

01b: Review of Basic Statistical Concepts and Introduction to SPSS

1. Descriptive and Inferential Statistics

Descriptive

Descriptive statistics are used to describe data, relationships among variables, and differences among groups.

Inferential

Inferential statistics are used to test whether relationships among variables or differences among groups appear to be real or due to random chance. With inferential statistics the goal is to collect data with a sample, then make inferences from that sample to the population. Inferential statistics help us decide whether what was found in the sample appears to be real in the population or, instead, a fluke due to chance.

Statistics vs. Parameters

(a) Parameters apply to populations, examples include

- mean (μ) age of everyone in this class (we define this class as a population),
- standard deviation (σ) of age in this class, and
- variance (σ^2) of age in this class.
- Note use of Greek symbols for population parameters.

(b) Statistics are estimates of population parameters, for example

- mean (M or \bar{X}) age for a sample of students in this class (not everyone; at least 1 less than all students in this class is a sample),
- standard deviation (s or SD) of age in this class, and
- variance (s^2 or VAR) of age in this class.
- Note use of Roman symbols for sample statistics.
- In statistical inference we attempt to infer the value of a population parameter from its corresponding statistic.

(c) Some differences to note:

- There will be no hypothesis testing – no inferential statistics – if one is working with a census (population data); no test statistics (t-ratios, F-ratios), p-values, confidence intervals, or standard errors. Why? Because inferential statistics are only applied to sample data to infer to population parameters.
- Some small differences in formulas between statistics and parameters, e.g., with variance one divides by $n-1$ with sample, but by n with census (population).

2. Central Tendency (in SPSS)

Below are sample scores. What are the mean (M), median (Md), and mode (Mo) for these scores?

Scores: 6, 1, 3, 1, 5, 4, 2

SPSS Data Entry screenshot below.

Data entry spreadsheet in SPSS.

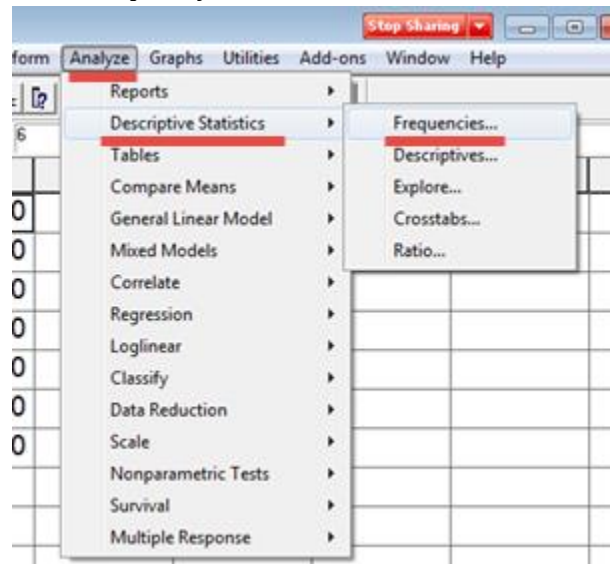
Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Ad

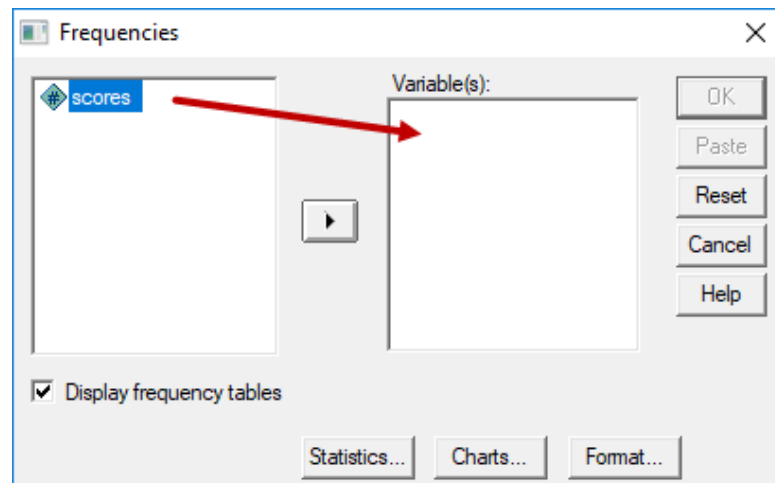
3: VAR00001 3

	VAR00001	var	var
1	6.00		
2	1.00		
3	3.00		
4	1.00		
5	5.00		
6	4.00		
7	2.00		
8			
9			
10			

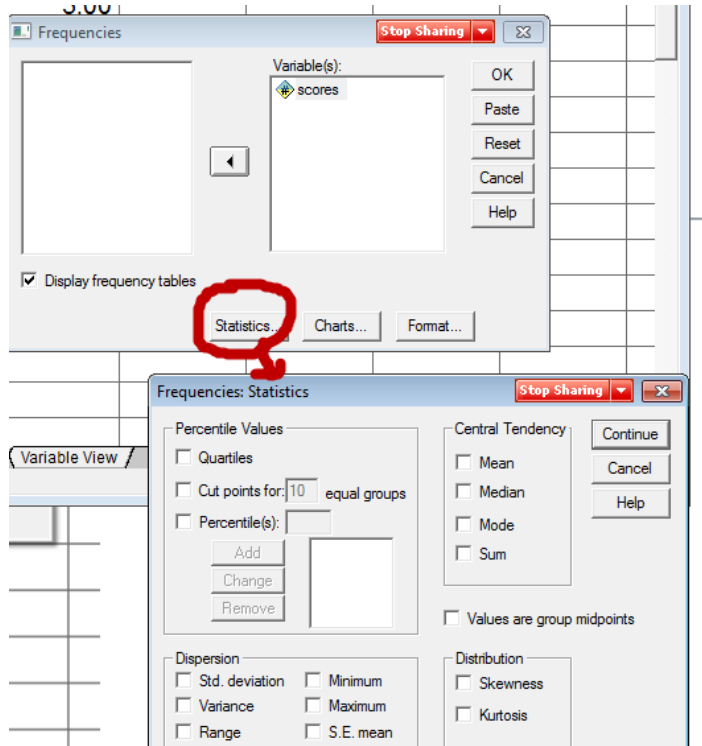
SPSS Frequency Command



Move the variable scores to the Variables box.



Click on the “Statistics” button then select the measures of central tendency needed.



Results from SPSS

Statistics

scores		
N	Valid	7
	Missing	0
Mean		3.1429
Median		3.0000
Mode		1.00
Std. Deviation		1.95180
Variance		3.810
Range		5.00

Mean = 3.1429

Median = 1 1 2 3 4 5 6

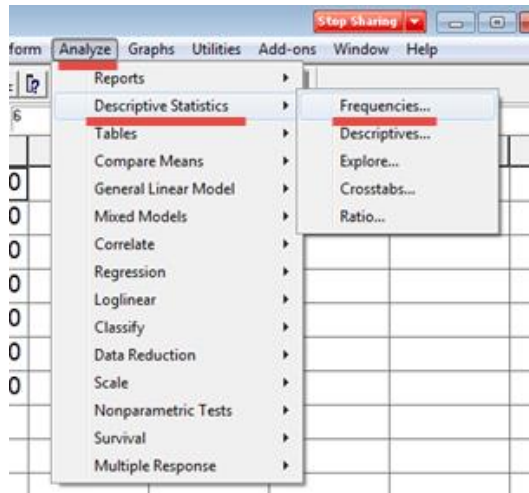
Mode = 1

3. Variability, Dispersion (in SPSS)

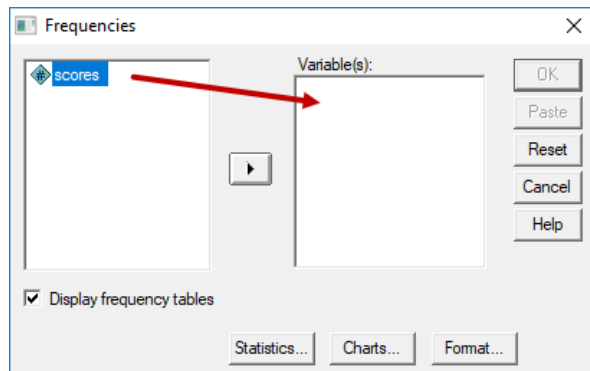
Use the same scores from above: 6, 1, 3, 1, 5, 4, 2. Find range (R), variance (VAR), and standard deviation (SD) for the above scores.

SPSS Commands

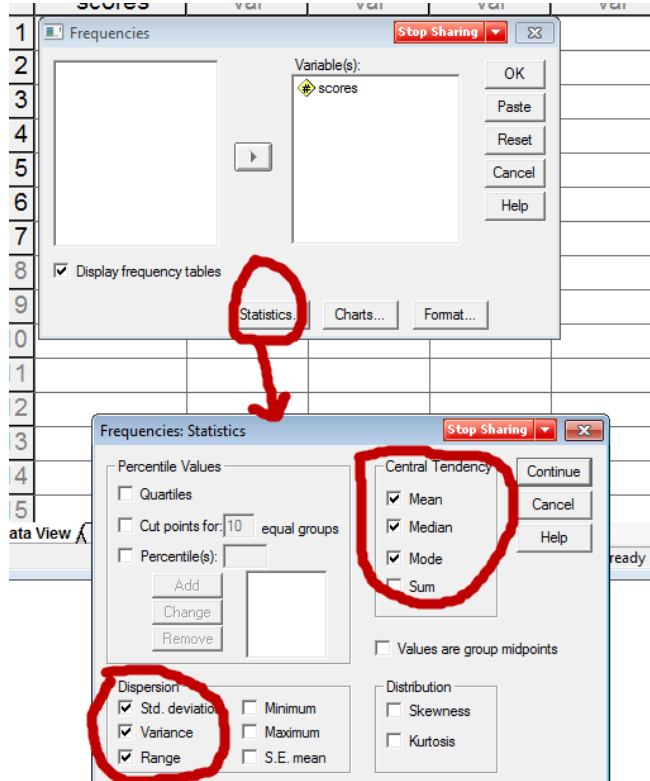
SPSS Frequency Command



Move the variable scores to the Variables box.



Click on the "Statistics" button then select the measures of variability (and central tendency) needed.



Results from SPSS

Statistics

scores		
N	Valid	7
	Missing	0
Mean		3.1429
Median		3.0000
Mode		1.00
Std. Deviation		1.95180
Variance		3.810
Range		5.00

Range = 5.00

Variance = 3.81

Standard Deviation = 1.9518

What does the SD represent?

Answer: Approximate mean (or average) of how far raw scores (X) deviate from the mean (M).

Example of deviation scores, SD is rough approximation to deviation score average (mean)

Scores	Mean	Deviations Scores (DS)	Squared DS
6	3.1429	2.8571	8.16302
1	3.1429	-2.1429	4.59202
3	3.1429	-0.1429	0.02042
1	3.1429	-2.1429	4.59202
5	3.1429	1.8571	3.44882
4	3.1429	0.8571	0.73462
2	3.1429	-1.1429	1.30622
	Sum	0	22.8571 = sums of squares (SS)
		VAR = SS/(n-1)	= 22.8571 / 6 = 3.8095
		SD = Square root of VAR	= SQRT(3.8095) = 1.9518

These values match the VAR and SD reported by SPSS above.

4. Frequencies, Percentile Ranks, and Quartiles

Frequencies

Example 1: Student Sex

We count the Sex of students in class. Below is a sample of students with their counts by sex.

M, M, F, F, F, F, M, F, M, F, F

Sex	Freq	Relative Freq.
Females	7	.6363 (64%)
Males	4	.3636 (36%)

N = 11 students,

$7/11 = .6363$,

$4 / 11 = .3636$


Frequency by Sex with SPSS

Below is screenshot of data entry. Note there are two Sex variables, one denoting sex with letters f and m, and a second denoting sex with numbers where female = 1 and male = 0. In SPSS it is usually better to use numbers to represent nominal variables like Sex because SPSS sometimes does not allow for processing of letters.

Data Entry Screen, SPSS

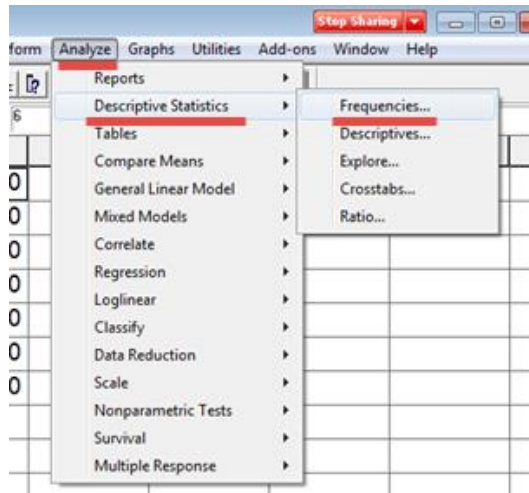
Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

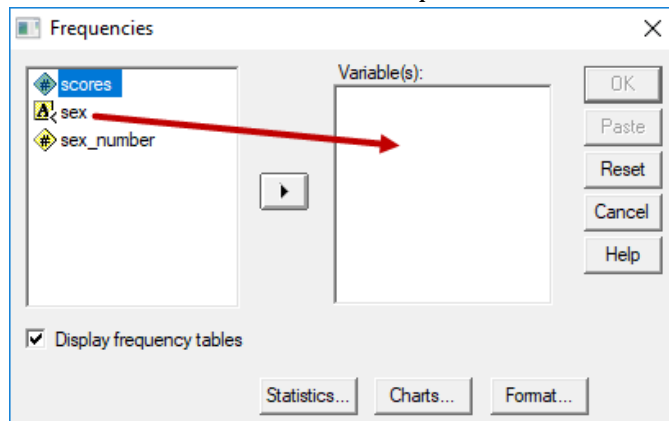


	scores	sex	sex_number	var
1	6.00	m	.00	
2	1.00	m	.00	
3	3.00	f	1.00	
4	1.00	f	1.00	
5	5.00	f	1.00	
6	4.00	f	1.00	
7	2.00	m	.00	
8	.	f	1.00	
9	.	m	.00	
10	.	f	1.00	
11	.	f	1.00	
12				

SPSS Frequency Command



Move the Sex variable to the Frequencies Variable box, then run the analysis.



Results from SPSS

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	f	7	63.6	63.6	63.6
	m	4	36.4	36.4	100.0
	Total	11	100.0	100.0	

Example 2: Frequencies for Numbers

Find frequencies and relative frequencies in SPSS for the scores in below.

6, 1, 3, 7, 5, 4, 2, 8

Results from SPSS

new_scores

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1.00	1	12.5	12.5	12.5
2.00	1	12.5	12.5	25.0
3.00	1	12.5	12.5	37.5
4.00	1	12.5	12.5	50.0
5.00	1	12.5	12.5	62.5
6.00	1	12.5	12.5	75.0
7.00	1	12.5	12.5	87.5
8.00	1	12.5	12.5	100.0
Total	8	100.0	100.0	

Assume response 5 was omitted from the 8 observations – note difference in percent vs. valid percent. Valid percent is the column that typically should be used since it is based upon obtained data and ignores missing values.

scores

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1.00	1	12.5	14.3	14.3
2.00	1	12.5	14.3	28.6
3.00	1	12.5	14.3	42.9
4.00	1	12.5	14.3	57.1
6.00	1	12.5	14.3	71.4
7.00	1	12.5	14.3	85.7
8.00	1	12.5	14.3	100.0
Total	7	87.5	100.0	
Missing System	1	12.5		
Total	8	100.0		

Note score 5 is missing and treated as a missing value.

Percentile Rank

What is a percentile rank? The most common definition and the one we will use:

PR = percentage (or proportion) of scores **at or** below a given score

Less common (and we won't use this one):

PR = proportion (or percentage) of scores below a given score.

The median is also the 50th percentile, i.e., PR = 50 = median.

The PR is appropriate for ranked, numeric data; it is not appropriate for categorical, nominal variables like sex, race, or types of flowers.

Percentile Rank with SPSS for Raw Data

In the Frequency display produced by SPSS, the column "Cumulative Percent" is the percentile rank for raw score data.

new_scores

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1.00	1	12.5	12.5	12.5
2.00	1	12.5	12.5	25.0
3.00	1	12.5	12.5	37.5
4.00	1	12.5	12.5	50.0
5.00	1	12.5	12.5	62.5
6.00	1	12.5	12.5	75.0
7.00	1	12.5	12.5	87.5
8.00	1	12.5	12.5	100.0
Total	8	100.0	100.0	

Percentile Rank column.

Example

For score of 5 using the frequency table above, 62.5 is the PR which means 62.5% of sample scored 5 or less.

SPSS Percentiles from Frequencies command

Statistics

new_scores

N	Valid	8
	Missing	0
Mean		4.5000
Median		4.5000
Mode		1.00 ^a
Std. Deviation		2.44949
Variance		6.000
Range		7.00
Percentiles	25	2.2500
	50	4.5000
	75	6.7500

a. Multiple modes exist. The smallest value is shown

The table above shows what SPSS produces when percentiles are requested in the Frequencies command. Note that scores of 2.25, 4.50, and 6.75 do not appear in the data. These represent calculated percentile scores for the ranks of 25, 50, and 75. They differ from the values provided by the Cumulative Percent column for the percentiles, and this discrepancy is common for small data files, and the even number of values creates the problem of the median of 4.00 vs. 4.50.

Quartiles

Quartiles are formed by the 25th, 50th, and 75th percentiles. Four sections with equal numbers of sampled units in each section.

To obtain quartiles, divide the score distribution into 4 sections with 25% of scores in each section based upon percentile ranks using these formulas:

1st quartile: median between lowest score and overall median of distribution

2nd quartile: median of distribution

3rd quartile: median between highest score and overall median of distribution

Also

- 1st quartile – 25th percentile
- 2nd quartile – 50th percentile (median)
- 3rd quartile – 75% percentile

Graphically:

Scores	1	2		3	4		5	6		7	8
			↑			↑			↑		
Quartiles =			1 st = 2.5			2 nd = 4.5			3 rd = 6.5		
Percentiles =			25			50			75		

SPSS reports different values for quartiles: 2.25, 4.50, and 6.75

Statistics

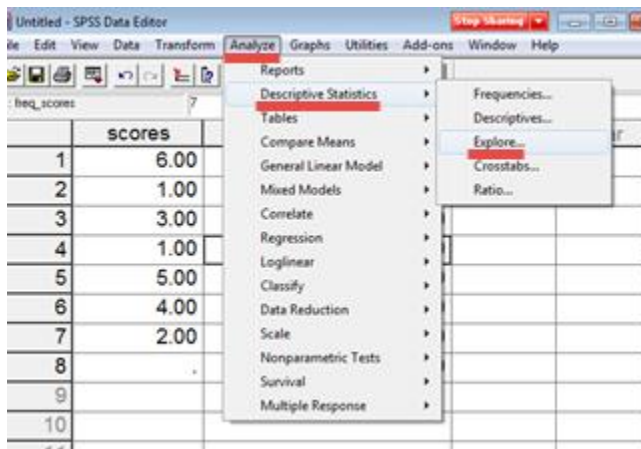
scores		
N	Valid	8
	Missing	0
Mean		4.5000
Median		4.5000
Mode		1.00 ^a
Std. Deviation		2.44949
Variance		6.000
Range		7.00
Sum		36.00
Percentiles	25	2.2500
	50	4.5000
	75	6.7500

a. Multiple modes exist. The smallest value is shown

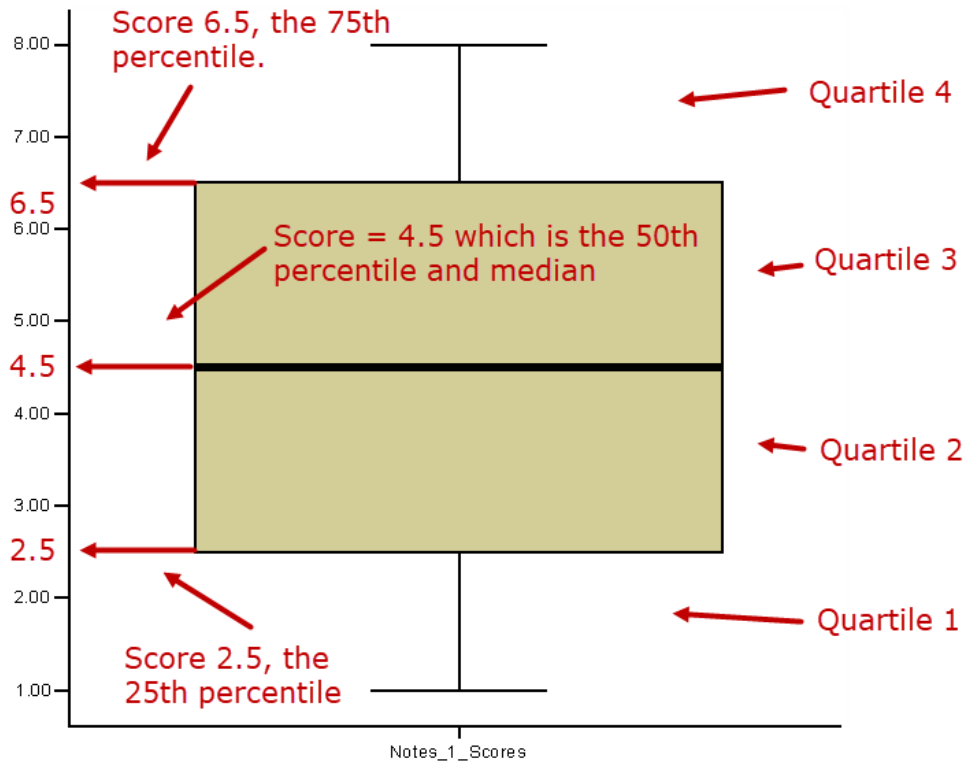
There are slight differences in quartile values reported by me above and by SPSS, so if you do it by hand use the formula above and if you rely on software report whatever values the software provides because all formulas for quartiles (and percentiles) provide close estimates.

5. Boxplot or Box and Whisker Plot

A boxplot is a graphical means of displaying central tendency and spread of scores. One may use the SPSS Explore command to obtain boxplots.



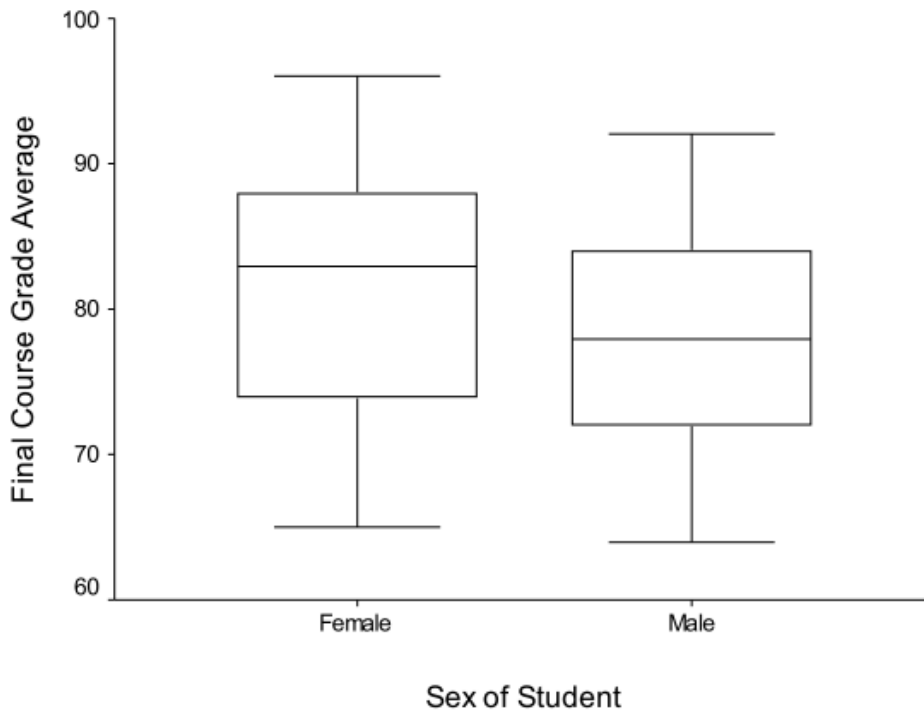
SPSS Boxplot



Note that the boxplot above uses values of 2.5 and 6.5 for the 25th and 75th percentiles, which are the same values I calculated, but are inconsistent with the SPSS Frequencies command result reported above.

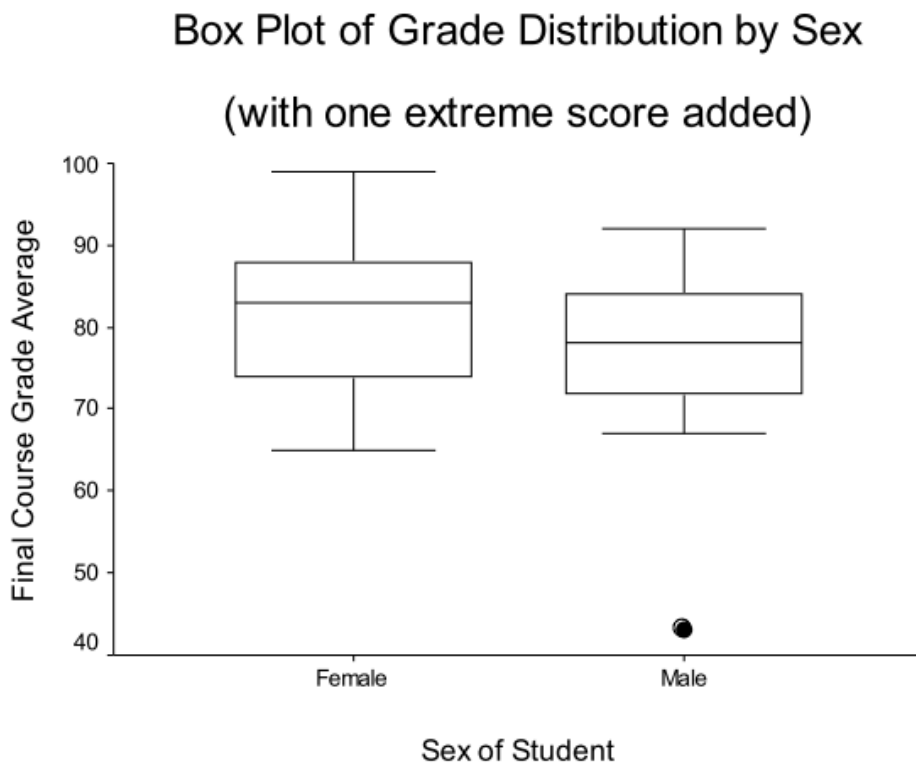
Box and Whisker plots need not show equal quartile sizes. See the example below for grade distribution by sex.

Box Plot of Grade Distribution by Sex



Boxplots are designed to display several summary indicators of data. For example, for females, the bottom of the box shows the score at the 25th percentile (symbolized as P_{25} which is roughly 74 in this sample); the top of the box is the 75th percentile (P_{75} , a score in this sample of about 88); and the thick line in the middle of the box represents the median (50th percentile, P_{50}). Note that the box is designed to describe the middle 50% of scores in the distribution.

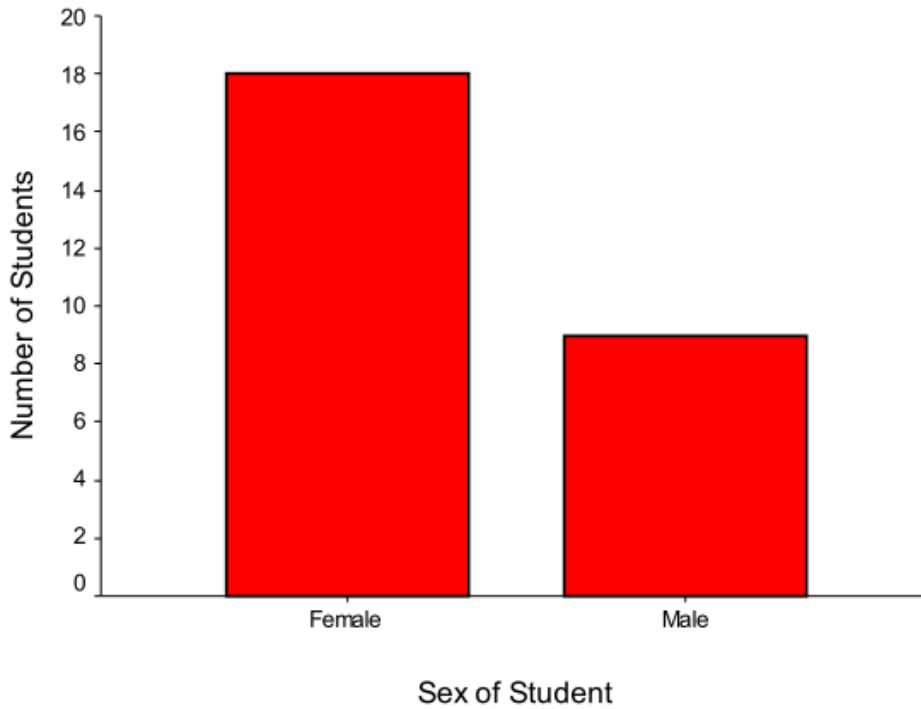
The whiskers extending from the box may represent several different things depending upon how they are implemented for given software. For the example listed below, the whiskers appear to show the upper and lower range for the sample of scores. The bottom whisker shows the lower range for the distribution of scores (a lower score of about 64); and the top whisker shows the upper range for the distribution of scores (a top score of about 96). In some software applications, whiskers extend to P_{10} and P_{90} (percentiles 10 and 90), and any scores beyond this range are represented as dots. The second box plot below illustrates a score, denoted by the black dot, that extends below the range of P_{10} .



6. Bar Charts

Bar charts are traditionally used to display frequency counts for qualitative variables. Below is an example showing the sex distribution of students in a class. As the frequency display below the bar chart shows, there were 18 females and 9 males enrolled in the class.

Distribution of Students by Sex

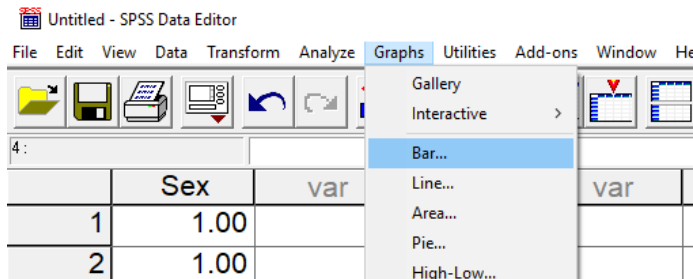


```
. tabulate Sex
```

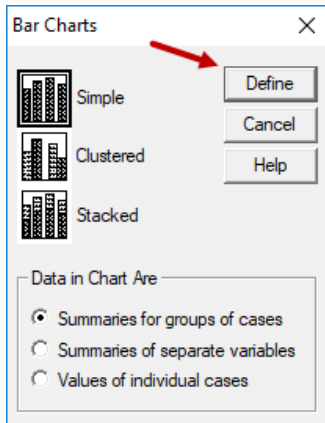
Sex of Student	Freq.	Percent	Cum.
f	18	66.67	66.67
m	9	33.33	100.00
Total	27	100.00	

SPSS Bar Graphs

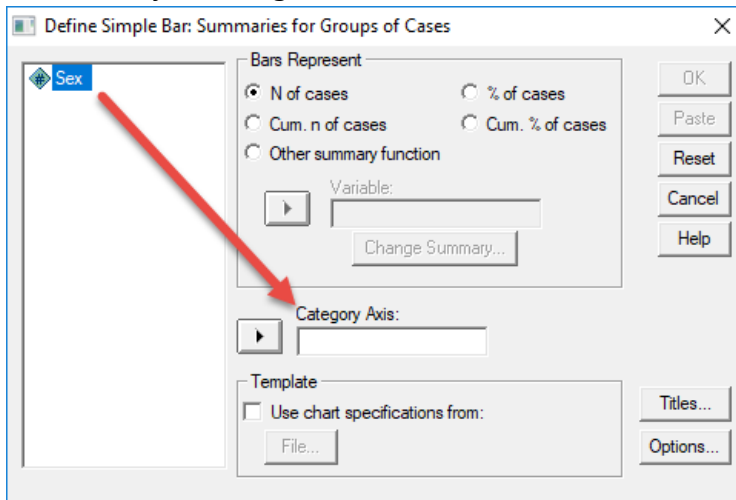
SPSS offers several options for obtaining bar graphs. Below is one using the Graphs->Bar command.



Next select Simple then select Define.

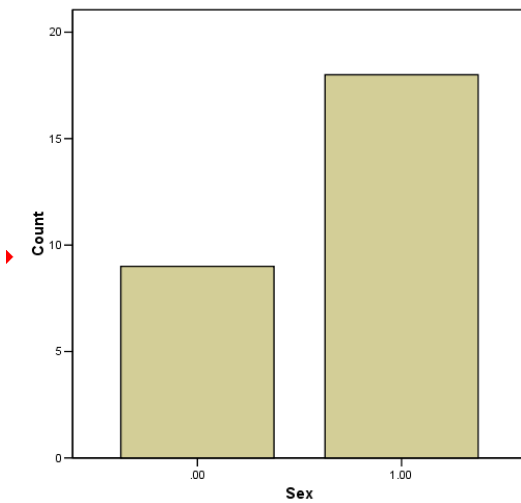


Next identify the categorical variable to be used and move it to the variable box.



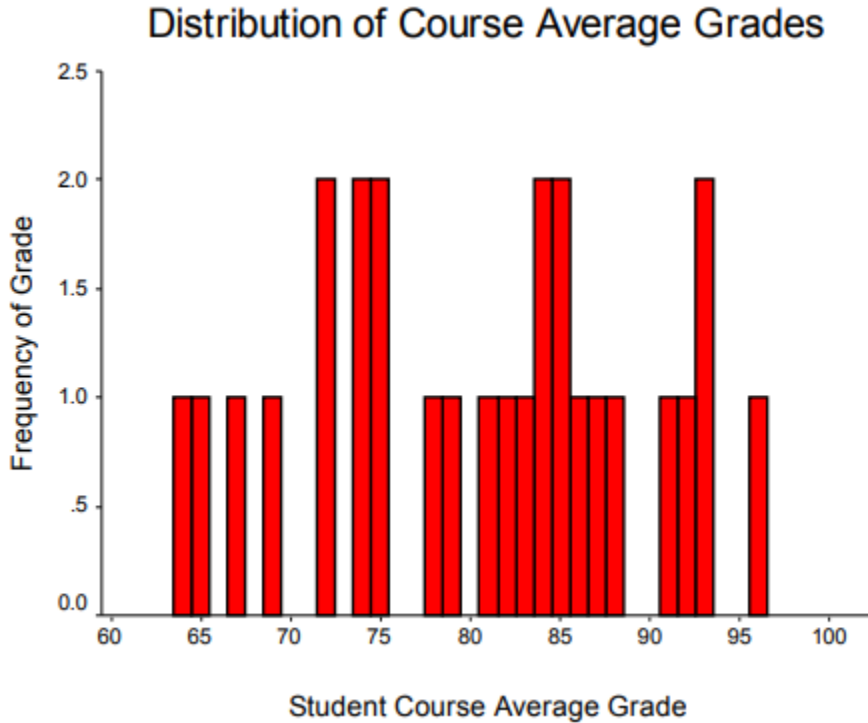
And here is the SPSS output.

Graph

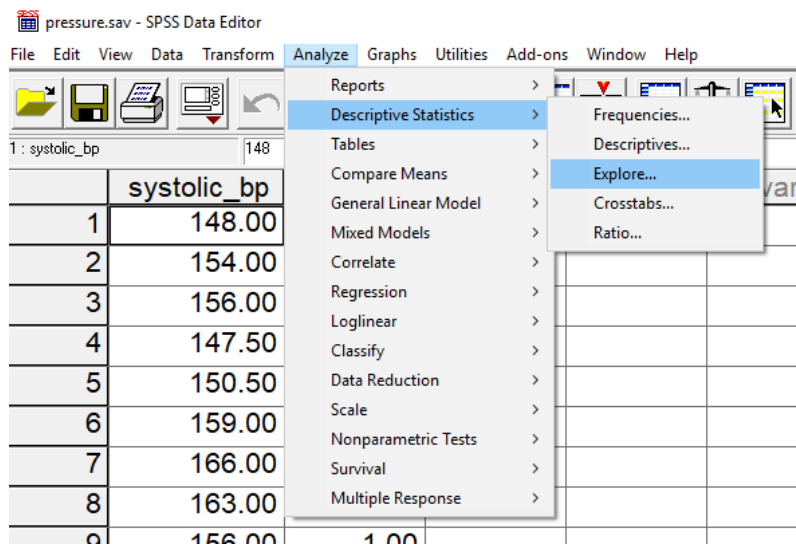


7. Histograms and Frequency Polygons

Similar to bar charts and stem-and-leaf displays, histograms may be used to show frequency information for quantitative variables. The primary difference between histograms and bar charts is that histograms are designed for quantitative data so the bars are allowed to touch when consecutive scores are presented. When gaps are present between bars in a histogram, that signals a frequency of zero for that particular score. Below is an example of a histogram for student grades. Smooth histograms are often used to present distributional shapes such as normal, F, t, chi-square, etc. These are called frequency polygons.

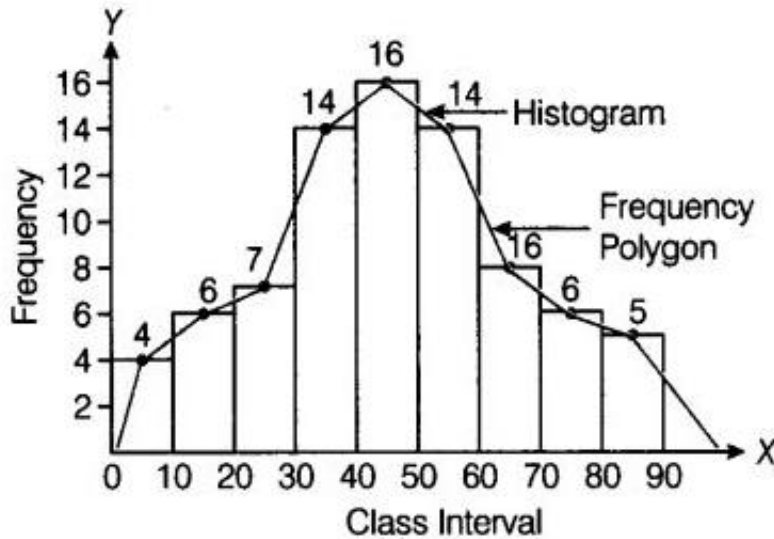


SPSS command option to obtain Histograms. Select Analyze->Descriptive Statistics-> Explore then select plots and histogram.



My version of SPSS does not produce frequency polygons, so below is an example taken from the linked source below.

<http://ask.learnbse.in/t/construct-a-frequency-polygon-with-histogram-for-the-following-data/16606>

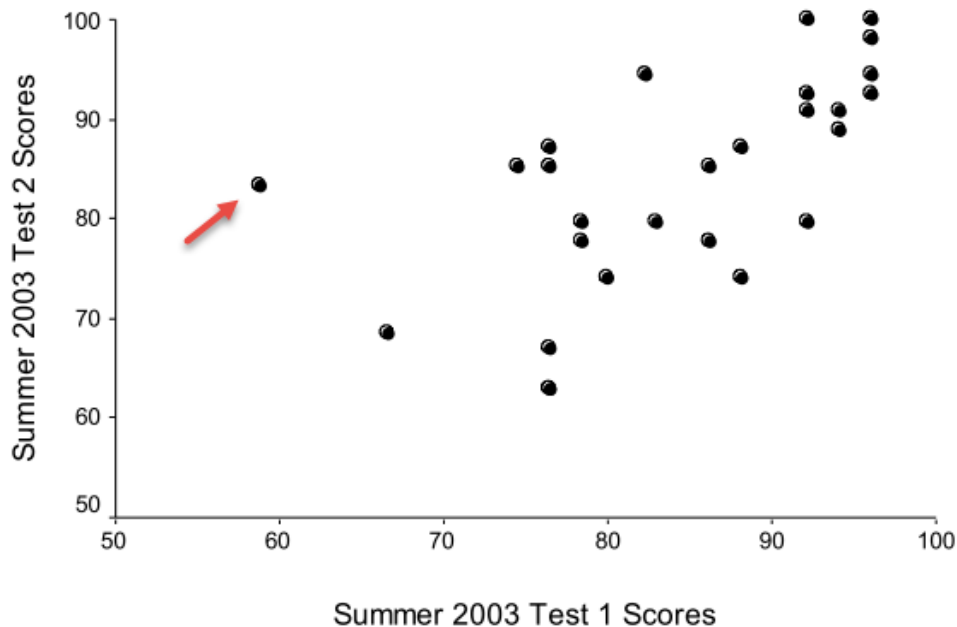


8. Scatterplots

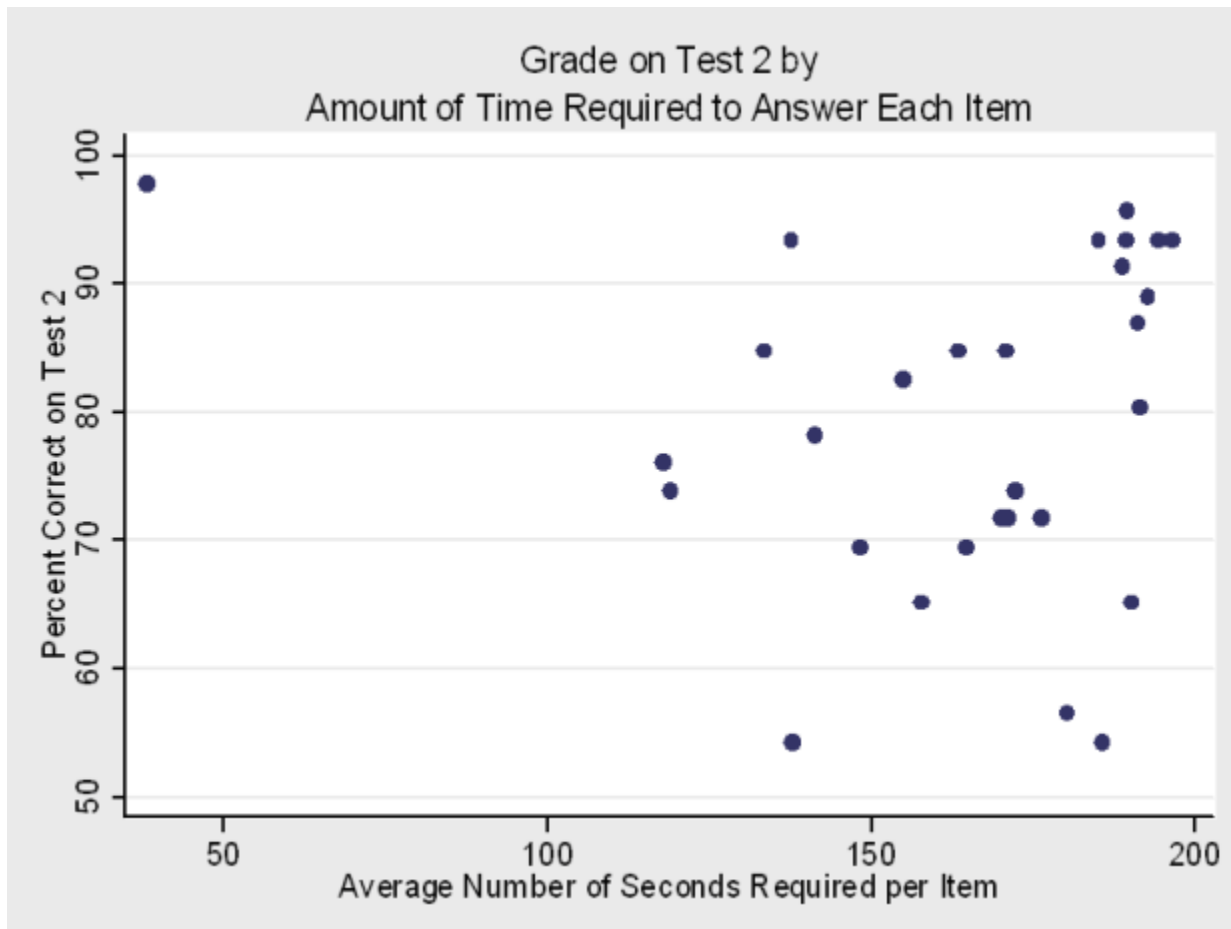
These graphs are useful for displaying the nature of relation between two quantitative variables. As the first scatterplot shows, there is a positive, linear trend between scores from tests 1 and 2 in educational research during the summer of 2003. Students who did well on test 1 tended also to perform well on test 2; similarly, those who performed poorly on test 1 also tended to perform poorly on test 2. There are, however, several exceptions to this trend. Note the student who scored just under 60 for test 1 but scored over 80 for test 2. This student is the isolated dot to the left of other dots in the scatter denoted by the red arrow.

Scatterplot of Test Scores from Students

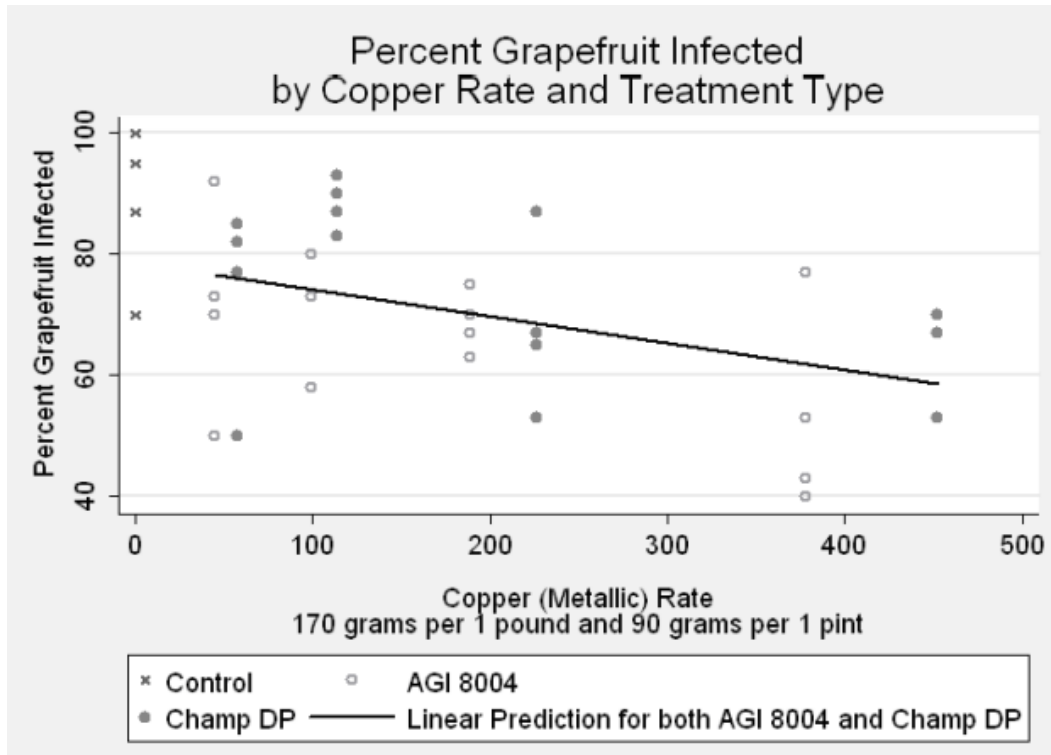
in Educational Research Summer 2003



The next scatterplot, displayed below, shows information pertaining to student performance on a test in educational research. The two variables considered are the average number of seconds spent per item completing the test and test score. The scatter of data to the right of the graph shows a slight positive relation between time spent on items and test score. Those students who spent more time per item tended to perform better on the test, although this pattern is not strong. Most students took between 120 and 195 seconds to answer each item (that's 2 to 3.5 minutes per item). There is one very clear exception to these data and that exception is symbolized by a student who spent an average of 38 seconds per item and scored 98% correct on this test. This student's performance represents what is known as an outlier, an observation that is clearly discrepant from other observations (data) in the distribution of scores.



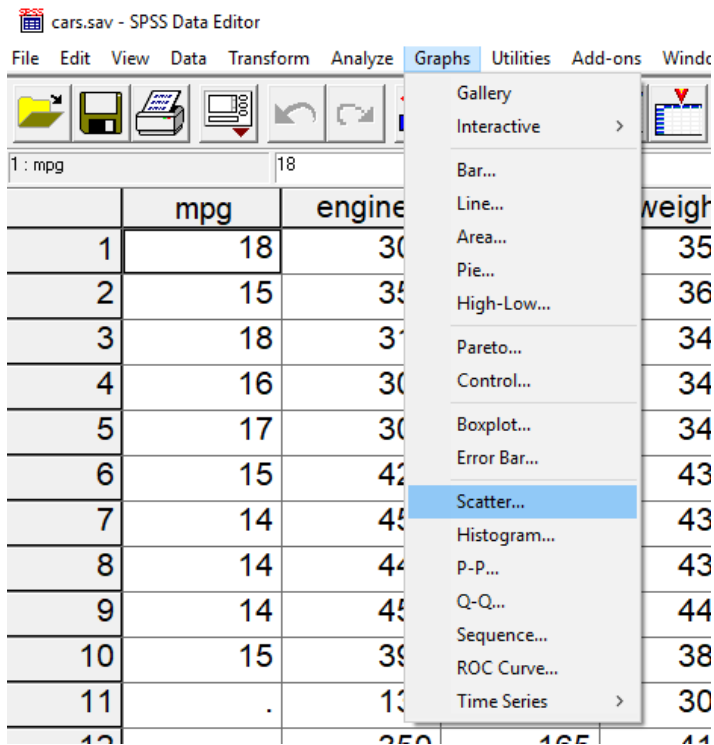
The next scatterplot displays data from an agricultural experiment in which grapefruit were treated with two types of fungicides and with varying amounts of the active ingredient (copper). The outcome of interest is the severity of the infection on grapefruit. The line in the graph represents a prediction line and can be used to estimate the change in severity of infection according to differing amounts of copper used.



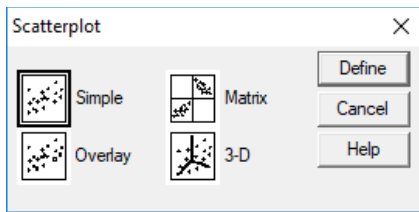
SPSS Scatterplots

To obtain scatterplots with SPSS, use the Graph->Scatterplot option, shown below. The linked data can be downloaded for those who wish to replicate this scatterplot.

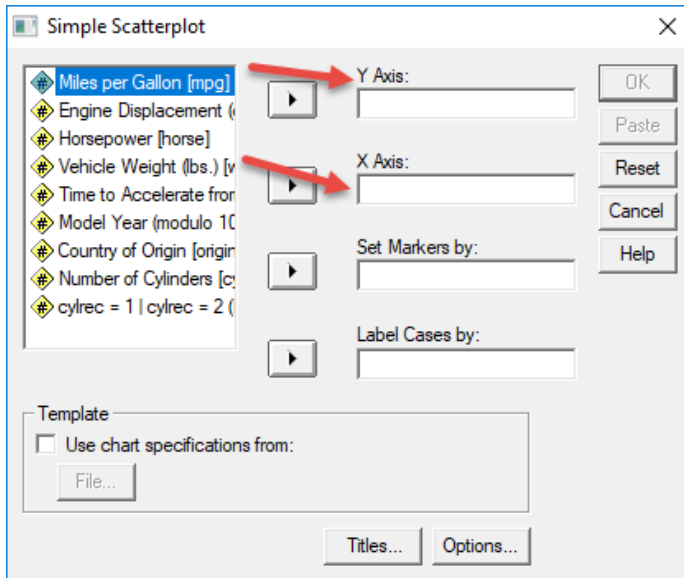
<http://www.bwgriffin.com/anova/cars.sav>



For the next window, select Simple and Define (if you want multiple scatterplots shown together for multiple variables, select Matrix or Overlay options).



For this example, we will plot the scatter between miles per gallon (MPG) and vehicle weight. The DV is MPG and IV is weight, so place MPG on the Y axis (vertical) and weight on the X axis (horizontal) of the scatterplot. Finish by clicking on the OK button.



Scatterplot of MPG and vehicle weight from SPSS. Results show a curvilinear relation and there is one strong outlier (low weight car with less than 10 MPG).

