

Spring 2009
Part IV: SPSS III - Regression and Graphing
Skill H

Objectives:

- 1) To perform a linear regression in SPSS
- 2) To learn about basic graphing functions in SPSS

Populations:

We will use the data set that we created for the 2^3 factorial ANOVA, 3 factors, each with 2 levels ($2 \times 2 \times 2 = 8$):

Factor 1: Brand

Orville Redenbacher Butter
Pop Secret

Factor 2: Time

1 minute 15 seconds (1.25)
1 minute 30 seconds (1.50)

Factor 3: Microwave

Dr. Burks' microwave
4th floor Student Lounge

Questions:

- 1) Can you predict the number of popped kernels by the number of leftover kernels?
- 2) What are the guidelines for graph making?

What we will do (numbers correspond to the questions):

- 1) Linear Regression
- 2) Graphing

Import Data from Excel to SPSS:

Open up the data file that you saved (SPSS2dataNAME.sav).

About linear regression:

Regression attempts to describe the dependence of a variable on one (or more) explanatory variables; it implicitly assumes that there is a one-way causal effect from the explanatory variable(s) to the response variable, regardless of whether the path of effect is direct or indirect. It is used as a measure of prediction.

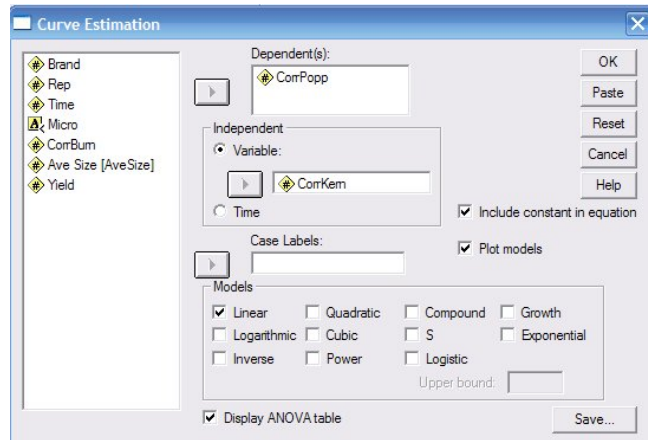
To conduct a regression in SPSS, there are two ways. We will use the way that has more built-in flexibility. We will focus on linear regressions but there are also other relationships that you may want to fit to data (logarithmic, exponential, etc...).

In our case, we would expect a negative relationship between the number of leftover kernels and the number of popped kernels. In other words, more leftover kernels would suggest less popped kernels.

A regression gives you two outputs: 1) a p-value that tells you whether or not the slope of the line is significantly different than 0; and 3) a R^2 value that tells you the extent to which one variable accounts for the variance found in the other. Be careful not to confuse R^2 with Pearson's r that has to do with correlation. The closer the value of R^2 to 1, the better relationship exists. If every data point fell exactly on the line, only then would you have $R^2 = 1$. In ecology, values > 0.80 usually represent strong relationships.

Objective 1 Regression Directions:

- 1) Choose Analyze
- 2) Choose Regression
- 3) Choose Curve Estimation
- 4) Note that the Linear Box is checked as the default and that you could do other types of regression, even simultaneously.
- 5) Add "Corr Popped" into Dependent Variable (it is what you want to predict)
- 6) Add "Corr Kernels" into Independent Variable (it is what you think will explain the data)
- 7) Keep the checked box for plot models and a graph of the relationship will automatically appear
- 8) Add the checked box for display ANOVA so that you can see if the slope is significantly different from 0 or not (remember the null hypothesis is that the slope does not differ from 0).



Here is what the output should resemble:

Linear

Model Summary

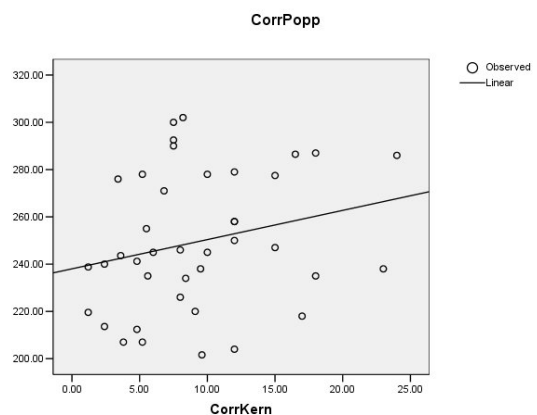
R	R Square	Adjusted R Square	Std. Error of the Estimate
.239	.057	.032	28.511

The independent variable is CorrKern.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	1875.430	1	1875.430	2.307	.137
Residual	30889.898	38	812.892		
Total	32765.328	39			

The independent variable is CorrKern.



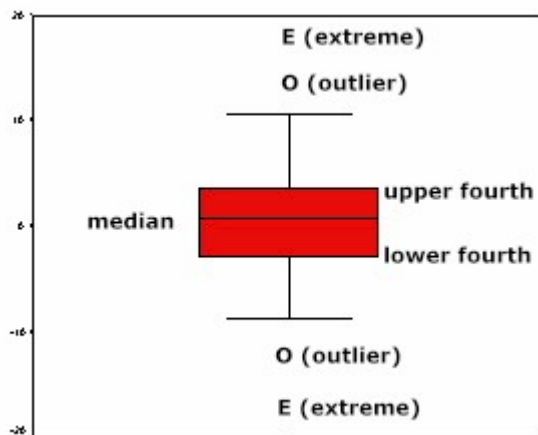
The results indicate that the slope does not differ from 0 ($p = 0.137$) and the number of leftover kernels only explains 5% of the number popped, very very weak result.

Application/Homework: Using your data, see if you come up with a similar result and then perform a regression with two variable that you think should have a predictive relationship.

Objective 2 Graphing Directions:

We will look at only 3 types of graphs in this lab using the 3 data files that you already have saved within SPSS. All of your output should be saved as SPSSOUT4NAME.spo:

- a) Boxplot - for showing a distribution
 - i) Will exhibit yield distributions of ORB and ORSP
 - ii) Data Set = TTESTSPSSNAME.sav
- b) Bar Graph - for comparing means
 - i) Will graph mean yields for ORSP, PSS and AB
 - ii) Data Set = SPSS1DATANAME.sav
- c) Scatter Plot - for showing the relation between two variables
 - i) Will graph relationship between popcorn size and yield by microwave
 - ii) Data set -SPSS2DATANAME.sav



Boxplots:

The basic elements of a boxplot is shown to the left. The top of the box is called the upper fourth. It is at the 75th percentile of the scores. The bottom of the box is called the lower fourth. It is at the 25th percentile of the scores. Therefore 50 % of the scores fall within the box. The interquartile range is the distance between the upper fourth and the lower fourth. The horizontal line through the box represents the median. The ends of the whiskers represent the largest and smallest values that are not

outliers. An *outlier*, O, is defined as a value that is smaller (or larger) than 1.5 box-lengths from the lower fourth (upper fourth). The box-length is defined as the interquartile range. An *extreme* value, E, is defined as a value that is smaller (or larger) than 3 box-lengths from the lower fourth (upper fourth). You should be wary of "outliers" or "extreme" values. Outliers and extreme values will tend to bias statistics that are based on "interval" level data.

To make a boxplot:

- 1) Open TTESTSPSSNAME.sav
- 2) Choose Graphs
- 3) Choose Boxplot
- 4) Select the icon for Simple and select Summaries for groups of cases.
- 5) Select Define.
- 6) Select the variable for which you want boxplots (YIELD), and move it into the Variable box.
- 7) Select a variable for the category axis (TYPE) and move it into the Category Axis box. This variable may be numeric, string, or long string.
 - Note - if you have multiple variables, you can select a variable and move it into the Label Cases by box. This variable can be numeric or string. If selected, the value labels

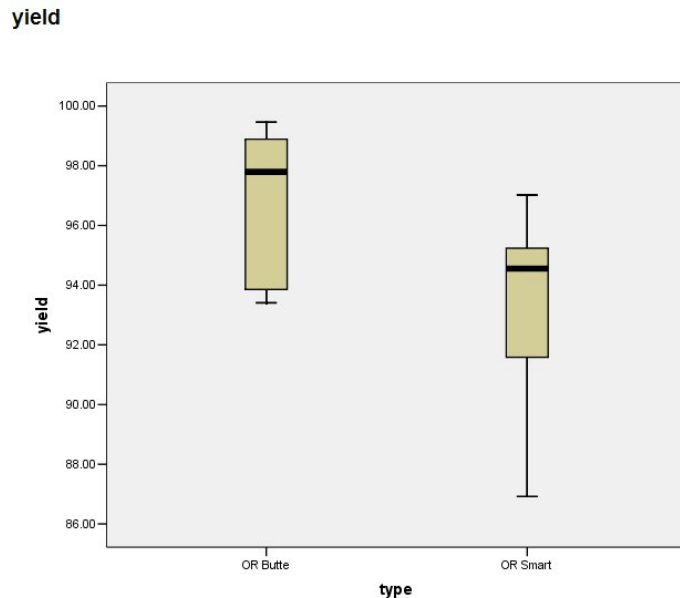
(or values if no labels are defined) of this variable can be used to label Outliers or extreme cases on the plot.

8) Hit OK

The first graph is what SPSS will produce as a standard box plot. However, as you might notice, this graph is not terribly attractive and several formatting steps can be done to make it better.

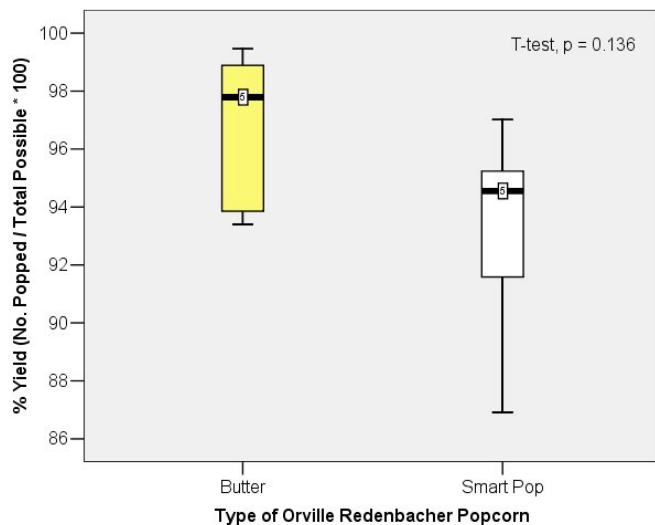
The ability to format graphs in SPSS is very extensive and it would take several pages to illustrate all of the screens. The best practice is just to work with the program. By double clicking on the boxplot, the Chart Editor in SPSS will appear. Then, by double clicking on the

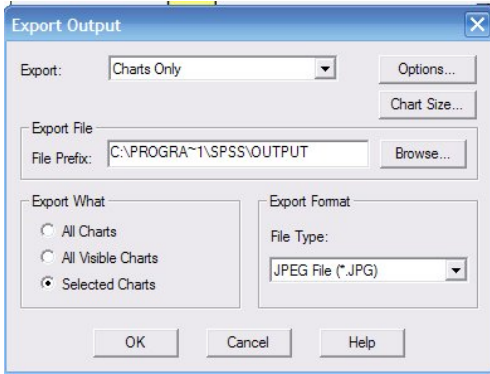
different parts of the graph, you will pull up different formatting options. A blue line will appear around the particular graphical element that you are trying to change. It takes some practice to learn what elements make better graphs, but you can always start with making specific Y and X axis labels and increasing the size of the fonts.



For the above example, I formatted the following:

- 1) Change Y-axis to be more descriptive.
- 2) Increased the size of the Y-axis description (12 pt).
- 3) Increased the size (12 pt) of the axis labels (numbers).
- 4) Changed the X-axis labels to be more descriptive.
- 5) Changed the quartile bar colors to yellow and white to indicate different types of popcorn.
- 6) Added statistical information as a text box.
- 7) Added data labels to indicate number of samples (N = 5).



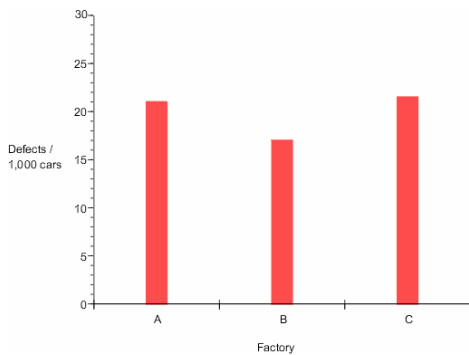


Play around with the boxplot graph until you can get the original to look like the edited. Remember to hit "Apply" and pay attention to what you are formatting. Only when you close the Chart Editor does the boxplot change in your output.

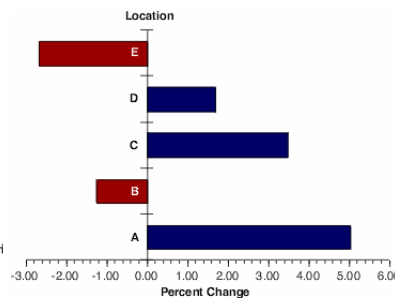
When you finish formatting the boxplot, Export it as a JPEG file by right-clicking on the chart, choosing Export, choosing Chart and then picking a location. Then you will be able to insert this Chart as a Picture into any file. Note - - - you can ONLY format the chart in SPSS. Once you have exited the Chart Editor, you have created the final product. On a separate sheet of paper, write a caption for your boxplot.

Bar Graphs: Good info @ <http://www.ncsu.edu/labwrite/res/gh/gh-bargraph.html>

Unlike the boxplot, which is typically in a standard format, there are many different types of bar graphs. Bar graphs are a very common type of graph best suited for a qualitative independent variable.



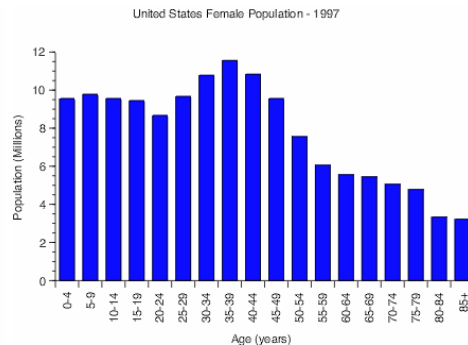
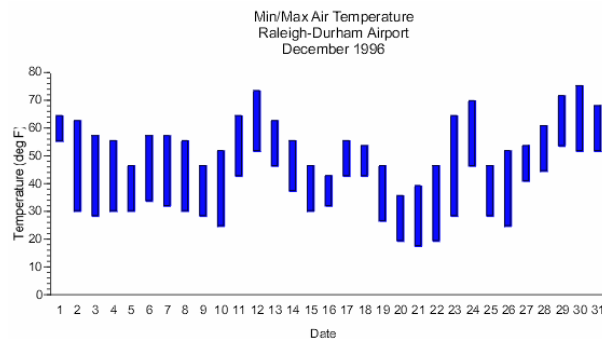
Simple Bar Graph



Horizontal Bar Graph

Range Bar Graph

Histogram



Vertical Bar: Since there is no uniform distance between levels of a qualitative variable, the discrete nature of the individual bars are well suited for this type of independent variable.

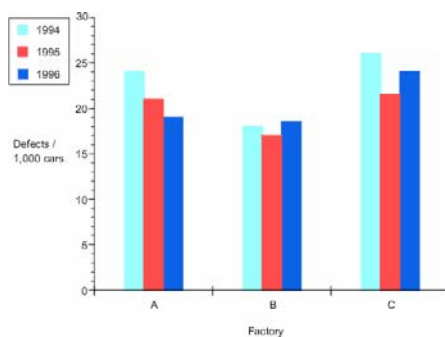
Though you can extract trends between bars (e.g., they are gradually getting longer or shorter), you cannot calculate a slope from the heights of the bars.

Horizontal bar: Bar graphs can be shown with the dependent variable on the horizontal scale. This type of bar graph is typically referred to as a horizontal bar graph. Otherwise the layout is similar to the vertical bar graph. Note in the example above, that when you have well-defined zero point (ratio and absolute values) and both positive and negative values, you can place your vertical (independent variable) axis at the zero value of the dependent variable scale.

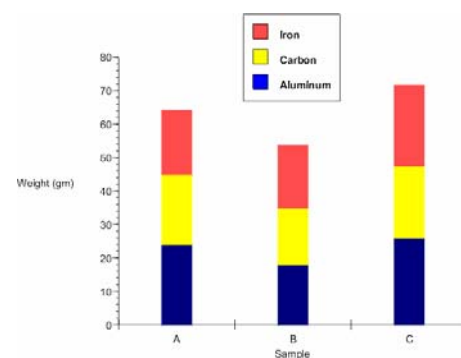
Range bar: Range bar graphs represents the dependent variable as interval data. The bars rather than starting at a common zero point, begin at first dependent variable value for that particular bar. Just as with simple bar graphs, range bar graphs can be either horizontal or vertical.

Histogram: Histograms are similar to simple bar graphs except that each bar represents a range of independent variable values rather than just a single value. What makes this different from a regular bar graph is that each bar represents a summary of data rather than an independent value. For this type of graph, the dependent variable is almost always a scalar scale representing the count, or number, of how many of a sample fall within each range of the independent variable.

Grouped or Clumped Bar Graph:



Stacked Bar Graph:



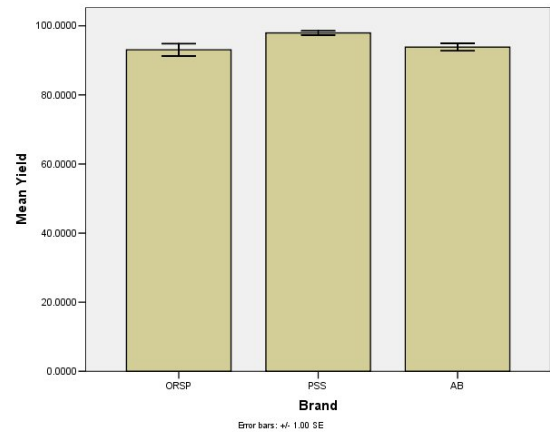
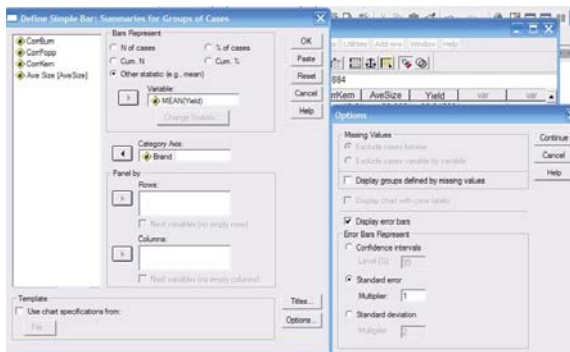
Depending on the organization of the data, you can create bar graphs that give even more information such as groups or composition. In all of the above examples, there is one key aspect missing: EXPRESSION OF ERROR. The big disadvantage of stacked bars is that you often lose the expression of error.

To make a bar graph:

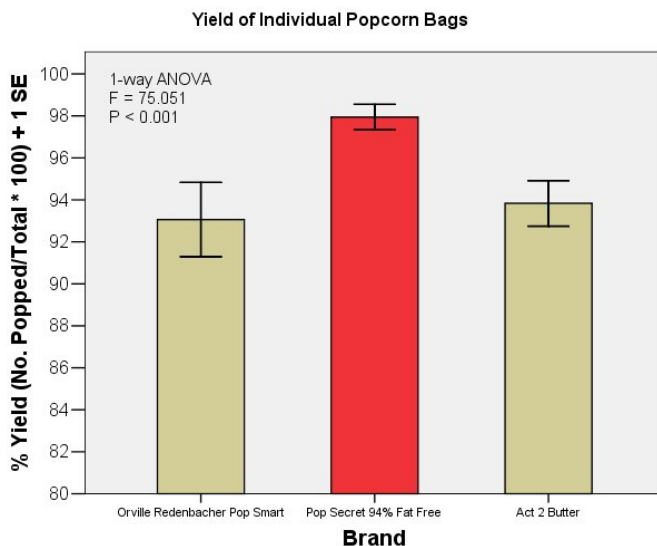
- 1) Open SPSSDATA1NAME.sav
- 2) Choose Graphs
- 3) Choose Bar
- 4) Choose Simple as you only have 3 groups and none of them are clustered
 - a) In contrast, think about your 2^3 ANOVA data in which you have 3 clusters
 - b) In that case, you would choose Clustered.
 - i) Then you should choose the most relevant group for the X-axis.
 - ii) Then the 2nd most important group for the cluster (bars side to side).
 - iii) Last, you chose the variable the exhibits the least difference for presentation in rows (two boxes on top) or columns (two side to side). See pg. 29 for an example.

- 5) Select the variable for bars represent and choose "Other Statistic" and insert YIELD.
- 6) Select a variable for the category axis (BRAND) and move it into the Category Axis box. This variable may be numeric, string, or long string.
- 7) Choose Options
- 8) Choose Display Error Bars
- 9) Choose Standard Error and change the number to 1
- 10) Hit Continue (note then you could format Titles if you wanted)
- 11) Then hit OK

→ Graph



The following setting will generate something similar to the graph you see on the right. Again, you will want to improve the appearance of that graph by using the Chart Editor. The idea is to make it look like this using these example edits:



- 1) Format axes to be more descriptive
- 2) Move information about error from small note at the bottom to Y axis
- 3) Change scale on Y axis to better show differences between brand, including major unit.
- 4) Change only significant difference bar to another color.
 - a) Note that you can also express statistical differences with small letters above bars.
 - b) Also, you can

distinguish between sets of bars also by using patterns. However, do not go overboard because then it makes the graph more difficult to read.

- 5) Change the width of the bars to make them more narrow.
- 6) Added statistical information in the form of a text box.

Scatter Plots:

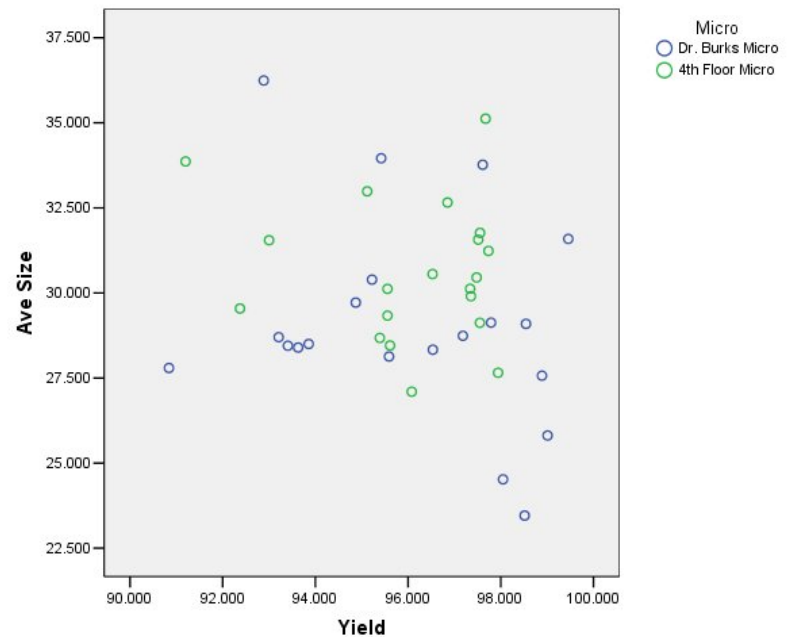
Scatter plots are similar to line graphs in that they use horizontal and vertical axes to plot data points. However, they have a very specific purpose. Scatter plots show how much one variable is affected by another. The relationship between two variables is called their correlation. Scatter plots usually consist of a large body of data. The closer the data points come when plotted to making a straight line, the higher the correlation between the two variables, or the stronger the relationship.

If the data points make a straight line going from the origin out to high x- and y-values, then the variables are said to have a positive correlation. If the line goes from a high-value on the y-axis down to a high-value on the x-axis, the variables have a negative correlation.

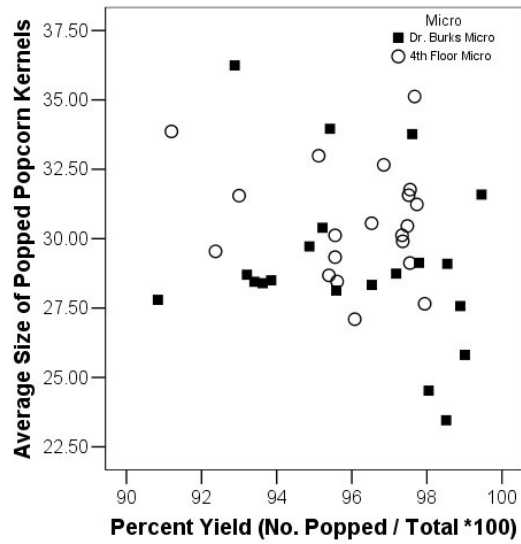
To make a bar graph:

- 1) Open SPSSDATA2NAME.sav
- 2) Choose Graphs
- 3) Choose Scatter/Dot
- 4) Choose Simple as you only have 2 variables
- 5) Add Average Size to Y-Axis Box
- 6) Add Yield to X-Axis Box
- 7) Set Markers by Micro
- 8) Hit OK

→ Graph



Again, there is a lot that you can do to format this graph so that it is more easy to read. Double click on the chart editor and format the graph so that it is easier to read and only black and white (Hint: if you click on the symbol next to the legend, then you can format at the data series). Make a list of the things that you had to change to format the scatter-plot.



Application/Homework: From your statistical analyses, determine what results would be most appropriate to graph. Then, produce one box plot (format B/W), one bar graph (Format color) and one scatter plot (format B/W), complete with captions. Export your graphs to PowerPoint and print full-size slides with captions.

Question Set 4:

1. Regression:
 - a. In your data, does a significant predictive relationship exist between the number of popped kernels and the number of leftover kernels?
 - b. Speculate about the mechanism behind the pattern that occurred.
 - c. What other dependent variables do you think might be predictive of each other?
2. Application: Test one prediction of what 2 other dependent variables you think might be correlated.
3. Types of graphs:
 - a. What is the advantage to a boxplot versus a bar graph?
 - b. Describe an example where you might want to use a horizontal bar graph versus a vertical bar graph.
 - c. What things did you have to format to improve your scatter plot?
4. Graphing: Prioritize and describe the 3 most important things that you think you learned about how to graph data in SPSS.

This concludes Part IV and the whole of Popcorn Statistics. Congratulations. Now you are ready to apply your statistical skills to your own data!

With this portion, you should turn in the following:

- 1) Regression output;
- 2) Graph 1 (boxplot) with captions
- 3) Graph 2 (bar graph) with captions
- 4) Graph 3 (scatter plot) with captions
- 5) Answer to Questions Set 4; double spaced.

Note that all SPSS output needs to be annotated. You can export your SPSS output to a Word Document (FILE then Export then file type .doc, browse (choose where you want the file and the name) and then OK). Format the word document for landscape and then print 2 pages per page.

