

SPSS Exercise 1 : Basics

SPSS is a program (application) that allows you to record, manage and store data, as well as to perform a variety of analyses on that data. This exercise is a brief review of statistical concepts you should be familiar with before we start using SPSS. In addition, we will also get a brief tour of SPSS and get a feel for the application's environment.

Types of Data

Measurement is the process of attaching values or labels to observations. When we do this, we do so using some type of measurement scale. It's important to know what scale of measurement you are using, since it will limit what you can do, and the types of questions you can answer with it.

- **Nominal** - Most limiting of all of the scales, it is a scale often used in Social Sciences. Nominal scales allow us to measure differences, but not direction or magnitude. For example, "RED" and "BLUE" both describe a difference (in color for this example), but "35 mph" and "50 mph" not only describes a difference (in speed), but how much of a difference (magnitude). Also, "50 mph" is faster than "35 mph", so it measures "more or less" (direction). So measuring color using terms like "RED" and "BLUE" is an example of a NOMINAL scale, while measuring speed in MPH would not be. Thus, NOMINAL scales allow us to categorize, but not to order or quantify observations. Examples: Sex, Ethnicity, Religion, Language.
- **Ordinal** - Measures rank ordering. More versatile than nominal, but does not tell us anything about the relative amount of a difference, just the direction of it. For example, In the Olympics they give out "gold", "silver" and "bronze". All it tells us is who were the best, second best, and third best, but not by how much. Say the *gold* medalist scored a 10.0, the *silver* medalist a 9.9 and *bronze* a 2.1 ! The difference between the three medals is only one step on our *gold* >> *silver* >> *bronze* scale, but it says nothing about the amount of difference between them. So ORDINAL scales allow you to measure a difference, and the direction of that difference, but not its magnitude. Examples: any rank-ordering, movie ratings, Income level, likert-type scales.
- **Interval** - More useful than either nominal or ordinal, interval scales allow us to measure not only the direction of a difference, but also its magnitude. The basic criteria that makes a scale an interval scale is that it has equal intervals between values, so the difference between a value of 5 and a value of 10 on some scale is the same as the difference between a value of 7 and a value of 12, that would be 5 ($10-5=5$; $12-7=5$) units. In the previous scale description (Ordinal) we used the example of Olympic medals. In that case the difference between "gold" and "silver" was not the same as between "silver" and "bronze". That's why it was only Ordinal, because the difference between the measurement points was not the same across the scale. Say you measured the temperature of an engine in degrees Fahrenheit. A difference in temperature between an engine running at "120 degrees" and one running at "140 degrees" is the same as the difference between an engine running at "100 degrees" and one running at "80 degrees". That is a difference of 20 degrees. So an INTERVAL scale allows you to measure three things now, a difference, the direction of that difference, and the magnitude of that difference. Examples: temperature, some psychological tests.

- **Ratio** - The best scale in terms of the range of questions it can answer. Ratio scales are defined by their having a true zero, that is to say that a value of 0 truly represents none of whatever you are measuring. For example, how many children someone has is a RATIO value. It tells us about differences (4 is not the same as 1 child), tells us about direction (4 is more than 1 child) tells us about magnitude (4 is 3 more than 1 child) and it also tells us that since 0 means no children (it's a true zero) we can make ratio statements with our data. If Joe has 4 children and Sue has 2, Joe has twice ($4/2=2$) as many. That's a ratio! If somebody in a TV commercial claims their pain reliever is twice as strong as the competitor's, you should wonder if their data is Ratio. It's the only way you could make such a statement (at least honestly). Examples: Reaction times, many demographics (number of children), physical measures (height), frequencies.

Descriptive Statistics

Statistics fall broadly into two categories, descriptive and inferential. We will consider three types of descriptive statistics: frequencies, measures of central tendency, and measures of variability.

Frequencies - Tell us how many. If you can count, you know about frequencies. They can be represented in tabular form, or graphical form. The normal curve is a good example of what a frequency distribution looks like when graphed. In tables, they tell us how many cases had a particular value (Example: How many people who took a survey checked their sex as female).

Measures of Central Tendency - Tell us about typical/average/common values.

- **Mean** - Average or Arithmetic Average. When you use the word "average" this is what you mean (no pun intended).
- **Median** - Middle value. Half of the records have higher values, half have smaller values. Also known as the 50th percentile (P_{50}). In a family with 5 children, the middle one would be the Median.
- **Mode** - Most common value. If in a sample you had 3 males and 2 females, male would be the mode. There are more of them.

Measures of Variability - Tell us about the range of our values. Are they similar, or very different?

- **Range** - Difference between the highest and the lowest values.
- **Standard Deviation** - The average difference from the mean. Say you had the ages for 10 people. If you calculated a Mean for their ages, and then calculated how far from that Mean each age was, and calculated the Mean for those differences, you would have the Standard Deviation of those ages. It is a wee bit more complicated, but that's the general idea. The bigger the number, the more spread out your values are.
- **Variance** - The squared Standard Deviation. Needed for some statistical tests. So if the Standard Deviation is 3, the Variance is 9 ($3 \times 3 = 9$). The term is also used in a broader sense to mean variability.

Inferential Statistics

Inferential statistics allow us to make inferences about a population based on a sample. Let's say you want to know who is going to win the next election, Mr. X, or Ms. Y. You could ask all the voters, but that's too hard and not necessary. You could get a sample of voters, ask them, and infer the results of the election based on your sample's responses.

All inferential statistics measure the probability of a particular outcome. Say our voter sample was made out of 100 randomly selected people, and 50 said they would vote for X and 50 would vote for Y. You can see that no winner could be predicted, since there was no winner in our sample. On the other hand, if all 100 voters in our sample say they will vote for Y, making a prediction is quite easy. If, however, 49 voters say X and 51 say Y, we again can't predict a winner because the difference is not large enough to suggest there is a real difference in the population our sample came from. This is the very important concept of *significance*. As the differences increase, the likelihood that they are not just a fluke of our sample, but a real difference in the population we drew the sample from, increases. It would be unlikely to randomly draw a bunch of people who all preferred one candidate from a population where support was split between two candidates evenly.

Significance

Central to interpreting the results of inferential tests, *significance* in statistics refers to rarity and not to importance or magnitude as the term is commonly used in everyday speech.

Imagine you have a container filled with red and green M&Ms. The M&Ms in this container represents a population. You want to see if there is the same number of red M&Ms as green ones. Instead of counting each M&M, you mix them up really well and randomly take out 10.

Say we get 5 Red, 5 Green

What is the probability of getting 5 of each color, if there were in fact the same number in the container? Well, pretty good actually. If the container had just as many red as green M&Ms, grabbing 10 randomly which were 5 red and 5 green would not be unlikely.

Say we get 0 Red, 10 Green

What is the probability of getting no red M&Ms, if there were in fact the same number of each color in the container? Low. If the container had just as many red as green M&Ms, grabbing 10 randomly which were all green would be unlikely.

As you can see, the bigger the difference between the number of red and green M&Ms in our sample, the lower the probability that the sample could have been drawn from a population with half red and half green M&Ms. The closer the counts are, the higher the probability that the sample could have been drawn from a population that was half of

each color. This probability is referred to as a ***p value***, and it is used to determine the significance of an inferential statistic.

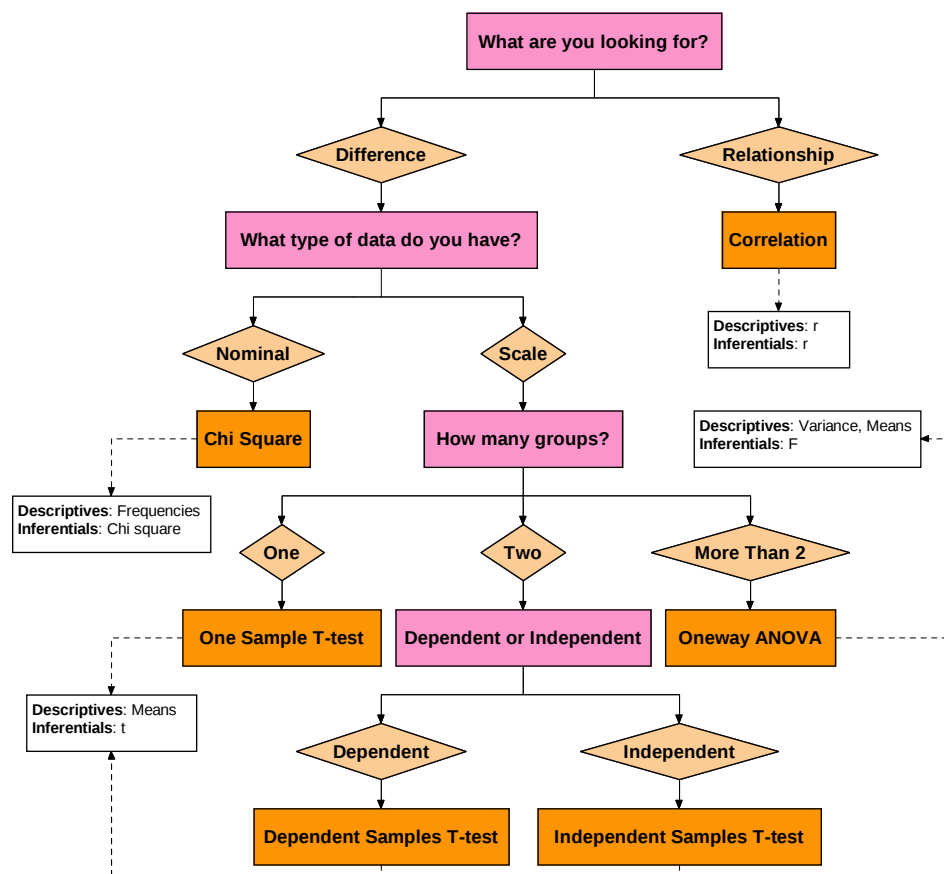
At some point, we draw a line and say something like "6 of one color and 4 of the other is not that unlikely (high p value), even 7 of one color and 3 of the other is not too unlikely (medium p value), but 8 or more of one color would be too unlikely (low p value) to come from a container that had equal amounts of red and green M&Ms."

Where we draw that line is called the ***level of significance***, that is, the level of the p value required to report your results as *significant*. **In this class, our level of significance will always be .05.**

When p values are **below (less than)** our level of significance (**$p < .05$**), we can claim that our results are *significant*.

The Tests

Different types of data, and different types of questions require different descriptive and inferential statistics. Inferential statistics most commonly use some aspect of a descriptive statistic (difference in frequencies, for example), and calculates one or more values that can be used to determine a p value. In other words, what is the probability that we would find the differences we find in our sample (hand-full of M&Ms) if there were really no differences in the population we were really interested in (bowl of M&Ms). The following image illustrates the inferential tests we will be working with, and the associated statistics.



Starting the Application

You will most likely find an icon on your desktop for SPSS. It looks something like this. Otherwise, you can start SPSS through "Start/All Programs/SPSS inc/SPSS 16.0/ SPSS 16.0".



Entering Data

Participant ID	
Sex	
Political Party	
Age	

Participant ID	
Sex	
Political Party	
Age	

Participant ID	
Sex	
Political Party	
Age	

Participant ID	
Sex	
Political Party	
Age	

Participant ID	
Sex	
Political Party	
Age	

We will be using three terms to refer to different aspects of our data: *variable*, *record* and *value*. As an example, let's assume we have given the brief survey above to 5 people.

- **Variable** - Our data will consist of 4 *variables*, Participant ID, Sex, Political Party and Age.
- **Record** - We will have 5 *records*, one for the responses from each survey.
- **Value** - The individual responses on the survey.

Assume this is how our surveys were returned. Participant ID is coded by us to identify the survey.

Participant ID	<u> 1 </u>
Sex	<u> M </u>
Political Party	<u> Dem </u>
Age	<u> 25 </u>

Participant ID	<u> 2 </u>
Sex	<u> F </u>
Political Party	<u> Dem </u>
Age	<u> 30 </u>

Participant ID	<u> 3 </u>
Sex	<u> F </u>
Political Party	<u> Rep </u>
Age	<u> 21 </u>

Participant ID	<u> 4 </u>
Sex	<u> M </u>
Political Party	<u> Ind </u>
Age	<u> 23 </u>

Participant ID	<u> 5 </u>
Sex	<u> F </u>
Political Party	<u> Rep </u>
Age	<u> 29 </u>

Our first **record** then would consist of the **values** entered for each **variable** by participant "A".

Participant ID = 1, Sex = M, Political Party = Dem, Age = 25

To enter our data into SPSS, we will need to identify the scale of measurement for each variable. Since Ratio and Interval scaled data are treated similarly, SPSS does not differentiate between ratio and interval scales, and groups them into a category called "Scale". In addition, SPSS prefers numbers, so we will have to code some variables.

Participant ID - Nominal

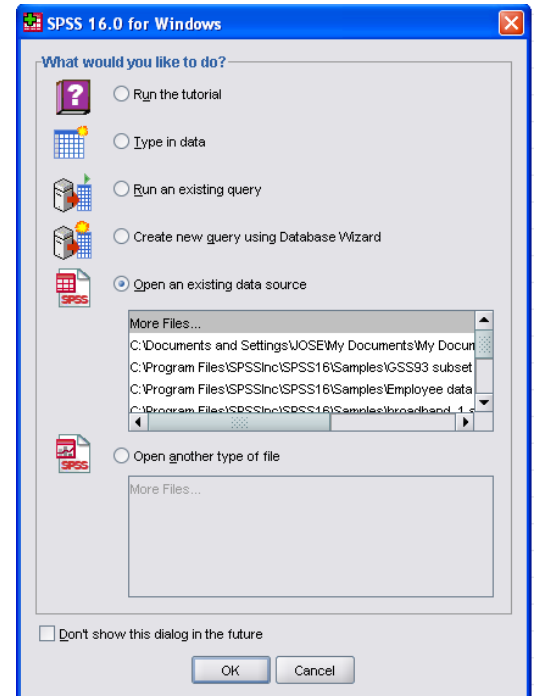
Sex - Nominal (Code as: M=1, F=2)

Political Party - Nominal (Code as: Dem=1, Rep=2, Ind=3)

Age - Scale

If you encountered the screen shown on the right when you started SPSS, select "Type in Data" and click OK.

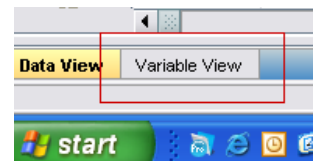
Otherwise, enter your data as shown below. Note that each row is a **record**, and each column is a **variable**.



SPSS Data Editor - Variable View

	VAR00001	VAR00002	VAR00003	VAR00004
1	1.00	1.00	1.00	25.00
2	2.00	2.00	1.00	30.00
3	3.00	2.00	2.00	21.00
4	4.00	1.00	3.00	23.00
5	5.00	2.00	2.00	29.00
6				
7				

Once our data is entered, we should define our variables in the Variable View screen. To begin, click on the "Variable View" tab (image on right) on the bottom left-hand corner of the screen. Your screen should then look like the image below.



SPSS Data Editor - Variable View

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	VAR00001	Numeric	8	2		None	None	8	Right	Scale
2	VAR00002	Numeric	8	2		None	None	8	Right	Scale
3	VAR00003	Numeric	8	2		None	None	8	Right	Scale
4	VAR00004	Numeric	8	2		None	None	8	Right	Scale
5										
6										

Here you can define several attributes for each variable. They are:

- **Name** – The name of your variable. Up to 8 characters, no weird stuff like spaces or & or #.
- **Type** – Numbers or text

- **Width** – Number of characters in field
- **Decimals** – How many decimal places should be used
- **Label** – Here you can attach an easy-to-read label (see below)
- **Values** – You can add labels for coded values (see below)
- **Missing** – What values define missing values
- **Columns** – How wide the column should show in “Data View” mode
- **Align** – Cell alignment in “Data View” mode
- **Measure** – Measurement scale (see previous discussion on measurement scales)

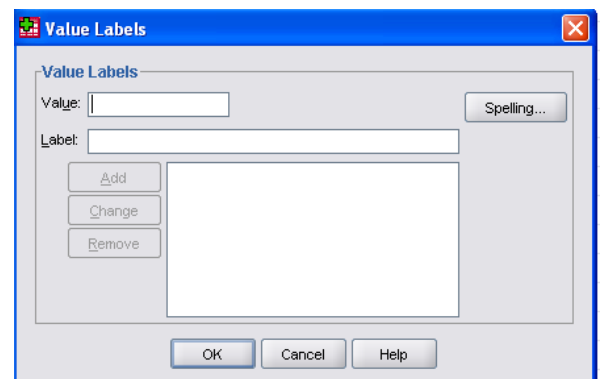
SPSS lets you attach a label to a variable to make your results more readable. Variable names are limited to 8 characters, cannot have spaces or special characters, and cannot start with a number. To enter a variable name, replace "var00001" in the Name column with "ID", enter "Participant ID" for the Label, and select "Nominal" from the drop-down in the Measure column, to specify the measurement scale.

Since this variable is only an identifier, there is nothing else we need to specify for it. Sex, however, is a coded variable. In other words, when we entered a "1" we really meant

"Male", so we need to tell SPSS what each value really means. Start by replacing "var00002" with "sex" in the Name column, and enter "Participant Sex" in the Label column. In the "Values" column, click on the small button on the right of the field (...). You should get the screen on the right.

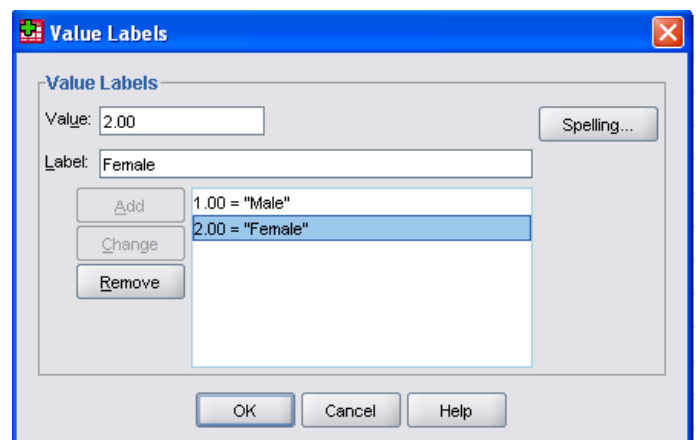
We would like to attach the label "Male" to the value "1", and "Female" to the value "2". To do this, enter "1" into the "Value" field, and "Male" into the "Label" field, and click the "Add" button.

The window should look like the image below.



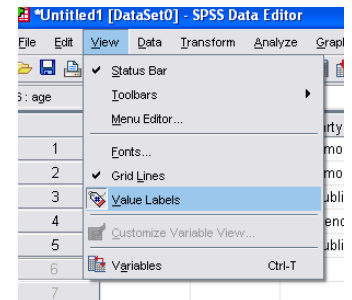
1 was 25 instead of the average age for Male was 25. So this will make things a lot easier to read later. The window should look like the one on the right. Click "OK". Proceed to define the other two variables (Political Party and Age) in the same way. Use the following values:

Now enter "2" into the "Value" field, and "Female" into the "Value Label" field, and click the "Add" button. The reason for this will become more apparent later when we see the results of some analysis. Imagine we left sex unlabeled, we would find ourselves looking at a report that said that the average age for



Variable Name	Variable Label	Measurement	Value Labels
party	Political Party	Nominal	1 = "Democrat" 2 = "Republican" 3 = "Independent"
age	Participant Age	Scale	None

When you are done, go back to the Data View screen (tab on lower left), your screen should look like one of the screens below. Note that the difference is that in one case it shows us the values we entered, and in the other it shows us the Labels we associated with those values. So in one it says 1, in the other 1 is translated into its Label "Male". To toggle back and forth between the two views, go to "View" on the top menu, then click on "Value Labels" to turn the checkmark on and off.



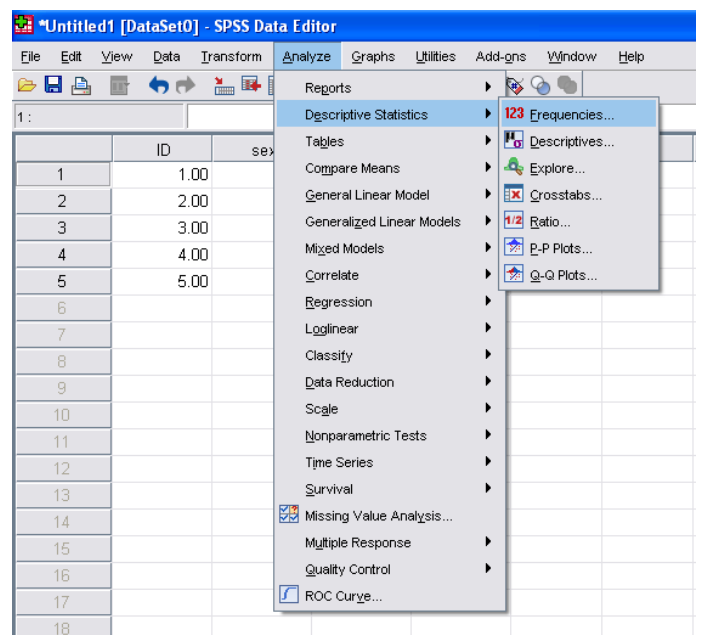
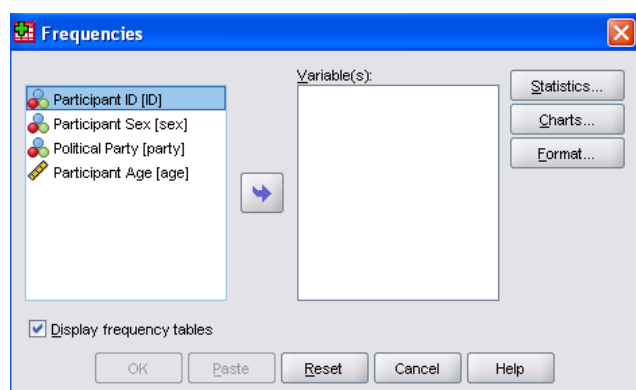
	ID	sex	party	age
1	1.00	Male	Democrat	25.00
2	2.00	Female	Democrat	30.00
3	3.00	Female	Republican	21.00
4	4.00	Male	Independent	23.00
5	5.00	Female	Republican	29.00
6				

	ID	sex	party	age
1	1.00	1.00	1.00	25.00
2	2.00	2.00	1.00	30.00
3	3.00	2.00	2.00	21.00
4	4.00	1.00	3.00	23.00
5	5.00	2.00	2.00	29.00
6				

Save your file.

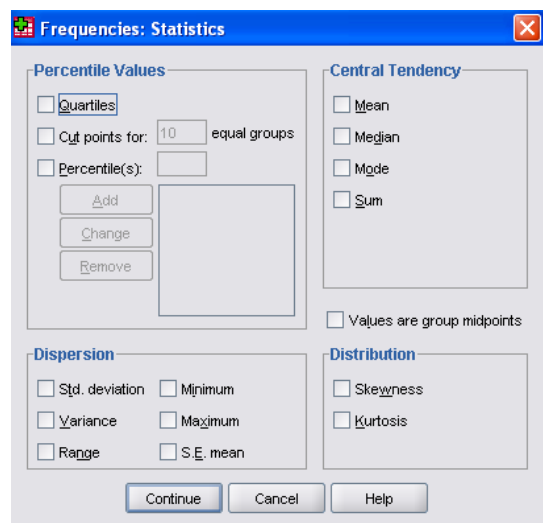
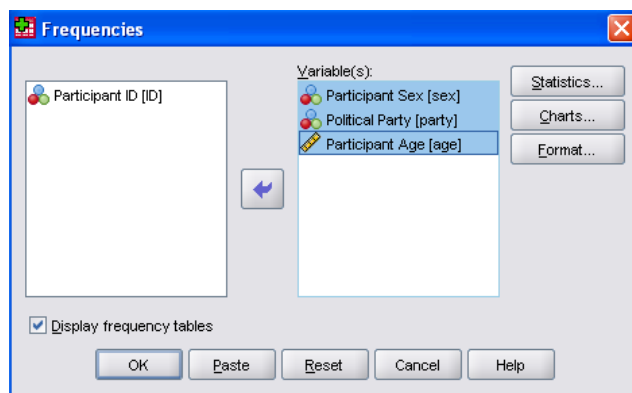
Using SPSS to calculate Descriptive Statistics

In the "Analyze" menu you will find "Descriptive Statistics". In its sub menu, select "Frequencies". You should see a window like the one below.



Select "Participant Sex", "Political Party" and "Participant Age" in the left-hand box and move them to the "Variable(s)" box by clicking on the right-pointing arrow between the boxes. You can select more than one variable at a time by either alt-clicking on each one, or shift-clicking to select a range of them. Check that the "Display frequency tables" check-box is checked. The window should then look like the one on the right.

Click on the "Statistics..." button. You should get the window below.



Check "Mean", "Median", "Mode", "Std. deviation" and "Range". Click "Continue", then click "OK". A new window will open when SPSS is done with the calculations. SPSS has several modes, the main two are the SPSS Data Editor and the SPSS Viewer. The spreadsheet you entered our data into and defined the variables in is the SPSS Data Editor. The window that just opened up with your results is the SPSS Viewer.

Look through the results and see if you can answer the assignment questions. They are easy enough to answer without SPSS, so use that information to help you find the answers within the SPSS results.

Assignment

For this first assignment there will be no "for points" component. Go to the web site, and on the Resources page you will find a link to Assignment 1 – For Practice & Points. Answer those questions using the results we calculated.

IMPORTANT: The online assignment will grade your answers and give you a score. In ALL of these assignments, when entering decimals (and not prompted as to how many to enter), be sure to round off to three decimal places if SPSS gives you more than 3. For simplicity, assume .0005 rounds up, so $4.0005 = 4.001$. Also, do not enter "%" signs when asked for a percentage, just the value.