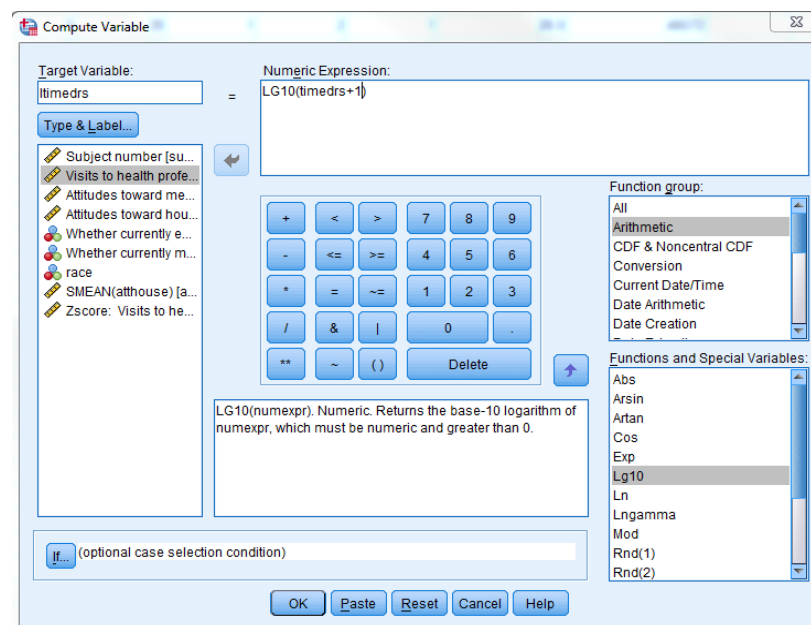
6.1.8. Click on “Visits to health professionals (**timedrs**)”

6.1.9. Click on the arrow to the right of the box of variables.

It will show up in the place under Numeric Expression: where the “?” was.


Since the data has zeros, you will need to add +1.

So, the Numeric Expression: should look like **LG10(timedrs+1)**



6.1.10. Click *OK*

The log10 transformed variable (“**ltime**drs”) will be in your Data View spreadsheet.



The screenshot shows the IBM SPSS Statistics Data Editor window with the file name '*2FRC-SCREEN.sav [DataSet1]'. The Data View spreadsheet is displayed with columns: subno, timedrs, attdrug, atthouse, emplmnt, mstatus, race, atthouse_1, Ztimedrs, ltime'drs, and var. The 'ltime'drs column is circled in red. The data rows are as follows:

	subno	timedrs	attdrug	athouse	emplmnt	mstatus	race	athouse_1	Ztimedrs	ltime'drs	var
1	1	1	8	27	1	2	1	27.0	-.63032	.30	
2	2	3	7	20	0	2	1	20.0	-.44765	.60	
3	3	0	8	23	0	2	1	23.0	-.72166	.00	

Now, let's see if the log10 transformation was successful in normalizing the distribution of the dependent variable.

6.2. We will repeat the exploration commands to look at the measures of normality on the transformed variable “**ltime**drs.”

6.2.1. Initiate

> *Analyze*

> *Descriptive Statistics*

> *Explore*

6.2.2. Click on *Reset*

6.2.3. Click over the transformed dependent variable (“**ltime**drs”) to Dependent List:

6.2.4. Click over independent variable (“Whether currently married [**mstatus**]”) to Factor List:

6.2.5. Do not change Display choices-leave on *Both*

6.2.6. In the upper right corner of the dialog box are three buttons. Click on *Plots*

6.2.7. Then, select *Histogram* and *Normality plots with tests*

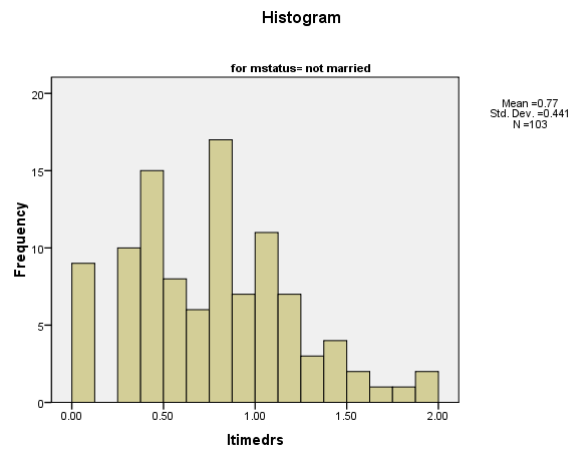
6.2.8. Click *Continue*

6.2.9. Click *OK*

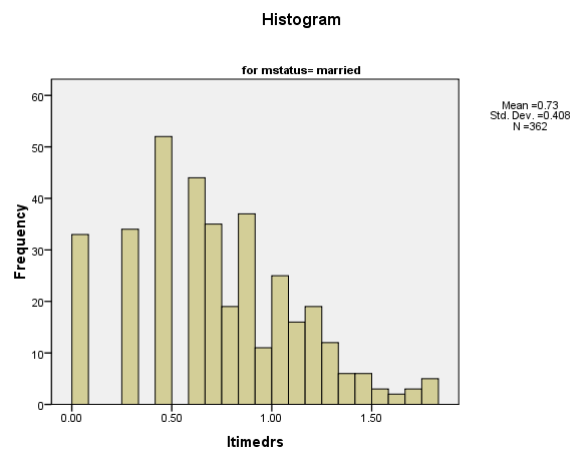
6.3. Provide interpretation information next.

6.3.1. *Histograms*

“for **mstatus** = not married”

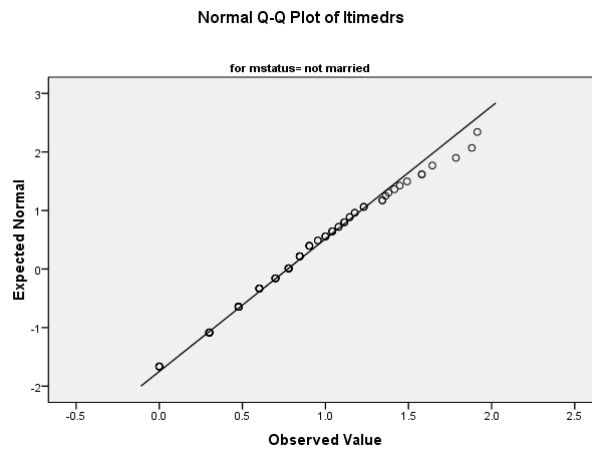


“for **mstatus** = married”

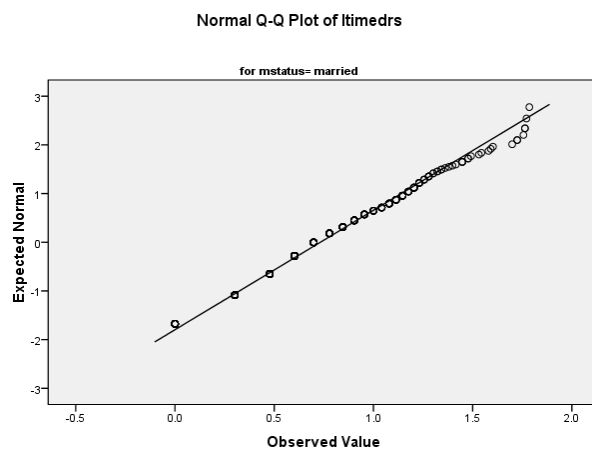


6.3.2. Normal Q-Q Plots

“for **mstatus** = not married”



“for **mstatus** = married”



6.3.3. Skewness

	Skewness Standard Score	Direction of the Skewness	Significant Departure? (yes, no)
“not married” $\frac{\text{Skewness Value}}{\text{Std. Error}} =$	_____	_____	_____
“married” $\frac{\text{Skewness Value}}{\text{Std. Error}} =$	_____	_____	_____

Descriptives

Whether currently married				Statistic	Std. Error
Itimedrs	1 not married	Mean		.7724	.04350
		95% Confidence Interval for Mean	Lower Bound	.6861	
			Upper Bound	.8587	
		5% Trimmed Mean		.7607	
		Median		.7782	
		Variance		.195	
		Std. Deviation		.44148	
		Minimum		.00	
		Maximum		1.91	
		Range		1.91	
		Interquartile Range		.56	
		Skewness		.296	.238
		Kurtosis		-.099	.472
	2 married	Mean		.7324	.02143
		95% Confidence Interval for Mean	Lower Bound	.6903	
			Upper Bound	.7746	
		5% Trimmed Mean		.7238	
		Median		.6990	
		Variance		.166	
		Std. Deviation		.40769	
		Minimum		.00	
		Maximum		1.79	
		Range		1.79	
		Interquartile Range		.53	
		Skewness		.195	.128
		Kurtosis		-.223	.256

6.3.4. *Kurtosis*

	Kurtosis Standard Score	Direction of the Kurtosis	Significant Departure? (yes, no)
“not married” $\frac{\text{Kurtosis Value}}{\text{Std. Error}} =$	_____	_____	_____
“married” $\frac{\text{Kurtosis Value}}{\text{Std. Error}} =$	_____	_____	_____

6.3.5. *Shapiro-Wilks' Test*

	S-W Statistic	Sig. Probabilities	Reject Null? (yes or no)	Normal? (yes or no)
“not married”	_____	_____	_____	_____
“married”	_____	_____	_____	_____

Tests of Normality							
Whether currently married		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
ltimehrs	1 not married	.078	103	.126	.975	103	.046
	2 married	.080	362	.000	.975	362	.000

a. Lilliefors Significance Correction

Since the dependent variable is now behaving more properly after the transformation, we can now answer the study question, Is there a significant difference between married and not married women in the number of visits they make to health professionals by testing the null hypothesis that there will be no differences between married and not married women in the number of visits they make to health professionals ($H_0: \mu_1 = \mu_2$).

6.4. Conduct the following analysis after clicking out of the output we have been using.

6.4.1. Initiate

> *Analyze*

> *General Linear Model*

> *Univariate*

6.4.2. Click on *Reset*

6.4.3. Click over the transformed dependent variable (“**ltimehrs**”) to **Dependent Variable:** list

6.4.4. Click over independent variable (“Whether currently married [**mstatus**]”) to **Fixed Factor(s):** list

6.4.5. Click on *Options*

6.4.6. Click on *Homogeneity tests* in the Display box.

6.4.7. Click *Continue*

6.4.8. Click *OK*

6.5. Provide the information below. We will use an alpha criterion of .01 to test the null hypothesis.

6.5.1. First, let’s see how the log10 transformation adjusted our variance.

Levene Statistic	Sig. Probability	Reject Null? (yes or no)	Homogeneous? (yes or no)
_____	_____	_____	_____

Levene's Test of Equality of Error Variances^a

Dependent Variable: ltimehrs			
F	df1	df2	Sig.
.589	1	463	.443

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + mstatus

6.5.2. Now provide the information for the analysis of variance:

Source *F* Sig.
mstatus _____

Tests of Between-Subjects Effects

Dependent Variable: ltimedrs

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.128 ^a	1	.128	.742	.390
Intercept	181.578	1	181.578	1052.433	.000
mstatus	.128	1	.128	.742	.390
Error	79.882	463	.173		
Total	335.529	465			
Corrected Total	80.010	464			

a. R Squared = .002 (Adjusted R Squared = -.001)

We retain the null hypothesis ($p > .01$) that there was no difference between married and not married women in the number of visits they make to health professionals.

Glossary of Terms

Mean: The average of a group of numbers.

Median: The middle number in a group of ordered numbers.

Mode: The number in a group that is most repeated.

Regression: A statistical measure that attempts to determine the strength of the relationship between one dependent variable and a series of independent variables.

Univariate analysis: The statistical model of analysis that compares one independent variable to a dependent variable.

Multivariate analysis: The statistical model of analysis that compares multiple independent variables to a dependent variable.

Attrition: The loss of participants throughout data gathering; participants who did not continue throughout data collection to completion. It may include dropout, nonresponse (not completing a survey or neglecting to answer most of the questions), or withdrawal.

One-Way ANOVA: A statistical test which examines the equality of three or more means at one time by using variances. It helps test multiple levels of an independent variable to determine main effects or interaction effects.

Independent t-test: A statistical test which determines whether a difference between two groups' averages is unlikely to have occurred because of random chance in sample selection. Its primary outputs include statistical significance and effect size.

Outliers (univariate/multivariate): An observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error. There are many options to deal with outliers, including transforming, replacing, or deleting.

Univariate Assumptions: The assumptions of normality, independence, and homogeneity of variance. When assumptions are met, they provide that the distributions in the populations from which the samples are selected have the same shapes, means, and variances. In other words, they are the same populations; the variances on the dependent variable are equal across groups.

- **Normality:** The assumption that the distributions of the populations from which the samples are selected are normal.
- **Independence:** The observations are random and independent samples from the populations.
- **Homogeneity of variance:** The assumption that the variances of the distributions in the populations from which the samples are selected are equal.

Goodness-of-fit: A description of how well a statistical model fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Tests for goodness-of-fit include the Kolmogorov-Smirnov test, sum of squares, or Pearson's chi-squared test.

Z-score: A transformed score that describes the difference between a raw score and the population mean in terms of its standard deviation. Z-scores have a mean of 0 and a standard deviation of 1. The z-score is negative when the score falls below the mean and positive when the score falls above the mean.

Answer KEY

5.2.1. Histograms

“for **mstatus** = not married”

seriously positively skewed, leptokurtic, does not appear normally distributed, potential concern with the two breaks in the distribution

“for **mstatus** = married”

seriously positively skewed, leptokurtic, does not appear normally distributed, potential concern with the two breaks in the distribution

5.2.2. Normal Q-Q Plots

“for **mstatus** = not married”

most points are off, not on or near, the straight (linear) line; does not appear to meet the assumption of normality

“for **mstatus** = married”

most points are off, not on or near, the straight (linear) line; does not appear to meet the assumption of normality

5.2.3. Skewness

	Skewness Standard Score	Direction of the Skewness	Significant Departure? (yes, no)
“not married” Skewness Value Std. Error =	3.277/.238 = <u>13.77</u>	<u>positive</u>	<u>Yes (> 3.29)</u>
“married” Skewness Value Std. Error =	2.989/.128 = <u>23.35</u>	<u>positive</u>	<u>Yes (> 3.29)</u>

5.2.4. Kurtosis

	Kurtosis Standard Score	Direction of the Kurtosis	Significant Departure? (yes, no)
“not married” $\frac{\text{Kurtosis Value}}{\text{Std. Error}} =$	$12.430/.472 =$ <u>26.33</u>	<u>leptokurtic</u>	<u>Yes (> 3.29)</u>
“married” $\frac{\text{Kurtosis Value}}{\text{Std. Error}} =$	$10.667/.256 =$ <u>41.67</u>	<u>leptokurtic</u>	<u>Yes (> 3.29)</u>

5.2.5. Shapiro-Wilks' Test

	S-W Statistic	Sig. Probabilities	Reject Null? (yes or no)	Normal? (yes or no)
“not married”	<u>.607</u>	<u>.000</u> $p < .001$	<u>yes</u>	<u>no</u>
“married”	<u>.660</u>	<u>.000</u> $p < .001$	<u>yes</u>	<u>no</u>

5.3. Screening for Homogeneity of Variance

5.3.1. Variance Ratio Analysis

$$\text{Variance Ratio} = \frac{\text{Largest Group Variance}}{\text{Smallest Group Variance}} = 191.891/99.192 = \underline{1.93}$$

5.3.2. Levene Test

Levene Statistic	Sig. Probability	Reject Null? (yes or no)	Homogeneous? (yes or no)
<u>4.421</u>	<u>.036</u>	<u>no</u>	<u>yes</u>

6.3.1. Histograms

“for **mstatus** = not married”

slight positive skew, more normally distributed, no major splits in the distribution

“for **mstatus** = married”

slight positive skew, more normally distributed, no major splits in the distribution

6.3.2. Normal Q-Q Plots

“for **mstatus** = not married”

more points are on or near the straight (linear) line, more normally distributed

“for **mstatus** = married”

more points are on or near the straight (linear) line, more normally distributed

6.3.3. Skewness

“not married” Skewness Value Std. Error =	Skewness Standard Score .296/.238 = <u>1.24</u>	Direction of the Skewness <u>Positive</u> <u>(More Normal)</u>	Significant Departure? (yes, no) <u>No (< 3.29)</u>
“married” Skewness Value Std. Error =	.195/.128 = <u>1.52</u>	<u>Positive</u> <u>(More Normal)</u>	<u>No (< 3.29)</u>

6.3.4. Kurtosis

“not married” Kurtosis Value Std. Error =	Kurtosis Standard Score -.098/.472 = <u>-0.21</u>	Direction of the Kurtosis <u>Platykurtic</u> <u>(More Normal)</u>	Significant Departure? (yes, no) <u>No (< 3.29)</u>
“married” Kurtosis Value Std. Error =	-.223/.256 = <u>-0.87</u>	<u>Platykurtic</u> <u>(More Normal)</u>	<u>No (< 3.29)</u>

6.3.5. Shaprio-Wilks' Test

	S-W Statistic	Sig. Probabilities	Reject Null? (yes or no)	Normal? (yes or no)
“not married”	<u>.975</u>	<u>.046</u>	<u>no</u>	<u>yes</u>
“married”	<u>.975</u>	<u>.000</u>	<u>yes</u>	<u>no</u>
		<u>p < .001</u>		

6.5.1. *Levene's Statistic*

Levene Statistic	Sig. Probability	Reject Null? (yes or no)	Homogeneous? (yes or no)
<u>.589</u>	<u>.443</u>	<u>no</u>	<u>yes</u>

6.5.2. Now provide the information for the analysis of variance:

Source	<i>F</i>	Sig.
mstatus	<u>.742</u>	<u>.390</u>

References

- Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.
- Norusis, M. J. (1994a). *SPSS 6.1 base system user's guide, part 2*. Chicago, IL: SPSS Inc.
- Norusis, M. J. (1994b). *SPSS advanced statistics 6.1*. Chicago, IL: SPSS Inc.
- Osterlind, S. J., & Tabachnick, B. G. (2001). *SPSS for windows workbook to accompany using multivariate statistics*. Needham Heights, MA: Allyn & Bacon.
- Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn & Bacon.
- Tabachnick, B. G., & Fidell, L. S. (2003, May). *Preparatory data analyses*. Paper presented at the annual meeting of the Western Psychological Association, Vancouver, BC, Canada.