GETTING COMFORTABLE WITH YOUR DATA

UNIVERSITY OF GUELPH

LUCIA COSTANZO

lcostanz@uoguelph.ca

REVISED SEPTEMBER 2012

CONTENTS

SPSS AVAILABILITY

Faculty, staff and students at the University of Guelph may access SPSS two different ways:

1. **Library computers**
   On the library computers, SPSS 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 SPSS at a cost. A free concurrent copy of SPSS is available to faculty, staff or graduate students at the University of Guelph. 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 developed in "Getting your data into SPSS". Participants are expected to have basic SPSS skills and statistical knowledge. Specific goals of this workshop are:

- To review reading data into SPSS
- To learn how to determine whether your data comes from a normal distributions
- How do we transform data if it is required
- Plotting your data

DATA FOR SPSS 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)

Variable PROV : Province of the respondent

Values	Categories
10	N.L.
11	P.E.I.
12	Nova Scotia
13	N.B.
24	Quebec
35	Ontario
46	Manitoba
47	Saskatchewan
48	Alberta
59	B.C.

Variable HHSIZE : # of people in the household

Values	Categories
1	
2	
3	
4	
5	5 or more

Variable SC_Q100 : What was the main reason you began to smoke again?

Values	Categories
1	To control body weight
2	Stress, need to relax or to calm down
3	Boredom
4	Addiction / habit
5	Lack of support or information
6	Going out more (bars, parties)
7	Increased availability
8	No reason / felt like it
9	Family or friends smoke
10	Other
96	Valid skip
97	Don't know
98	Refusal
99	Not stated

Variable DVURBAN : Characteristic of community

Values	Categories
1	Urban
2	Rural
9	Not stated

Variable DVMARST : Grouped marital status of respondent

Values	Categories
1	Common-law/Married
2	Widow/Divorced/Separated
3	Single
9	Not stated

Variable PS_Q30 : Age smoked first cigarette
Variable PS_Q40 : Age begin smoking cigarettes daily
Variable HS_Q20 : # of people that smoke inside the home
Variable WP_Q10A : # of cigarettes smoked-Monday
Variable WP_Q10B : # of cigarettes smoked-Tuesday
Variable WP_Q10C : # of cigarettes smoked-Wednesday
Variable WP_Q10D : # of cigarettes smoked-Thursday
Variable WP_Q10E : # of cigarettes smoked-Friday
Variable WP_Q10F : # of cigarettes smoked-Saturday
Variable WP_Q10G : # of cigarettes smoked-Sunday

Values	Categories
96	Valid skip
97	Don't know
98	Refusal
99	Not stated

Statistical Refresher

Types of Statistics

Two broad types of statistics exist 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 measures. Inferential statistics allow you to make conclusions regarding the data i.e. significant difference, relationships between variables, etc.

Here are some examples of descriptives and inferential statistics:

Descriptives	**Inferential**
• Frequencies • Means • Standard Deviations • Ranges • Medians • Modes	• t-tests • Chi-squares • ANOVA • 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 measurements 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 educations or satisfaction/attitude questions.

Interval

Are measured on a scale 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 values of the 50^{th} percentile)
Interval/Scale	Mean – The arithmetic average of a distribution.

SUMMARY STATISTICS USING FREQUENCIES

Summaries of individual variables provide an important "first look" at your data. Some of the tasks that these summaries help you to complete are listed below:

- Determining "typical" values of the variables. What values occur most often? What range of values are you likely to see?
- Checking the assumptions for statistical procedures. Do you have enough observations? For each variable, is the observed distribution of values adequate?
- Checking the quality of the data. Are there missing or mis-entered values? Are there values that should be recoded?

The Frequencies procedure is useful for obtaining summaries of individual variables. The following examples show how Frequencies can be used to analyze variables measured at nominal, ordinal, and scale levels.

USING FREQUENCIES TO STUDY NOMINAL DATA

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

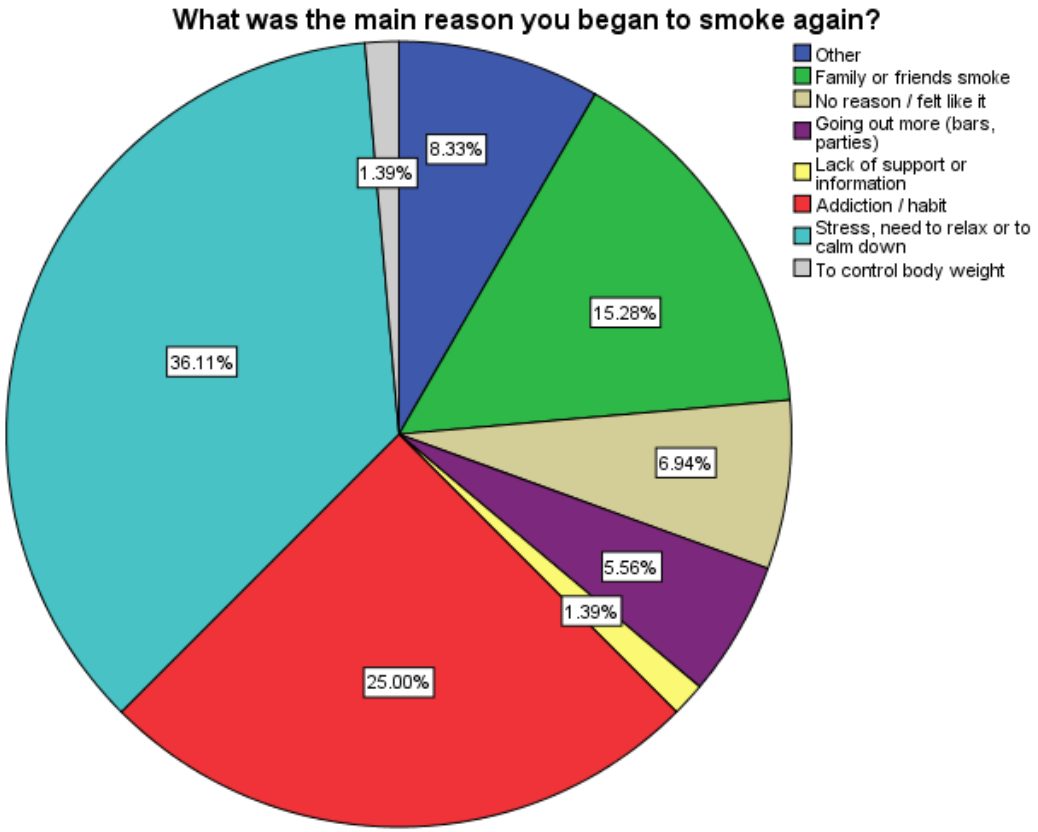
Exercise: Using Frequencies to Study Nominal Data
Lets run frequencies on the variable SC_Q100 which records the main reason the respondent began to smoke again.

GENERATING A FREQUENCIES TO STUDY NOMINAL DATA

1. Click on `Analyze>Descriptive Statistics>Frequencies`.
2. Select which variable you wish to analyze.
3. Select the `Charts` button and select `Pie charts`.

Results from Frequencies for Nominal Data

A pie chart is a good visual tool for assessing the relative frequencies of each category.



At a glance, you see that the main reason that the respondent began to smoke again with to relieve stress and needed to relax/calm down.

Statistics

What was the main reason you began to smoke again?

N	Valid	72
	Missing	268
Mode		2

What was the main reason you began to smoke again?

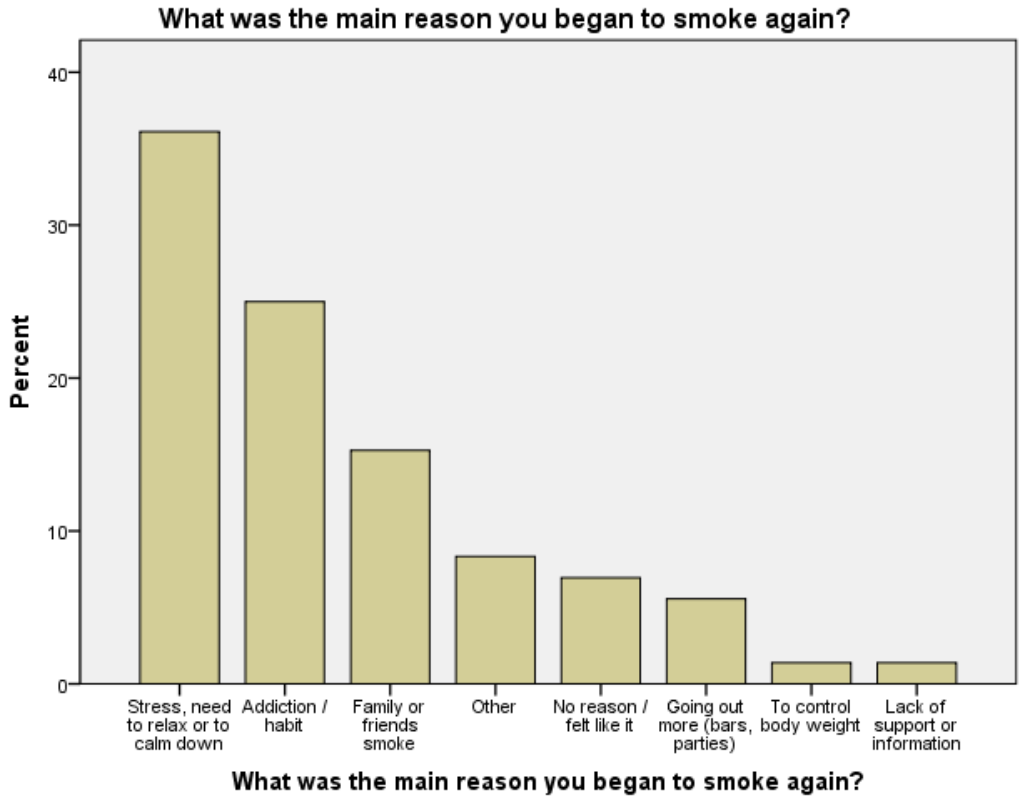
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Other	6	1.8	8.3	8.3
	Family or friends smoke	11	3.2	15.3	23.6
	No reason / felt like it	5	1.5	6.9	30.6
	Going out more (bars, parties)	4	1.2	5.6	36.1
	Lack of support or information	1	.3	1.4	37.5
	Addiction / habit	18	5.3	25.0	62.5
	Stress, need to relax or to calm down	26	7.6	36.1	98.6
	To control body weight	1	.3	1.4	100.0
	Total	72	21.2	100.0	
Missing	99	9	2.6		
	97	3	.9		
	96	256	75.3		
	Total	268	78.8		
Total		340	100.0		

Generating a Bar Chart to Study Nominal Data

A bar chart, ordered by descending frequencies, quickly helps you to find the mode and also to visually compare the relative frequencies.

1. Click on `Analyze>Descriptive Statistics>Frequencies`.
2. Select which variable you wish to analyze.
3. Select the `Charts` button and select `Bar charts`.
4. Click on the `Format` button and choose `Descending counts` then click on `Continue`.

The following bar chart is produced and is another way to visually display the data.



USING FREQUENCIES TO STUDY ORDINAL DATA

Ordinal data have 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 in equality. Examples include: results of a horse race, level of education or satisfaction/attitude questions.

Exercise: Using Frequencies to Study Ordinal Data
Let's run summary statistics on the variable (HHSIZE) that represents the number of people in the household.

GENERATING A FREQUENCIES TO STUDY ORDINAL DATA

1. Click on `Analyze>Descriptive Statistics>Frequencies`.
2. Select which variable you wish to analyze.
3. Select the `Charts` button and select `Bar charts`.
4. Click on the `Format` button and choose `Descending values` then click on `Continue`.

RESULTS FROM FREQUENCIES FOR ORDINAL DATA

- The frequency table for the ordinal data serves much the same purpose as the table for nominal data. For example, you can see from table that 34% of respondents come from a household with 2 members.
- However when studying ordinal data, the Cumulative Percent is much more useful. The table has been ordered by descending values, shows that 50.0% of the contacts have 2 or less members in the household.

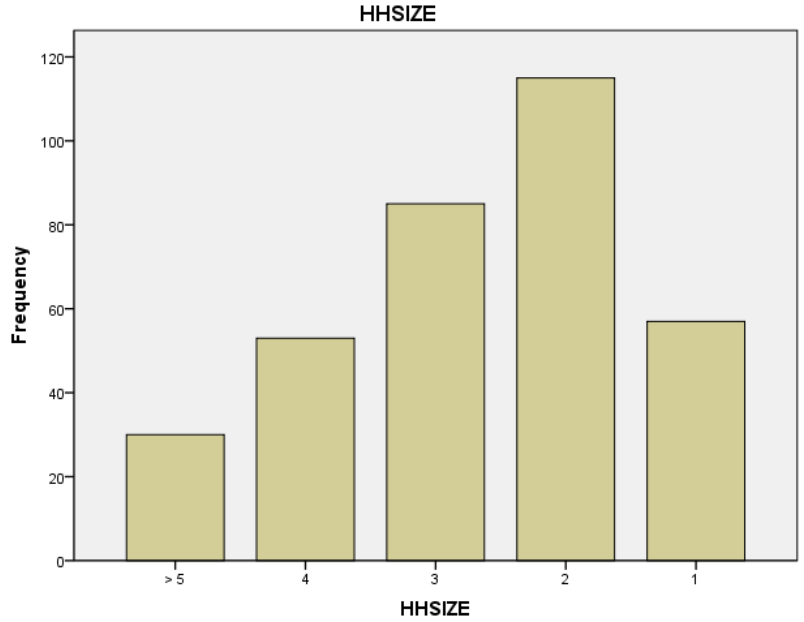
Statistics

HHSIZE

N	Valid	340
	Missing	0
Median		2.00

HHSIZE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	57	16.8	16.8	16.8
	2	115	33.8	33.8	50.6
	3	85	25.0	25.0	75.6
	4	53	15.6	15.6	91.2
	> 5	30	8.8	8.8	100.0
	Total	340	100.0	100.0	



As long as the ordering of values remains intact, reversed or not, the pattern in the bar chart contains information about the distribution of number of people in the household. The frequencies of number of people in household increases from >5 to 2 members in the household, then drops off.

Using Frequencies to Study Scale Data

There are two types of scale data:

Interval

Are measured on a scale 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.

Exercise: Using Frequencies to Study Scale Data
Lets run summary statistics on the variable (totcig), the total number of cigarettes in a week.

Generating a Frequencies to Study Ordinal data

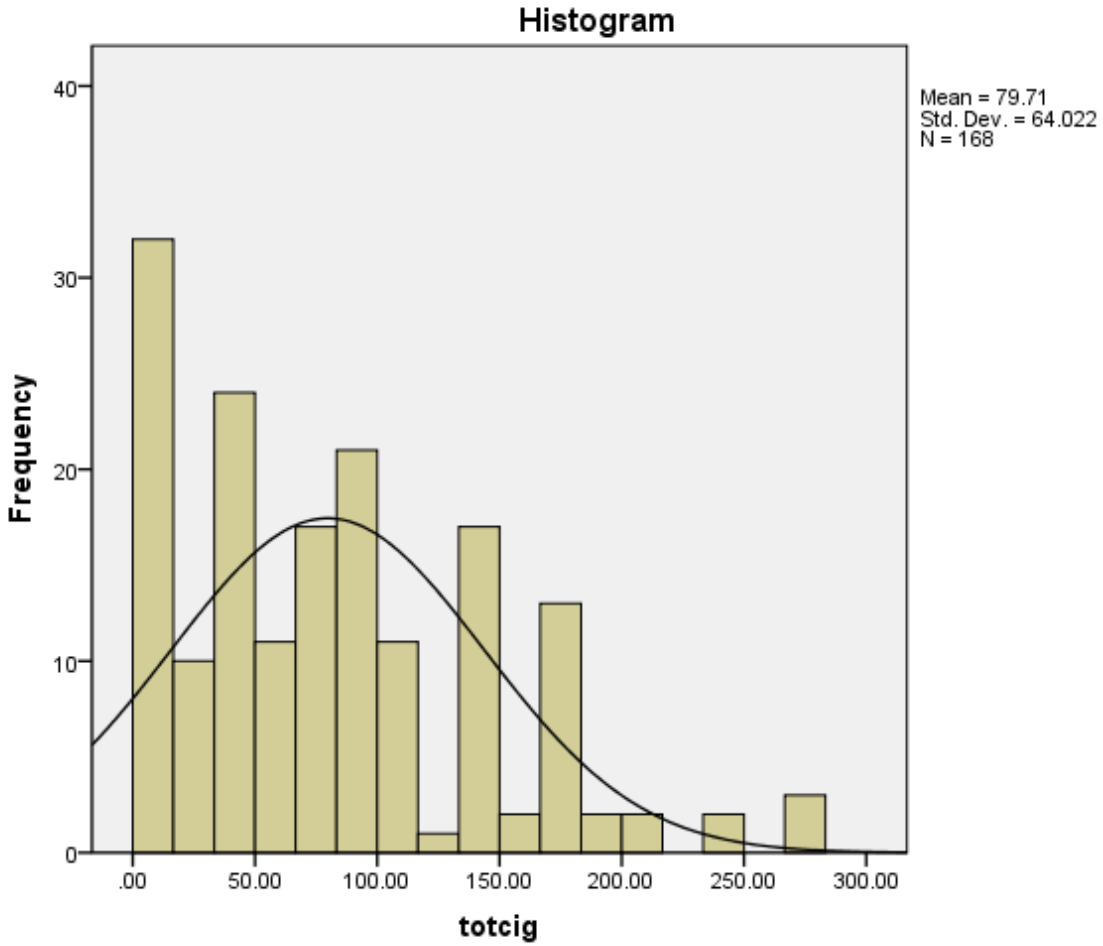
1. Click on `Analyze>Descriptive Statistics>Frequencies`.
2. Select which variable you wish to analyze.
3. Select the `Charts` button and select Histogram with Normal Curve.
4. Click on the `Statistics Button and select Mean, Median, Quartiles, Min, Max and Std Deviation`.

RESULTS FROM FREQUENCIES FOR SCALE DATA

Statistics

totcig

N	Valid	168
	Missing	172
Mean		79.7143
Std. Deviation		64.02190
Minimum		.00
Maximum		280.00
Percentiles	25	32.0000
	50	70.0000
	75	118.0000

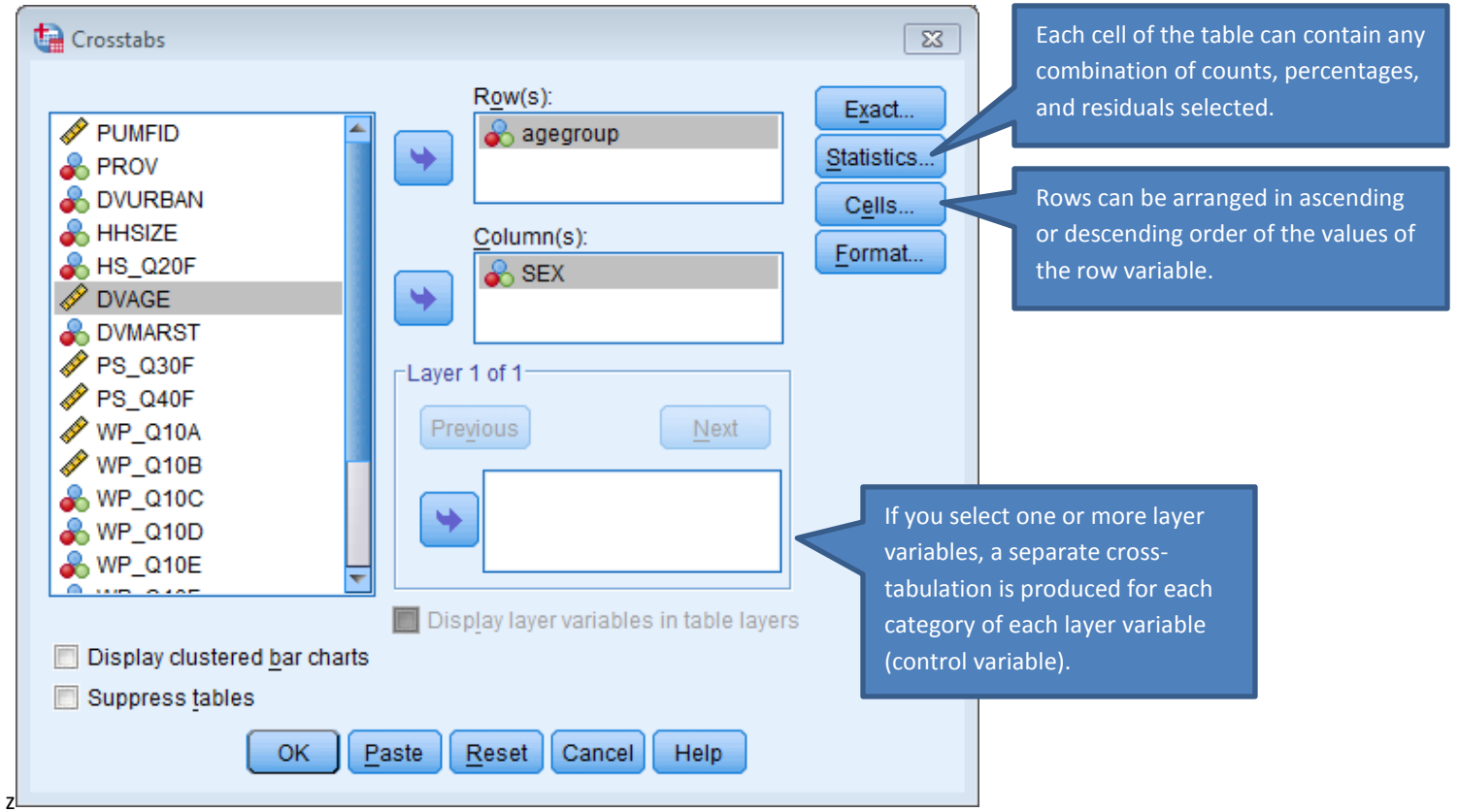

Histogram
Mean = 79.71
Std. Dev. = 64.022
N = 168
Frequency
40
30
20
10
0
.00
50.00
100.00
150.00
200.00
250.00
300.00
totcig

CROSSTABULATION TABLES

(Adapted from SPSS Help Files)

A crosstabulation (crosstab) table also referred to as contingency table allows you to display the relationship between two or more categorical variables. These categorical variables can be either nominal or ordinal in nature. The data format can be either numeric or alphanumeric. Crosstabs can be thought of as joint frequency distribution for more than one variable. The size of the table is determined by the number of distinct values for each variable, with each cell in the table representing a unique combination of values.

Within each cell, a percentage or counts can be reported accordingly. There are several statistical tests allow you to determine whether there is a relationship between the variables in a crosstab. If a relationship exists, the strength can be determined.



GENERATING A CROSSTABULATION

1. Click on `Analyze>Descriptive Statistics>Crosstabs`.
2. Select which variable(s) you wish to have as the row variable(s).
3. Select which variable(s) you wish to have as the column variables(s).

TIP!

The minimum specification for Crosstabs is one row variable and one column variables. It is somewhat arbitrary as to which variable is placed in the row or column.

Exercise: Creating a crosstabs
Let's create a crosstabs to show us the age distribution of the 2 genders.

RESULTS FROM CROSSTABS USING ROW PERCENTAGE

- There are 6 females between the ages of 25-34 years of age.

agegroup * Respondent sex Crosstabulation

Count

		Respondent sex		
		Male	Female	Total
agegroup	15-24	46	45	91
	25-34	9	6	15
	35-44	7	13	20
	45-54	10	8	18
	55-64	13	11	24
	65-74	3	1	4
	75-84	1	0	1
Total		89	84	173

Crosstab Cell Display

To help you uncover patterns in the data that contribute to a significant chi-square test, the Crosstabs procedure displays expected frequencies and three types of residuals (deviates) that measure the difference between observed and expected frequencies. Each cell of the table can contain any combination of counts, percentages and residuals selected.

Counts

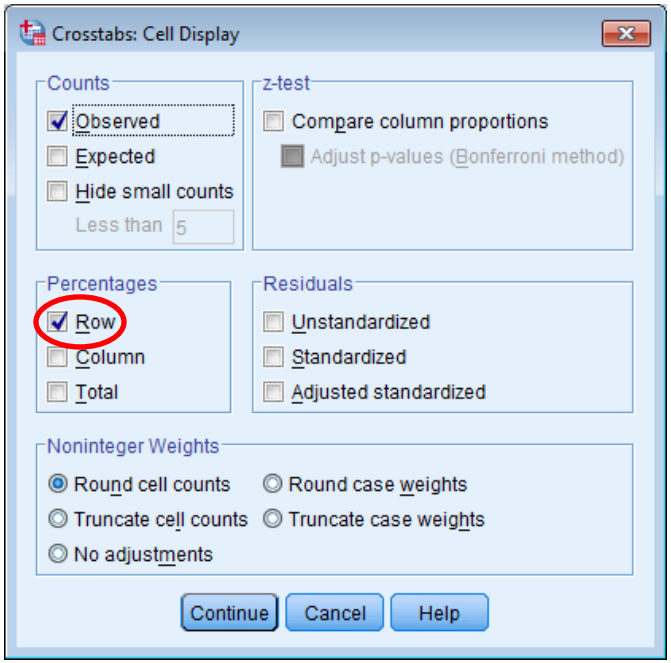
The number of cases actually observed and the number of cases expected if the row and column variables are independent of each other.

Percentages

The percentages can add up across the rows or down the columns. The percentages of the total number of cases represented in the table (one layer) are also available.

Generating a Crosstabulation

1. Click on `Analyze>Descriptive Statistics>Crosstabs`.
2. Then click on the `Cell` button.
3. Check off the row checkbox.



Results from Crosstabs Using Row Percentage

- Of all the individuals between the ages of 25-34 years, 40% are females.

agegroup * Respondent sex Crosstabulation

			Respondent sex		
			Male	Female	Total
agegroup	15-24	Count	45	43	88
		% within agegroup	51.1%	48.9%	100.0%
	25-34	Count	9	6	15
		% within agegroup	60.0%	40.0%	100.0%
	35-44	Count	6	12	18
		% within agegroup	33.3%	66.7%	100.0%
	45-54	Count	10	8	18
		% within agegroup	55.6%	44.4%	100.0%
	55-64	Count	13	11	24
		% within agegroup	54.2%	45.8%	100.0%
	65-74	Count	3	1	4
		% within agegroup	75.0%	25.0%	100.0%
	75-84	Count	1	0	1
		% within agegroup	100.0%	.0%	100.0%
Total		Count	87	81	168
		% within agegroup	51.8%	48.2%	100.0%

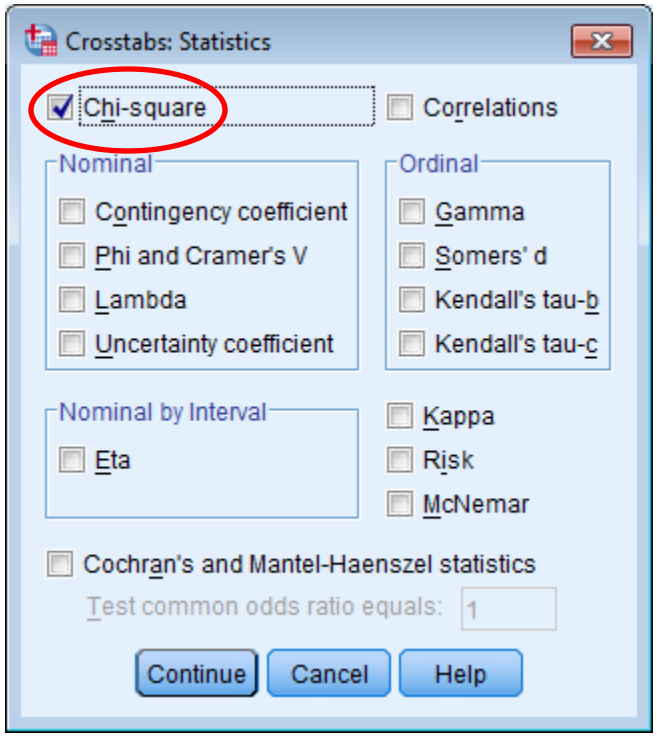
Significance Testing for Crosstabs

The Chi-Square is used to test whether the relationship between two cross tabulated variables is significant. The Chi-square is based on two assumptions. Firstly, the individual observation must be independent of each other. Secondly, the expected frequencies should be greater that 5. In a larger table, not more than 20% of then variables can have expected frequencies less than 5.

For the Chi-Square, the null hypothesis that the row variable is unrelated (that is, only randomly related) to the column variable. The alternative hypothesis is not rejected when the variables have an associated relationship.

Generating a Chi-Square for a Crosstabulation

1. Click on `Analyze>Descriptive Statistics>Crosstabs`.
2. Then click on the `Statistics` button.
3. Check off the Chi-square checkbox.



Results from Significance Testing

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. For example, create new groupings or re-examine your choice of statistical test.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.826[a]	6	.566
Likelihood Ratio	5.296	6	.506
Linear-by-Linear Association	.358	1	.550
N of Valid Cases	168		

a. 4 cells (28.6%) have expected count less than 5. The minimum expected count is .48.

Exercise: Creating a crosstabs
Let's create a crosstabs to show us the marriage distribution of the 2 genders.

RESULTS FROM SIGNIFICANCE TESTING

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

Grouped marital status of respondent * Respondent sex Crosstabulation

			Respondent sex		
			Male	Female	Total
Grouped marital status of respondent	Common-law/Married	Count	31	28	59
		% within Grouped marital status of respondent	52.5%	47.5%	100.0%
	Widow/Divorced/Separated	Count	5	8	13
		% within Grouped marital status of respondent	38.5%	61.5%	100.0%
	Single	Count	50	45	95
		% within Grouped marital status of respondent	52.6%	47.4%	100.0%
Total		Count	86	81	167
		% within Grouped marital status of respondent	51.5%	48.5%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.959[a]	2	.619
Likelihood Ratio	.965	2	.617
Linear-by-Linear Association	.006	1	.939
N of Valid Cases	167		

UNIVARIATE ANALYSIS

(Adapted from Univariate Analysis and Normality Test Using SAS, SPSS and STATA by Hun Myoung Park)
(http://www.indiana.edu/~statmath/stat/all/normality/normality.pdf)

Descriptive statistics provide important information about variables. Mean, median, and mode measure the central tendency of a variable. Measures of dispersion include variance, standard deviation, range, and interquartile range (IQR). Researchers may draw a histogram, a stem-leaf plot or a box plot to see how a variable is distributed.

Statistical methods are based on various underlying assumptions. One common assumption is that a random variable is normally distributed. In many statistical analyses, normality is often conveniently assumed without any empirical evidence or test. But normality is critical in many statistical methods. When this assumption is violated, interpretation and inference may not be reliable or valid.

There are two ways of testing normality (Table 1). Graphical methods display the distributions of random variables or differences between an empirical distribution and a theoretical distribution (e.g. the standard normal distribution). Numerical methods present summary statistics such as skewness and kurtosis, or conduct statistical test of normality. Graphical methods are intuitive and easy to interpret, while numerical methods provide more objective ways of examining normality.

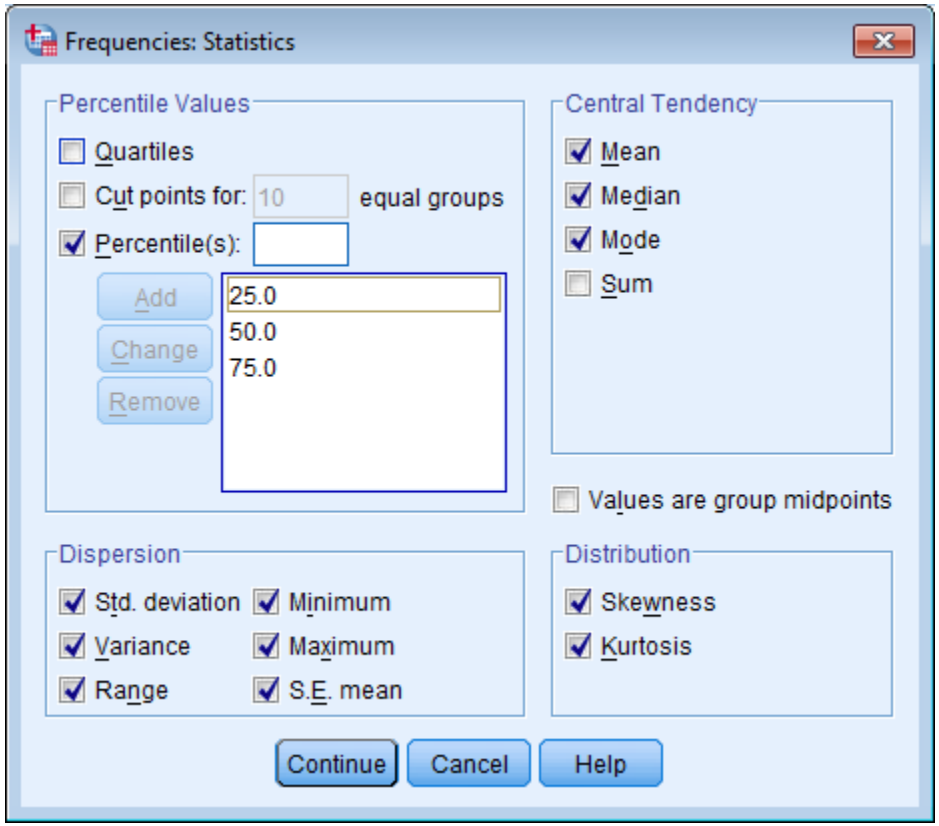
	Graphical Methods	**Numerical Methods**
Descriptive	stem-leaf plot box plot histogram	skewness kurtosis
Theory Drive	P-P plot Q-Q plot	Shapiro-Wilk, Shapiro-Francia, Kolomogorov-Smirnov, Jarque-Bera, Skewness-Kurtosis, (Lillefors) Anderson-Darling/Cramer-vonMises

Graphical and numerical methods are either descriptive or theory –driven. The dot plot and histogram, for instance, are descriptive graphical methods, while skewness and kurtosis are descriptive numerical methods. The P-P and Q-Q plots are theory driven graphical methods for testing normality, whereas the Shapiro-Wilks and Jarque-Bera tests are theory-driven methods.

TESTING NORMALITY USING SPSS

SPSS has the DESCRIPTIVES and FREQUENCIES commands to produce descriptive statistics. DESCRIPTIVES is usually applied to continuous variables, but FREQUENCIES is also able to produce various descriptive statistics including skewness and kurtosis. The CHART BUILDER command draws histograms and box plots. The EXAMINE command can product both descriptive statistics and various plots, such as a stem-leaf plot, histogram, box plot, P-p plot and Q-Q plot. This command is able to draw the detrended Q-Q plot that SAS and STATA do not support. EXAMINE also performs the Kolmogorov-Smirnov and Shapiro-Wilk tests for normality.

For this example we will be looking at the variable totcigs (total number of cigarettes smoked in a week). The first step is to generate summary statistics using `Analyze>Frequencies`.



RESULTS FROM THE FREQUENCIES

The variable totcigs has a mean of 79.71 and variance of 4098.8. The kurtosis is 0.526 and the skewness is 0.904.

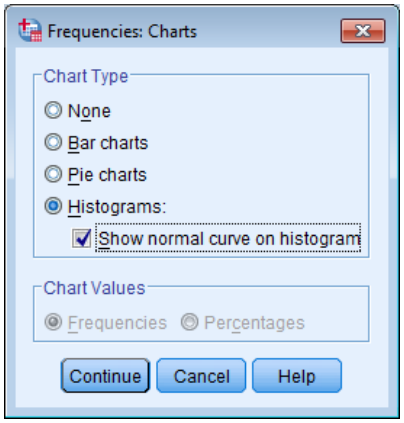
Statistics

totcig

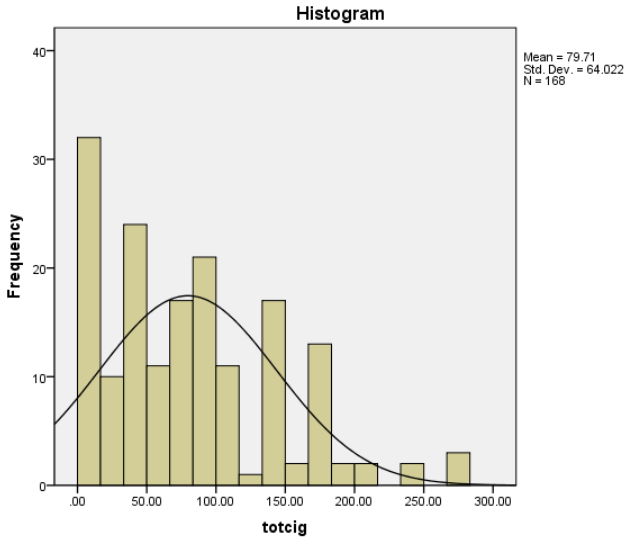
N	Valid	168
	Missing	0
Mean		79.7143
Std. Error of Mean		4.93940
Median		70.0000
Mode		84.00
Std. Deviation		64.02190
Variance		4098.804
Skewness		.904
Std. Error of Skewness		.187
Kurtosis		.526
Std. Error of Kurtosis		.373
Range		280.00
Minimum		.00
Maximum		280.00
Percentiles	25	32.0000
	50	70.0000
	75	118.0000

GRAPHICAL METHODS

A histogram is the most widely used graphical method. The histogram option with Frequencies Dialog (`Analyze>Frequencies`) box allows you to add a normal density curve to the histogram. A histogram can also be generated using the `Graphs>Chart Builder`.



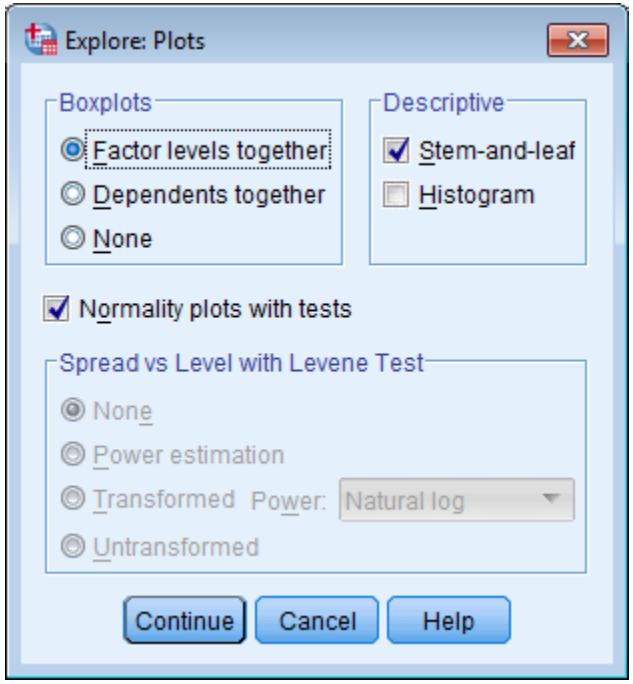
The histogram that is generated suggests that the variable is not normally distributed.



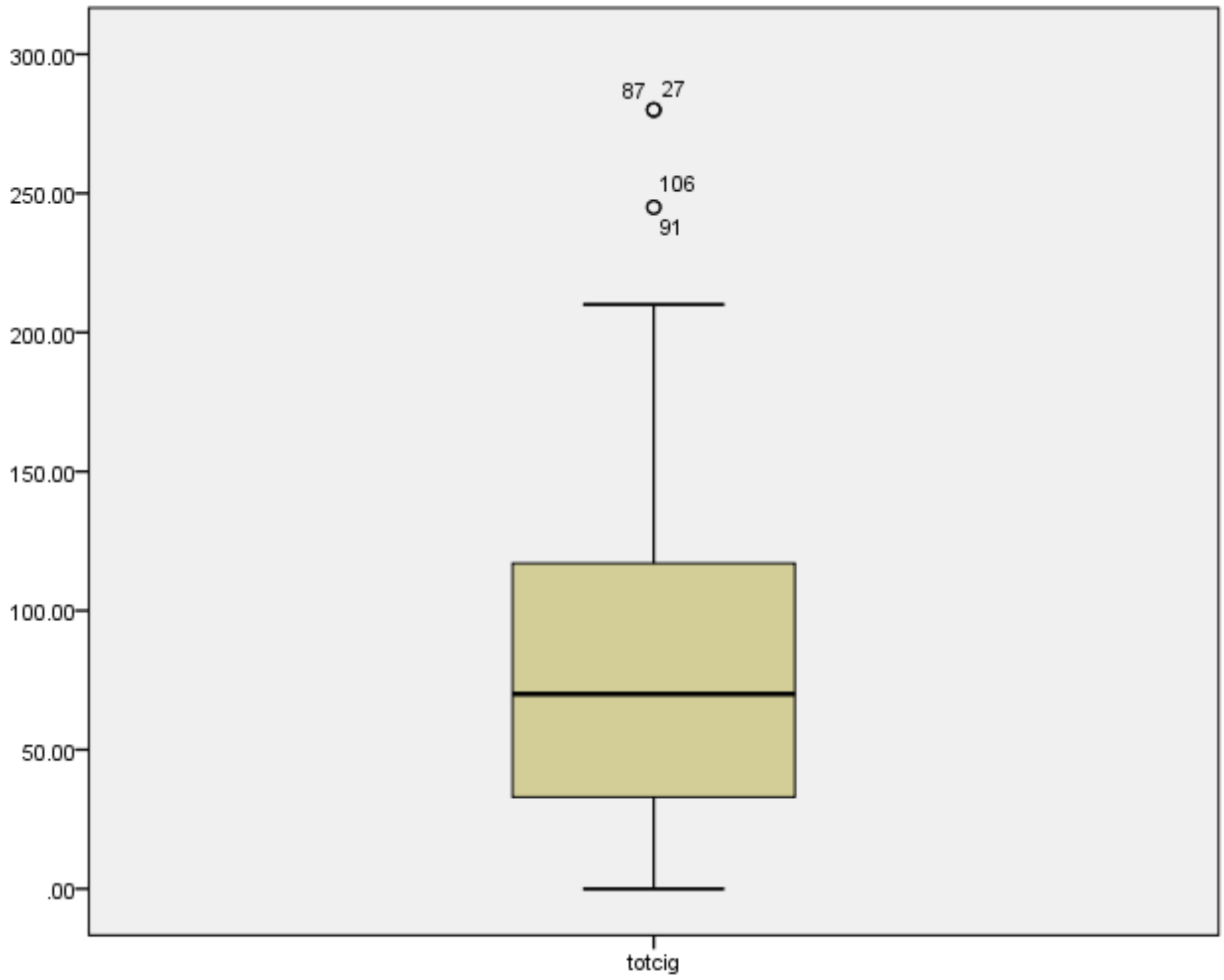
BOX PLOT

A box plot represents the minimum, 25th percentile (1st quartile), 50th percentile (median), 75th percentile (3rd quartile) and maximum in a box and lines. Outliers if any appear at the outsides of the (adjacent) minimum and maximum lines. As such, a box plot effectively summarizes these major percentiles using a box and lines. If a variable is normally distributed, its 25th and 75th percentile are symmetric and its median and mean are located at the same point exactly in the center of the box.

To generate a box plot, click on `Analyze>Descriptives>Explore` and then click on `Plot` button. Then choose Stem and Leaf & Normality plots with tests.



It can be seen that both extremes (i.e. minimum and maximum), the 25th, 50th and 75th percentiles are symmetrically arranged in the box plot.

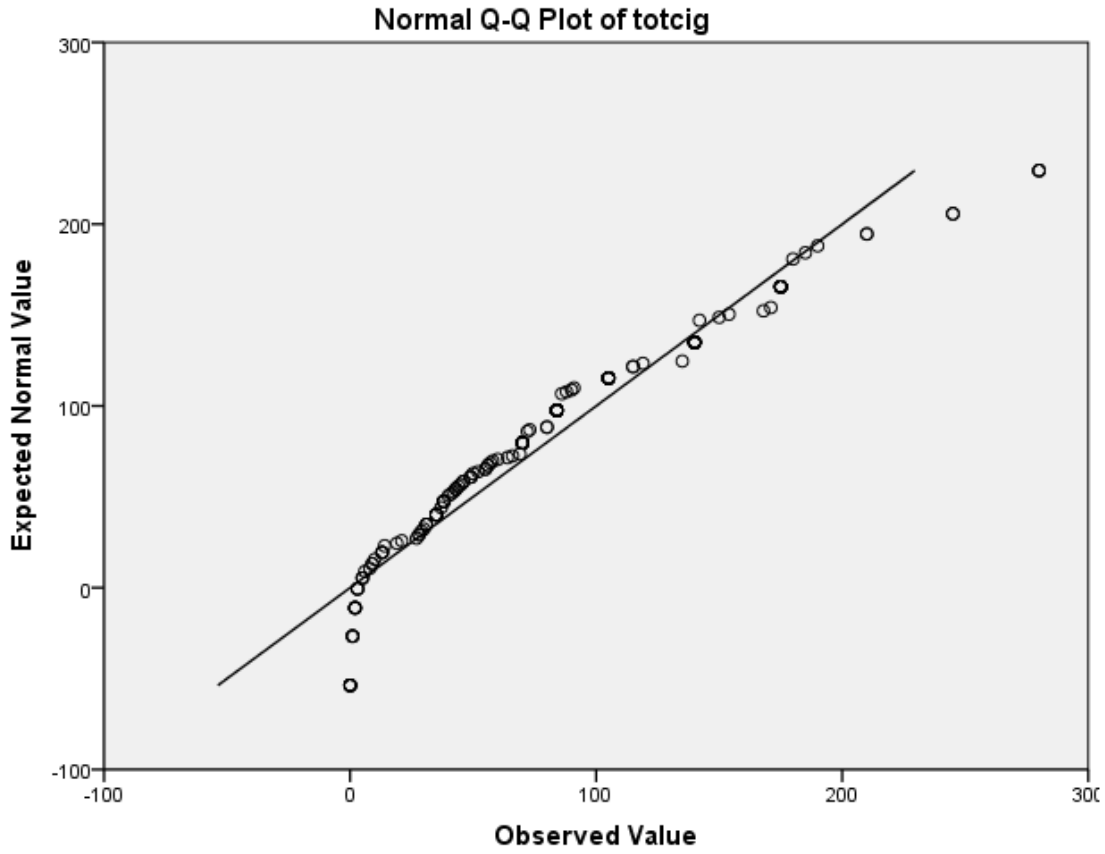


Q-Q PLOT

The quantile-quantile plot (Q-Q plot) compares ordered values of a variable with quantiles of a specific theoretical distribution (ie the normal distribution). If two distributions match, the points on the plot will form a linear pattern passing through the origin with a unit slope. P-P and Q-Q plots are used to see how well a theoretical distribution models the empirical data.

To generate a box plot, click on Analyze>Descriptives>Q-Q plot and then click on the Plot button.

The Q-Q plot indicates that is a significant deviation from the fitted line.



THEORY DRIVEN STATISTICS

Skewness and kurtosis are based on the empirical data. The numerical methods for testing normality compare empirical data with a theoretical distribution. Widely used methods include the Kolmogorov-Smirnov (K-S) D test (Lilliefors test), Shapiro-Wilk test, Anderson-Darling test and Cramer-von Mises test. The K-S, D test, and Shapiro-Wilk W test are commonly used.

To generate Test of Normality, click on `Analyze>Explore`.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
totcig	.134	168	.000	.922	168	.000

a. Lilliefors Significance Correction

Since N is less than 2,000, we have to read the Shapiro-Wilk statistic that does reject the null hypothesis of normality ($p<0.05$).