

Section 3

DESCRIPTIVE STATISTICS

Table of Contents

SECTION 3	1
DESCRIPTIVE STATISTICS	1
TABLE OF CONTENTS	1
FIGURES	2
DESCRIPTIVE STATISTICS WITH LAZSTATS	3
CENTRAL TENDENCY AND VARIABILITY	3
FREQUENCIES	4
CROSS-TABULATION	7
BREAKDOWN	7
NORMALITY TESTS	10
X VERSUS Y PLOT	11
GROUP FREQUENCY HISTOGRAMS	13
REPEATED MEASURES BUBBLE PLOT	14
COMPARISONS WITH THEORETICAL DISTRIBUTIONS	17
THREE DIMENSIONAL ROTATION	18
BOX PLOTS	21
PLOT X VERSUS MULTIPLE Y VALUES	24
STEM AND LEAF PLOT	26
MULTIPLE GROUP X VERSUS Y PLOT	29
RESISTANT LINE FOR BIVARIATE DATA	31

Figures

Figure 1. The Dialog for Central Tendency and Variability	3
Figure 2. The Frequencies Dialog	5
Figure 3. Specifying the Interval Size and Number of Intervals for the Frequency Analysis	5
Figure 4. A Plot of Frequencies in the Cansas.LAZ File	6
Figure 5. Specification of a Cross-Tabulation	7
Figure 6. The Breakdown Form	8
Figure 7 Normality Test Dialog.....	10
Figure 8 X Versus Y Dialog	11
Figure 9 A Plot of Two Variables	12
Figure 10 Specification Dialog for a Frequency Analysis	13
Figure 11 A Sample Frequency Plot.....	13
Figure 12 Repeated Measures Bubble Plot Dialog.....	14
Figure 13 Bubble Plot of School Achievement.....	15
Figure 14 Plot of Teacher-Student Ratio to Achievement.....	16
Figure 15 Comparison of Cumulative Distributions.....	17
Figure 16 Cumulative Normal vs. Cumulative Observed Values.....	17
Figure 17 Scatter Plot of Values for Three Variables.....	19
Figure 18 Rotated Variables to Examine Relationship Between Two Variables.....	20
Figure 19 Box Plot Dialog.....	21
Figure 20 Box Plot of the Slice Variable.....	23
Figure 21 Plot X Versus Multiple Y Dialog.....	24
Figure 22 Teacher Salaries Versus SAT Achievement.....	25
Figure 23 Stem and Leaf Plot Dialog	26
Figure 24 The Multiple Group X vs. Y Plot Dialog.....	29
Figure 25 An X vs. Y Plot for Multiple Groups.....	30
Figure 26 Resistant Line Dialogue Form.....	31
Figure 27 Plot of Resistant Line Slopes.....	32

Descriptive Statistics With LazStats

This section demonstrates the use of LazStats to obtain descriptive statistics for data that you have entered in a file on the main form's grid. In many cases, a graphical picture of one's data is highly useful in understanding the distribution of the values for one or more variables. In some procedures, the data of one or more variables must be defined as an integer. In other procedures, the data should be defined as a floating point variable. Be sure to define your variables as needed for each procedure.

Central Tendency and Variability

Click on the Analyses menu and place your mouse on the Descriptive option. The sub-option for Central Tendency and Variability is then chosen by clicking that option. To demonstrate, we will use the file labeled *cansas.LAZ* and obtain the descriptive statistics for the variable "Weight".

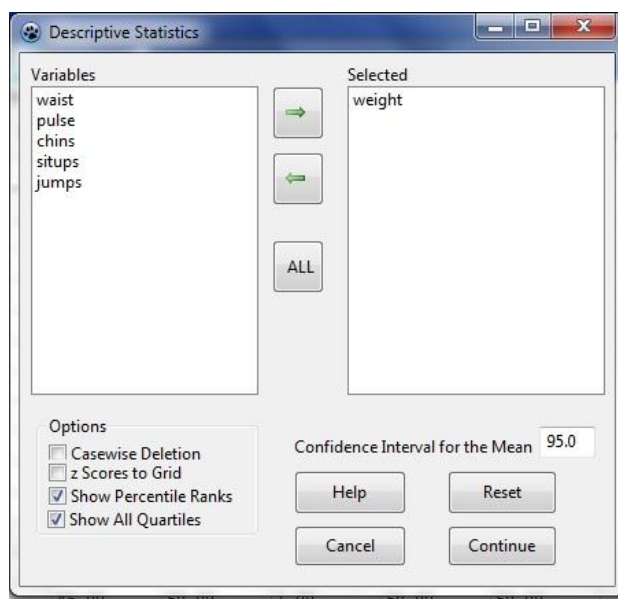


Figure 1. The Dialog for Central Tendency and Variability

When you click the OK button on the above form you will see the output displayed on the "Output Form". Notice that there are several options that may have been selected. The CaseWise Deletion option lets you obtain the results for only those cases in which there are no missing values. The z Scores to Grid option lets you create new variables that are the standardized z scores (mean of 0 and standard deviation of 1.0) for the variables you selected to analyze. Shown below is the result of our analysis:

DISTRIBUTION PARAMETER ESTIMATES

```
weight (N = 20)  Sum =          3572.000
Mean =          178.600  Variance =          609.621  Std.Dev. =          24.691
Std.Error of Mean =          5.521
95.00 percent Confidence Interval for the mean = 167.083 to 190.117
Range =          109.000  Minimum =          138.000  Maximum =          247.000
Skewness =          0.970  Std. Error of Skew =          0.512
Kurtosis =          1.802  Std. Error Kurtosis =          0.992
```

PERCENTILE RANKS

Score Value	Frequency	Cum.Freq.	Percentile Rank
138.000	1	1	2.50
154.000	2	3	10.00
156.000	1	4	17.50
157.000	1	5	22.50
162.000	1	6	27.50
166.000	1	7	32.50
167.000	1	8	37.50
169.000	1	9	42.50
176.000	2	11	50.00
182.000	1	12	57.50
189.000	2	14	65.00
191.000	1	15	72.50
193.000	2	17	80.00
202.000	1	18	87.50
211.000	1	19	92.50
247.000	1	20	97.50

First Quartile = 158.250
Median = 176.000
Third Quartile = 192.500
Interquartile range = 34.250

Alternative Methods for Obtaining Quartiles

	Method 1	2	3	4	5	6	7	8
Pcntile								
Q1	157.000	158.250	157.000	159.500	160.750	157.000	160.750	157.000
Q2	176.000	176.000	176.000	176.000	176.000	176.000	176.000	176.000
Q3	191.000	192.500	191.000	192.000	191.500	191.000	191.500	193.000

NOTES:

Method 1 is the weighted average at X[np] where n is no. of cases, p is percentile / 100

Method 2 is the weighted average at X[(n+1)p] This is used in this program.

Method 3 is the empirical distribution function.

Method 4 is called the empirical distribution function - averaging.

Method 5 is called the empirical distribution function = Interpolation.

Method 6 is the closest observation method.

Method 7 is from the TrueBasic Statistics Graphics Toolkit.

Method 8 was used in an older Microsoft Excel version.

See the internet site <http://www.xycoon.com/> for the above.

=====

Frequencies

Another way to examine data is to obtain the frequency of cases that fall within categories determined by a range of score values. To do this, click on the Frequencies option under the Descriptive menu. You will see the form shown below:

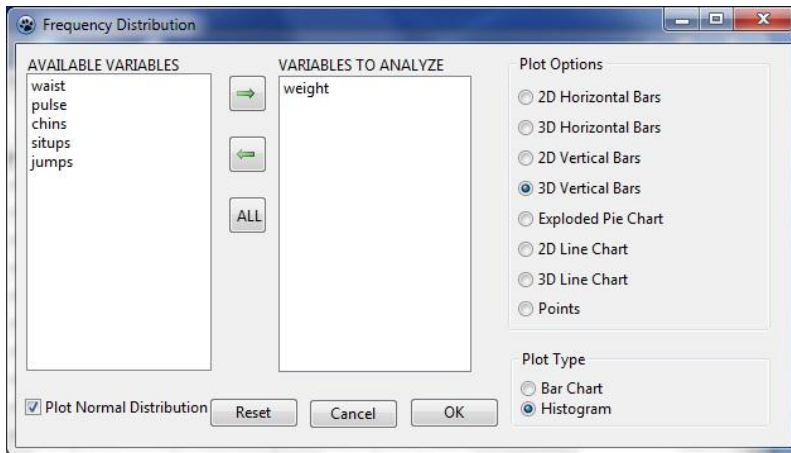


Figure 2. The Frequencies Dialog

Notice that we have selected the variable “Weight” from the cansas.LAZ file. We have also elected to obtain a three dimensional, vertical bar chart of the obtained frequencies and to plot the normal distribution for corresponding frequencies behind the bar chart. Also elected was to create a new variable in the grid that contains an integer value of the frequency group. This could be useful for other graphical plots like the box plot procedure. When we click the OK button above, we first are presented with a dialog box that asks us to define the interval size and the number of intervals. One must enter an interval size that produces a number of intervals equal to or less than the number of cases. You simply click on that box and enter the new value. When you press the return key after entering a new value, you will see a change in the number of intervals. You can repeat that process until the number of intervals is acceptable. If you attempt to create more intervals than the number of cases, you will receive a warning and be returned to this dialog:

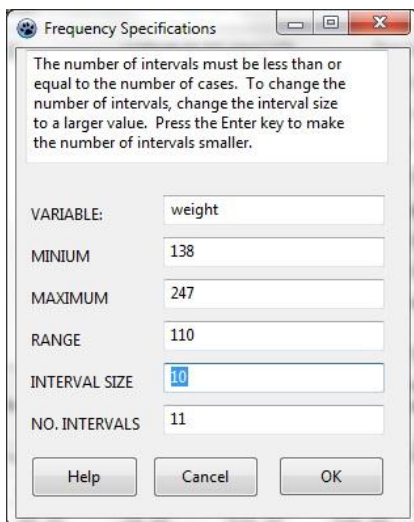


Figure 3. Specifying the Interval Size and Number of Intervals for the Frequency Analysis

Notice we have changed the interval size to 10 which resulted in the number of intervals that is less than the number of cases. Clicking the OK button results in the following:

FREQUENCY ANALYSIS BY BILL MILLER

Frequency Analysis for weight

FROM	TO	FREQ.	PCNT	CUM.FREQ.	CUM.PCNT.	%ILE RANK
138.00	148.00	1	0.05	1.00	0.05	0.03
148.00	158.00	4	0.20	5.00	0.25	0.15
158.00	168.00	3	0.15	8.00	0.40	0.33
168.00	178.00	3	0.15	11.00	0.55	0.47
178.00	188.00	1	0.05	12.00	0.60	0.57
188.00	198.00	5	0.25	17.00	0.85	0.72
198.00	208.00	1	0.05	18.00	0.90	0.88
208.00	218.00	1	0.05	19.00	0.95	0.93
218.00	228.00	0	0.00	19.00	0.95	0.95
228.00	238.00	0	0.00	19.00	0.95	0.95
238.00	248.00	1	0.05	20.00	1.00	0.97

Interval ND Freq.

1	1.16
2	1.90
3	2.63
4	3.12
5	3.14
6	2.70
7	1.97
8	1.23
9	0.65
10	0.30
11	0.11
12	0.04

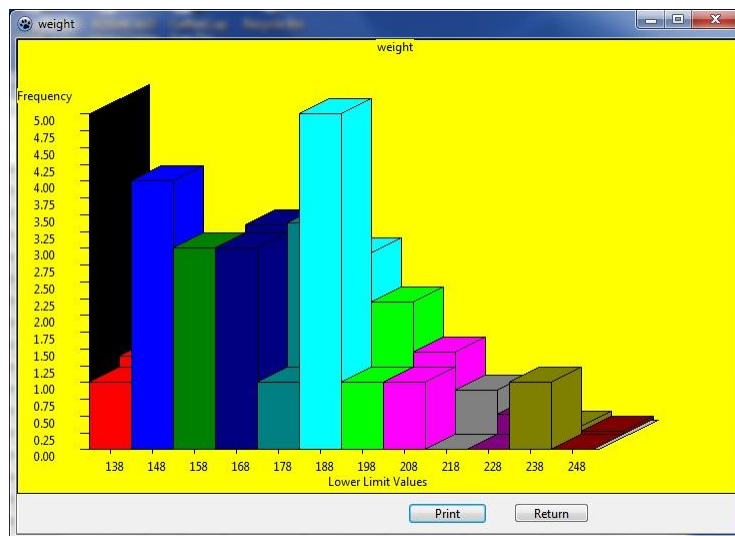


Figure 4. A Plot of Frequencies in the Cansas.LAZ File

Notice that the bars in the front of the plot represent the frequency of scores in the intervals of our data while the bars behind represent frequencies expected in the normal distribution.

Cross-Tabulation

When you have entered data that represents cases classified by two or more categorical variables, it is useful to count the number of cases classified in those categories. The Cross Tabulation option of the Descriptives option gives you those results. We will use a file labeled “twoway.LAZ” to demonstrate. We have loaded the file into the grid and elected the cross tabulation option. Below are the results:

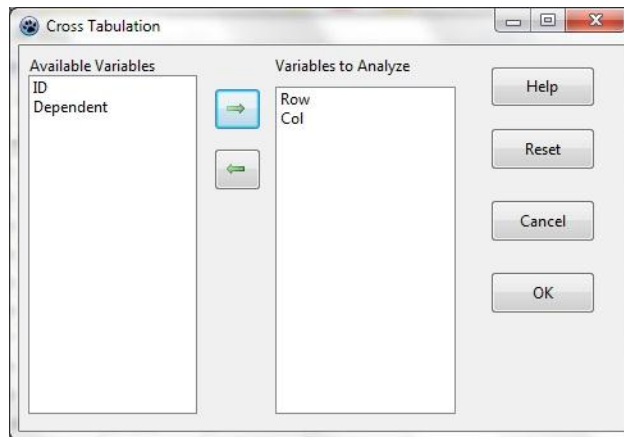


Figure 5. Specification of a Cross-Tabulation

CROSSTAB RESULTS

Analyzed data is from file : C:\lazarus\Projects\LazStats\LazStatsData\twoway.LAZ

Row min.= 1, max.= 2, no. levels = 2

Col min.= 1, max.= 2, no. levels = 2

FREQUENCIES BY LEVEL:

For cell levels: Row : 1 Col: 1 Frequency = 3

For cell levels: Row : 1 Col: 2 Frequency = 3

Sum across levels = 6

For cell levels: Row : 2 Col: 1 Frequency = 3

For cell levels: Row : 2 Col: 2 Frequency = 3

Sum across levels = 6

Cell Frequencies by Levels with 12 cases.

Variables

	Col: 1	Col: 2
Block 1	3.000	3.000
Block 2	3.000	3.000

Grand sum across all categories = 32

Breakdown

A procedure related to the Cross-Tabulation procedure described above lets you analyze a continuous (floating point) variable broken down into categories of one or more categorical variables.

Using the same file as above (twoway.LAZ) we will demonstrate this procedure. Below is the form and the results.

Figure 6. The Breakdown Form

BREAKDOWN ANALYSIS PROGRAM

VARIABLE SEQUENCE FOR THE BREAKDOWN:

Row (Variable 1) Lowest level = 1 Highest level = 2
 Col (Variable 2) Lowest level = 1 Highest level = 2

Variable levels:

Row level = 1
 Col level = 1

Freq.	Mean	Std. Dev.
3	3.000	1.000

Variable levels:

Row level = 1
 Col level = 2

Freq.	Mean	Std. Dev.
3	6.000	1.000

Number of observations accross levels = 6
 Mean accross levels = 4.500
 Std. Dev. accross levels = 1.871

Variable levels:

Row level = 2
 Col level = 1

Freq.	Mean	Std. Dev.
3	10.000	2.646

Variable levels:

Row level = 2
 Col level = 2

Freq.	Mean	Std. Dev.
3	12.000	2.646

Number of observations accross levels = 6
 Mean accross levels = 11.000
 Std. Dev. accross levels = 2.608

Grand number of observations accross all categories = 12
 Overall Mean = 7.750
 Overall standard deviation = 4.025

ANALYSES OF VARIANCE SUMMARY TABLES

Variable levels:

Row level = 1
 Col level = 1

Variable levels:

Row level = 1
 Col level = 2

SOURCE	D.F.	SS	MS	F	Prob.>F
GROUPS	1	13.50	13.50	13.500	0.0213
WITHIN	4	4.00	1.00		
TOTAL	5	17.50			

Variable levels:

Row level = 2
 Col level = 1

Variable levels:

Row level = 2
 Col level = 2

SOURCE	D.F.	SS	MS	F	Prob.>F
GROUPS	1	6.00	6.00	0.857	0.4069
WITHIN	4	28.00	7.00		
TOTAL	5	34.00			

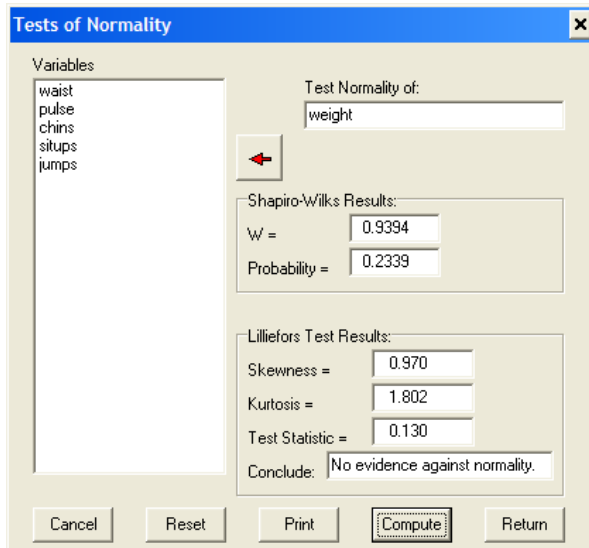
ANOVA FOR ALL CELLS

SOURCE	D.F.	SS	MS	F	Prob.>F
GROUPS	3	146.25	48.75	12.188	0.0024
WITHIN	8	32.00	4.00		
TOTAL	11	178.25			

FINISHED

Normality Tests

One can test the assumption that the distribution of values in a variable are a random sample from a normally distributed population. The dialog form is shown below:



The image shows a software dialog box titled "Tests of Normality". On the left, a list of variables includes "waist", "pulse", "chins", "situps", and "jumps". A red arrow points from this list to a text field labeled "Test Normality of:" which contains the word "weight". Below this, the "Shapiro-Wilks Results:" section shows "W =" with a value of 0.9394 and "Probability =" with a value of 0.2339. The "Lilliefors Test Results:" section shows "Skewness =" with a value of 0.970, "Kurtosis =" with a value of 1.802, and "Test Statistic =" with a value of 0.130. A "Conclude:" label is followed by the text "No evidence against normality." At the bottom, there are five buttons: "Cancel", "Reset", "Print", "Compute" (which is highlighted with a dashed border), and "Return".

Test	Statistic	Value
Shapiro-Wilks	W	0.9394
Shapiro-Wilks	Probability	0.2339
Lilliefors	Skewness	0.970
Lilliefors	Kurtosis	1.802
Lilliefors	Test Statistic	0.130

Figure 7 Normality Test Dialog

In this example we have utilized the `cansas.LAZ` file and analyzed the weight variable. The two tests both support the assumption that weight is obtained from a normally distributed population.

X Versus Y Plot

One of the best way to examine the relationship between two variables is to plot the values of one against the other. We have selected the cansas.LAZ file and have plotted two of the variables. Shown below is the dialog form for this procedure. You can see the variables that have been selected and the options for the output that have been selected.

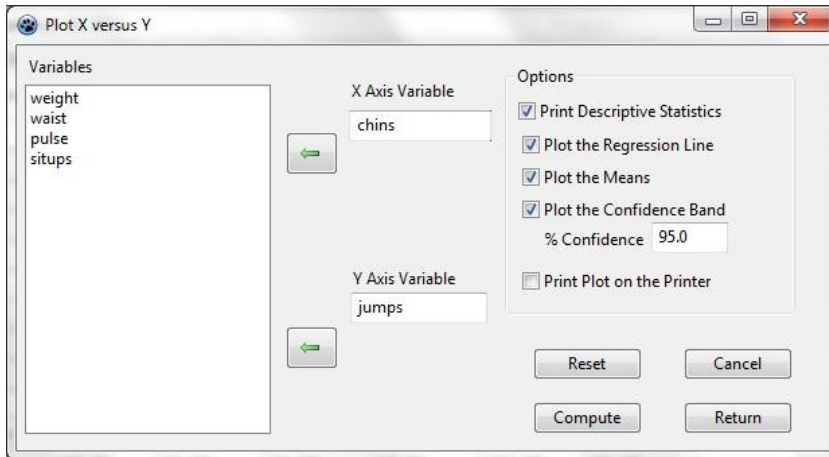


Figure 8 X Versus Y Dialog

The output obtained when you click the OK button is shown below:

X versus Y Plot

X = chins, Y = jumps from file:

C:\lazarus\Projects\LazStats\LazStatsData\cansas.LAZ

Variable	Mean	Variance	Std.Dev.
chins	9.45	27.94	5.29
jumps	70.30	2629.38	51.28

Correlation = 0.4958, Slope = 4.81, Intercept = 24.86
Standard Error of Estimate = 45.75
Number of good cases = 20

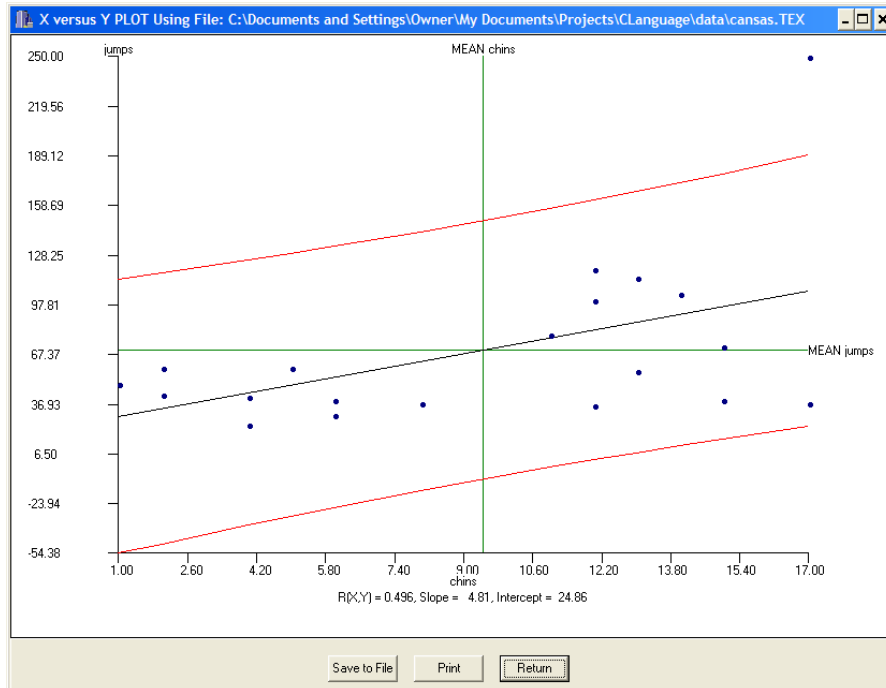


Figure 9 A Plot of Two Variables

The results indicate a moderate correlation of 0.496 with considerable scatter of points. In particular, notice the “outlier” at the Y value of 17 and the jumps of 250. Elimination of that point might change the correlation quite a bit. We also notice that the pattern of points does not seem to form a symmetric oval that is expected for a bivariate-normal distribution. Notice the values below the means form a somewhat flat distribution while those above the mean for chins is more rounded. One could speculate that there might be a curvilinear relationship between these two variables. The two red curves on the border of the plots indicate the 95% confidence limits. Notice the point we mentioned lies quit a bit outside this interval.

Group Frequency Histograms

When data values have been classified as members of various groups, one can obtain a plot of the frequency of cases in each group. The frequency variable should be defined as an integer variable, typically with values from 1 to the highest group number. We have selected the file chisqr.LAZ as an example in which cases have been classified into both rows and columns. In our example we have chosen to plot the frequency of cases in the various columns and have chose the three dimensional vertical plot.

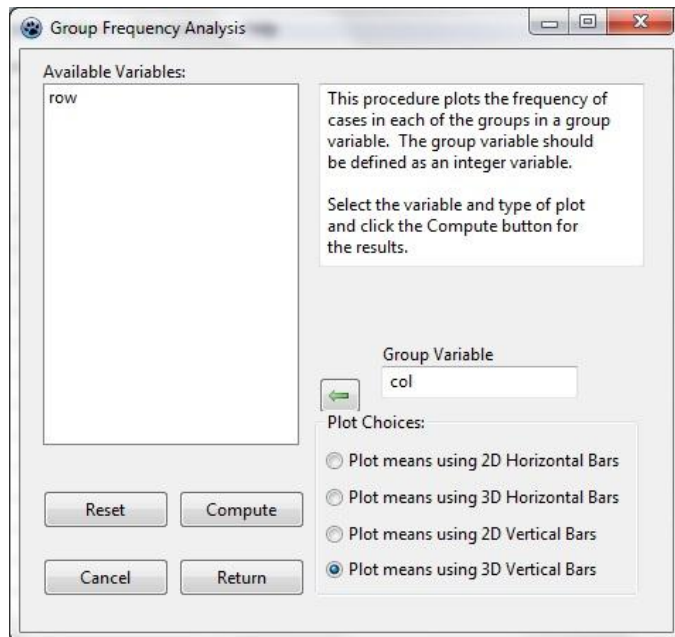


Figure 10 Specification Dialog for a Frequency Analysis

The plot obtained is:

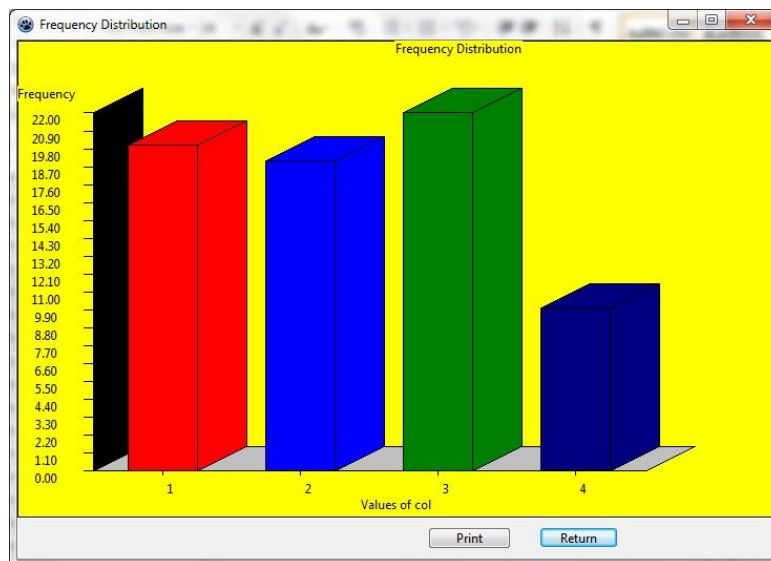


Figure 11 A Sample Frequency Plot

Repeated Measures Bubble Plot

Teachers, physicians, economists and other professionals often collect the same measure repeatedly over time for various classes of subjects. One of the ways to examine trends in this data is to plot these repeated values with bubbles that are colored for the groups. In our example we are going to use some school data that shows achievement of students as a function of both the year the data were collected and the ratio of teachers to students. Our specifications are shown in the following dialog:

Repeated Measures Bubble Plot

Directions:

1. Select the variable containing the bubble Identification number - an integer in the range of 1 to N objects.
2. Select the variable representing the X axis integer value for the object. This is the repeated measures variable.
3. Select the variable representing the Y axis. This should be a floating point value.
4. Select the variable representing the size of the bubble for each object to be plotted at the X and Y locations.

Note: Each data line represents one replication (X value) of the object to be plotted. See the example data file labeled BubblePlot.tex

Available Variables

Bubble Identification Number Variable: school [Reset]

X Value Variable: Year [Compute]

Y Value Variable: Achieve [Cancel]

Bubble Size Variable: Ratio [Return]

Main Title: School Achievement Over Time

Your X Label: School ID Your Y Label: Achievement

Options:

☐ Transform Data Grid for ANOVA (Treatments by Subjects ANOVA)

Figure 12 Repeated Measures Bubble Plot Dialog

When the Compute button is clicked, the following plot is obtained:

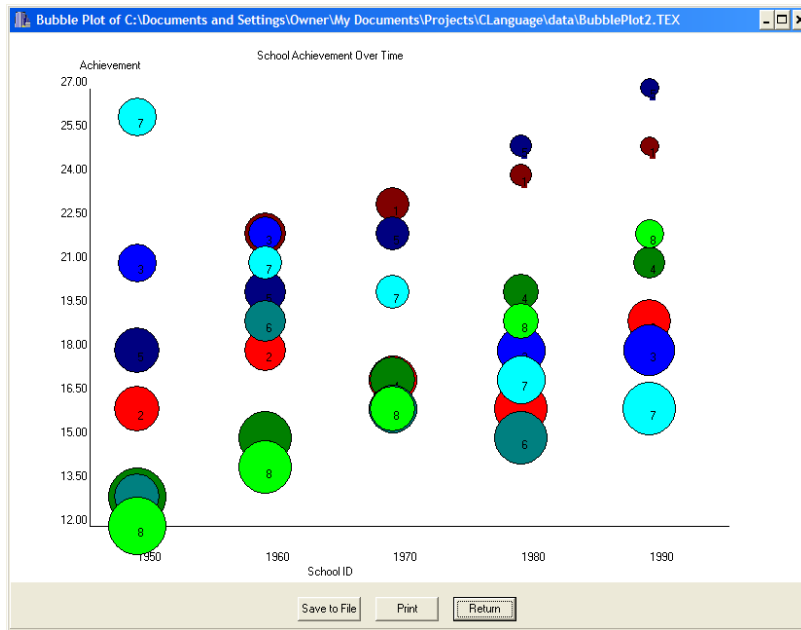


Figure 13 Bubble Plot of School Achievement

Notice in this plot that as the number of students to teacher ratio increases, the achievement goes down (group 7 as an example.) Also notice the increase in achievement as the ratio decreases as demonstrated by group 8. One would most likely want to obtain the correlation between the ratio and achievement across all the years! Additional output obtained is:

MEANS FOR Y AND SIZE VARIABLES

Grand Mean for Y = 18.925
Grand Mean for Size = 23.125

REPLICATION MEAN Y VALUES (ACROSS OBJECTS)

Replication	1	Mean =	17.125
Replication	2	Mean =	18.875
Replication	3	Mean =	18.875
Replication	4	Mean =	19.250
Replication	5	Mean =	20.500

REPLICATION MEAN SIZE VALUES (ACROSS OBJECTS)

Replication	1	Mean =	25.500
Replication	2	Mean =	23.500
Replication	3	Mean =	22.750
Replication	4	Mean =	22.500
Replication	5	Mean =	21.375

MEAN Y VALUES FOR EACH BUBBLE (OBJECT)

Object	1	Mean =	22.400
Object	2	Mean =	17.200
Object	3	Mean =	19.800
Object	4	Mean =	17.200
Object	5	Mean =	22.400
Object	6	Mean =	15.800
Object	7	Mean =	20.000
Object	8	Mean =	16.600

MEAN SIZE VALUES FOR EACH BUBBLE (OBJECT)

Object	1	Mean =	19.400
Object	2	Mean =	25.200
Object	3	Mean =	23.000
Object	4	Mean =	24.600
Object	5	Mean =	19.400
Object	6	Mean =	25.800
Object	7	Mean =	23.200
Object	8	Mean =	24.400

We have plotted the ratio of student to teachers against achievement and obtained the following:

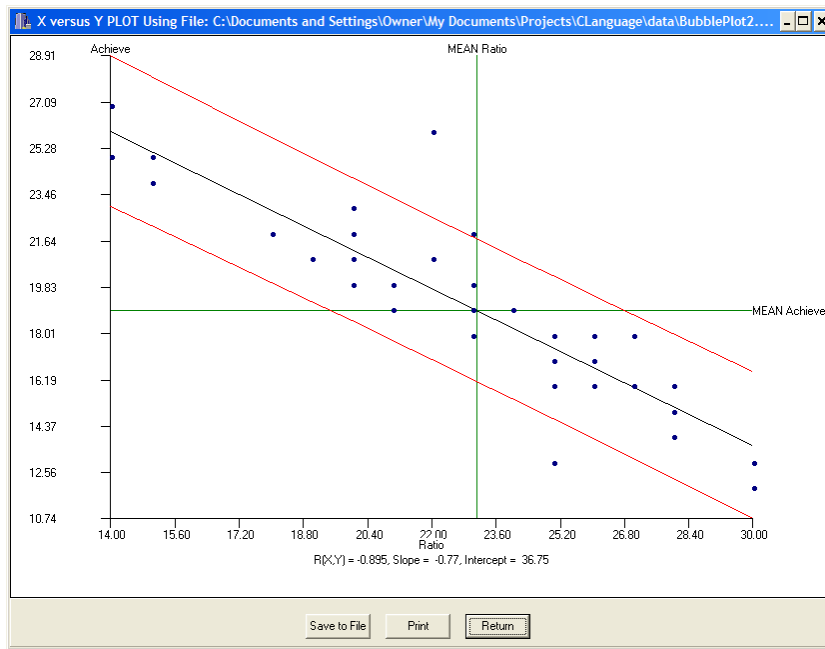


Figure 14 Plot of Teacher-Student Ratio to Achievement

The above plot verifies our bubble plot which suggested a high degree of relationship between these two variables. In effect, the bubble plot is a way of viewing three dimensions of your data. In the above example we viewed the relationship among achievement (the Y axis), year (the X axis), and student to teacher ratio (the bubble size) for a number of schools (the bubbles.) You may also want to consider the three dimensional plot procedure which lets you rotate your data around the X, Y or Z axis.

Comparisons With Theoretical Distributions

LazStats lets you view the distribution of your data against a theoretical distribution in several ways. This procedure lets you plot the cumulative distribution of your data values and show the theoretical cumulative distribution of a theoretical curve. In addition, you can also plot the frequency distribution of your values against the theoretical frequency distribution. A variety of theoretical distributions are available for comparison. We will demonstrate the use of this procedure to plot the same data used previously, that is, the weight variable from the cansas.LAZ file. Show below is the dialog form:

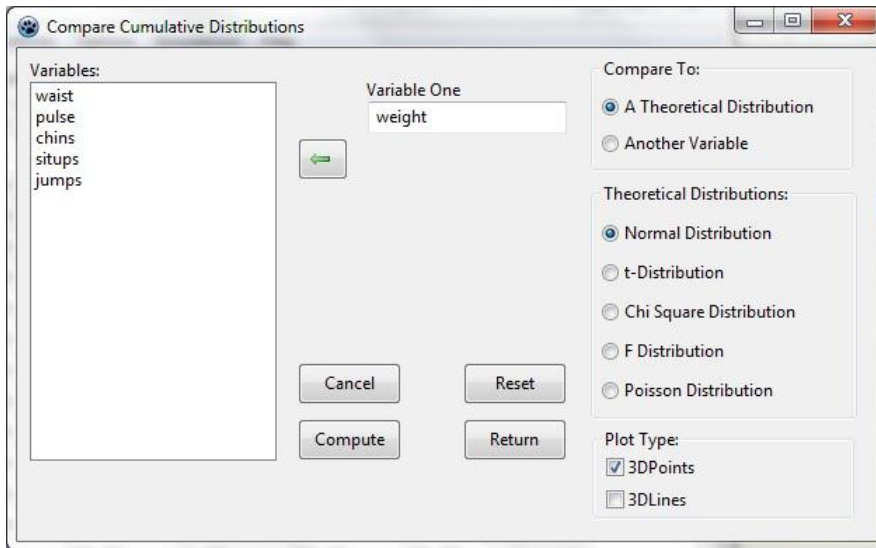


Figure 15 Comparison of Cumulative Distributions

The results are:

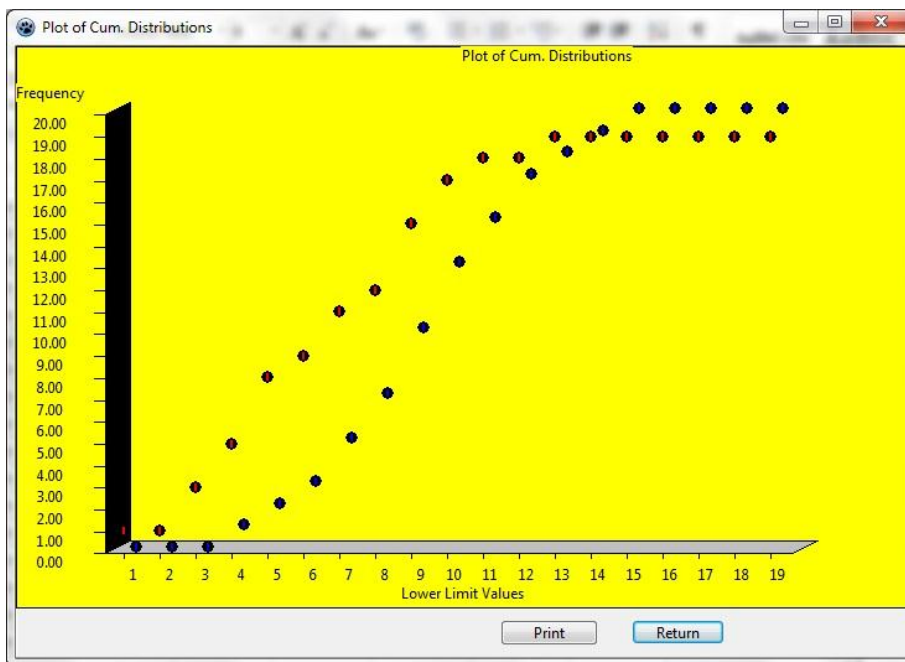


Figure 16 Cumulative Normal vs. Cumulative Observed Values

Notice that the observed data seem to follow the cumulative distribution of the normal curve fairly well.

The printout for the above analysis is:

Distribution comparison by Bill Miller					
weight	weight	weight	Normal	Normal	Normal
X1 Value	Frequency	Cum. Freq.	X2 Value	Frequency	Cum. Freq.
138.000	1	1.000	-3.000	0	0.000
144.000	0	1.000	-2.667	0	0.000
150.000	2	3.000	-2.333	0	0.000
156.000	2	5.000	-2.000	1	1.000
162.000	3	8.000	-1.667	1	2.000
168.000	1	9.000	-1.333	1	3.000
174.000	2	11.000	-1.000	2	5.000
180.000	1	12.000	-0.667	2	7.000
186.000	3	15.000	-0.333	3	10.000
192.000	2	17.000	-0.000	3	13.000
198.000	1	18.000	0.333	2	15.000
204.000	0	18.000	0.667	2	17.000
210.000	1	19.000	1.000	1	18.000
216.000	0	19.000	1.333	1	19.000
222.000	0	19.000	1.667	1	20.000
228.000	0	19.000	2.000	0	20.000
234.000	0	19.000	2.333	0	20.000
240.000	0	19.000	2.667	0	20.000

Kolmogorov Probability = 0.765763173908239, Max Dist = 0.222222222222222

Three Dimensional Rotation

One gains an appreciation for the relationship among two or three variables if one can view a plot of points for three variables in a 3 dimension space. It helps even more if one can rotate those points about each of the three axis. To demonstrate we have elected three variables from the cansas.LAZ file. Show below is the dialog and plot:

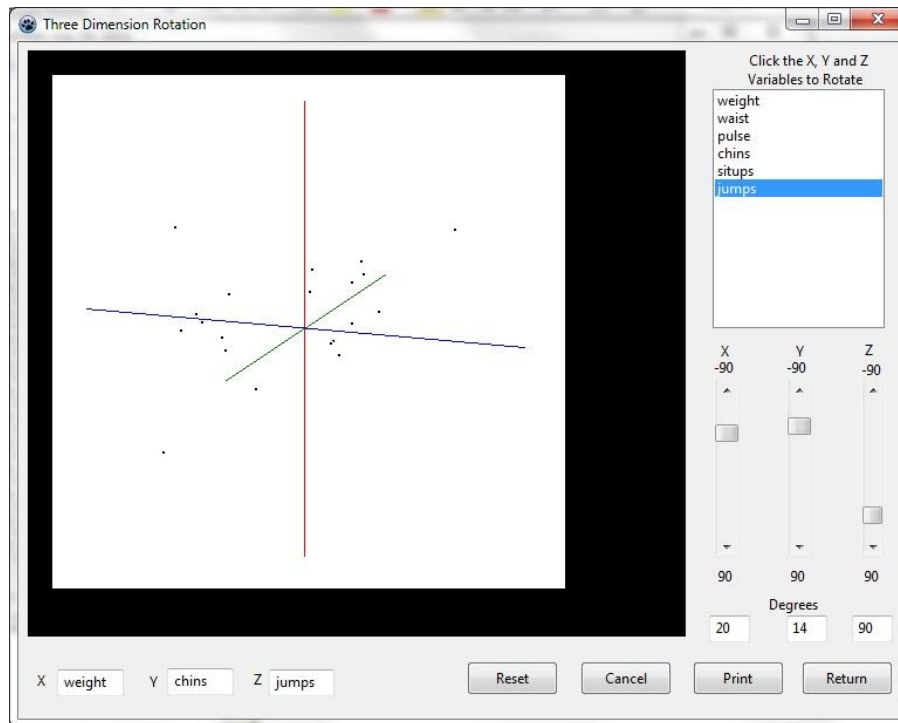


Figure 17 Scatter Plot of Values for Three Variables

You can place the mouse on one of the three “scroll” bar buttons (squares in the slider portion) and drag the button down while holding down the left mouse button. This will let you see more clearly the relationships among the three variables. To demonstrate, we have rotated the Y axis to -90 and the Z axis to nearly 0 degrees to examine the relationship between X and Y variables (weight and chins.)

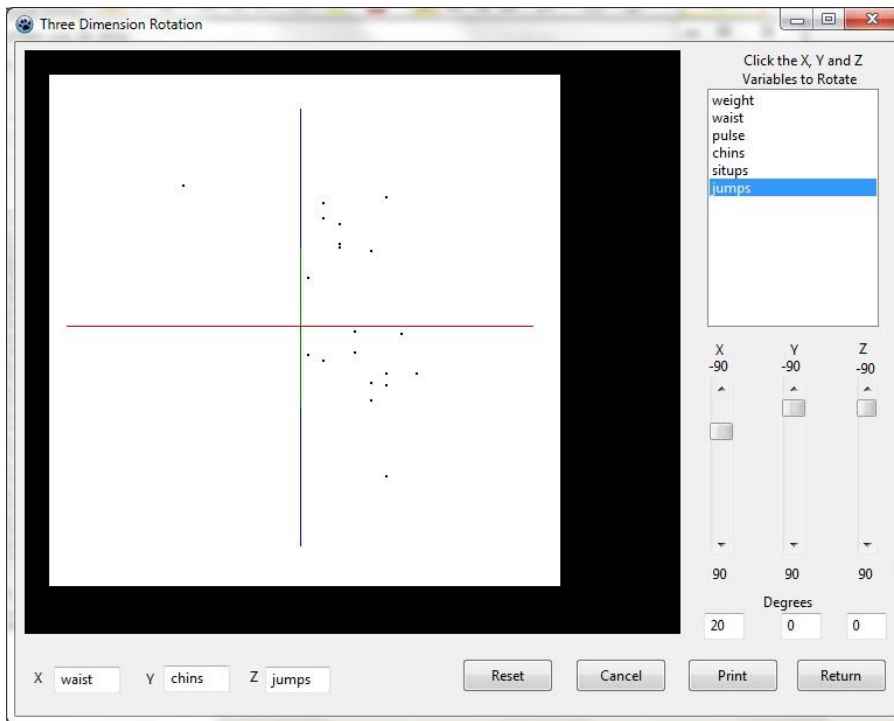


Figure 18 Rotated Variables to Examine Relationship Between Two Variables

Essentially, you can rotate the points around any one of the three axis until one of the axis is hidden. This lets you see the points projected for just two of the variables at a time.

Box Plots

Box plots are a way of visually inspecting the distribution of scores within various categories. As an example, we will use a file labeled anova2.LAZ which contains data for an analysis of covariance with row, column, slice, X, covar1 and covar2 variables. We have selected to do a box plot of the X variable (the dependent variable) for the three slice categories. Shown below is the dialog box for our analysis.

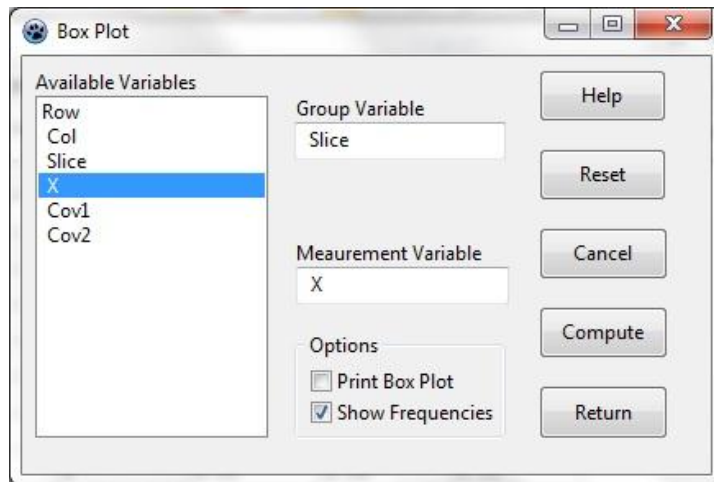


Figure 19 Box Plot Dialog

Since we have elected the option of showing the frequencies within each category, we first obtain the following output:

Box Plot of Groups

```
Results for group 1, mean =    3.500
Centile      Value
Ten          1.100
Twenty five  2.000
Median       3.500
Seventy five 5.000
Ninety       5.900
Score Range  Frequency Cum.Freq. Percentile Rank
```

0.50 - 1.50	2.00	2.00	8.33
1.50 - 2.50	2.00	4.00	25.00
2.50 - 3.50	2.00	6.00	41.67
3.50 - 4.50	2.00	8.00	58.33
4.50 - 5.50	2.00	10.00	75.00
5.50 - 6.50	2.00	12.00	91.67
6.50 - 7.50	0.00	12.00	100.00
7.50 - 8.50	0.00	12.00	100.00
8.50 - 9.50	0.00	12.00	100.00
9.50 - 10.50	0.00	12.00	100.00
10.50 - 11.50	0.00	12.00	100.00

```
Results for group 2, mean =    4.500
Centile      Value
Ten          2.600
Twenty five  3.500
```

Median 4.500
 Seventy five 5.500
 Ninety 6.400
 Score Range Frequency Cum.Freq. Percentile Rank

0.50 - 1.50	0.00	0.00	0.00
1.50 - 2.50	1.00	1.00	4.17
2.50 - 3.50	2.00	3.00	16.67
3.50 - 4.50	3.00	6.00	37.50
4.50 - 5.50	3.00	9.00	62.50
5.50 - 6.50	2.00	11.00	83.33
6.50 - 7.50	1.00	12.00	95.83
7.50 - 8.50	0.00	12.00	100.00
8.50 - 9.50	0.00	12.00	100.00
9.50 - 10.50	0.00	12.00	100.00
10.50 - 11.50	0.00	12.00	100.00

Results for group 3, mean = 4.250

Centile Value
 Ten 1.600
 Twenty five 2.500
 Median 3.500
 Seventy five 6.500
 Ninety 8.300
 Score Range Frequency Cum.Freq. Percentile Rank

0.50 - 1.50	1.00	1.00	4.17
1.50 - 2.50	2.00	3.00	16.67
2.50 - 3.50	3.00	6.00	37.50
3.50 - 4.50	2.00	8.00	58.33
4.50 - 5.50	1.00	9.00	70.83
5.50 - 6.50	0.00	9.00	75.00
6.50 - 7.50	1.00	10.00	79.17
7.50 - 8.50	1.00	11.00	87.50
8.50 - 9.50	1.00	12.00	95.83
9.50 - 10.50	0.00	12.00	100.00
10.50 - 11.50	0.00	12.00	100.00

You can see that the procedure has obtained the centiles and percentiles for the scores in each category of our slice variable. The plot for our data is shown next:

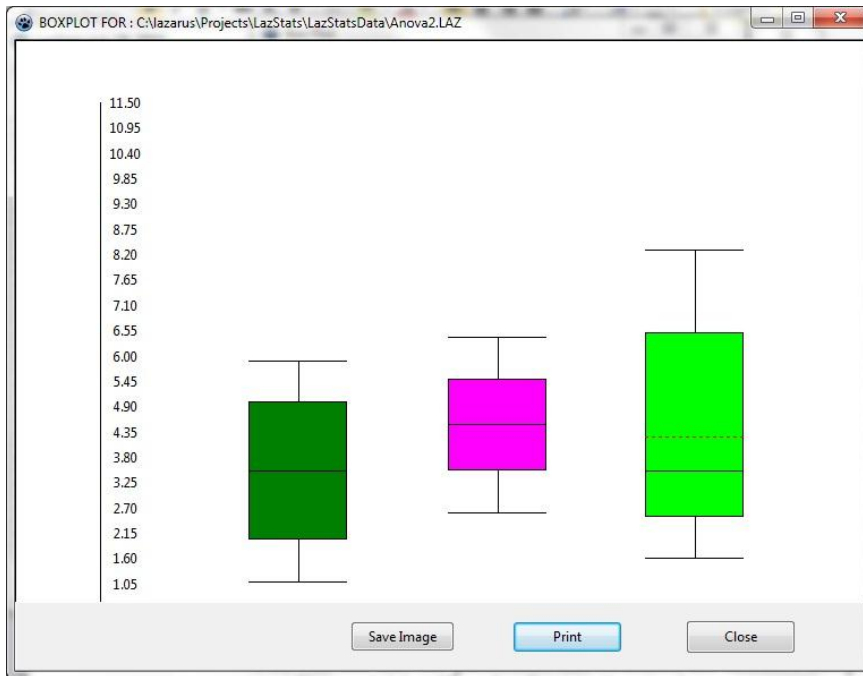


Figure 20 Box Plot of the Slice Variable

The “whiskers” for each box represent the 10th and 90th percentiles. The shaded box itself represents the scores within the interquartile range. The mean and the median (50th percentile) are also plotted. In the above plot one can see that there is skewed data in the third group. The mean and median are visibly separate. The mean is the dotted line and the median is the solid line.

Plot X Versus Multiple Y Values

One often has multiple dependent measures where the measures are on a common scale of measurement or have been transformed to z scores. It is helpful to visually plot these multiple variables against an X variable common to these measures. As an example, we will use a file labeled SchoolData.LAZ. We will examine the relationship between teacher salaries and student achievement on the Scholastic Aptitude Verbal and Math scores (N = 135.) The dialog form is shown below:

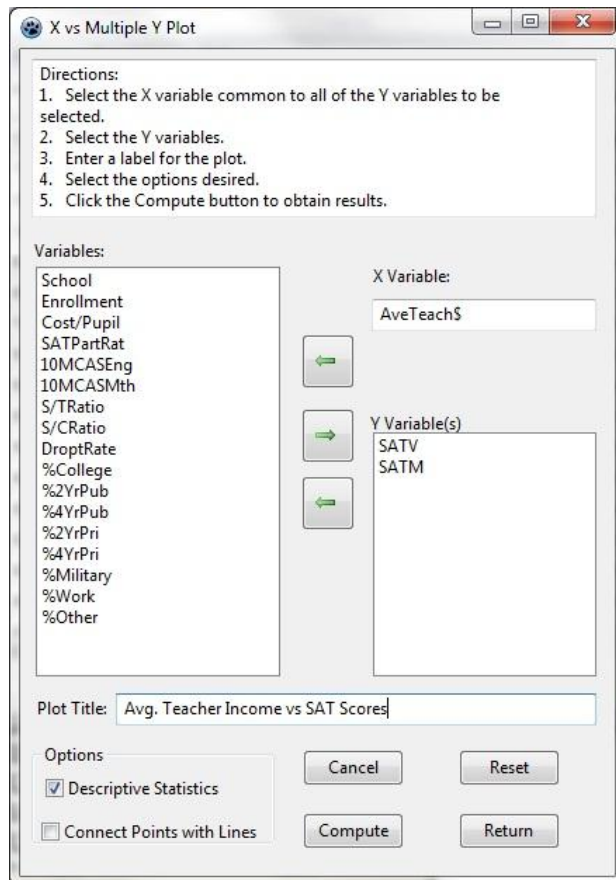


Figure 21 Plot X Versus Multiple Y Dialog

Since we chose the option to show related statistics, we first obtain:

X VERSUS MULTIPLE Y VALUES PLOT

CORRELATION MATRIX

Correlations			
	SATV	SATM	AveTeach\$
SATV	1.000	0.936	0.284
SATM	0.936	1.000	0.353
AveTeach\$	0.284	0.353	1.000

Means

Variables	SATV	SATM	AveTeach\$
-----------	------	------	------------

512.637 518.252 46963.230

Standard Deviations

Variables	SATV	SATM	AveTeach\$
	41.832	44.256	4468.546

No. of valid cases = 135

Next we get the plot:

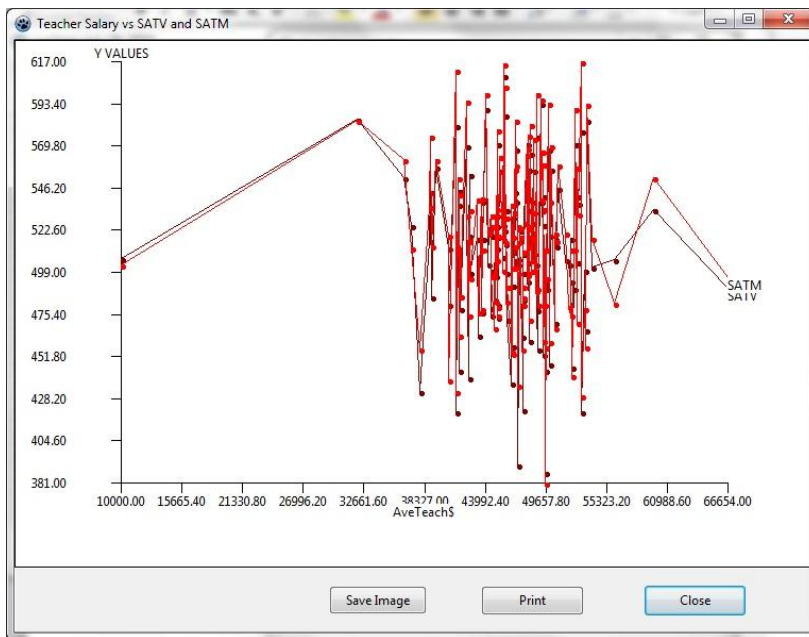


Figure 22 Teacher Salaries Versus SAT Achievement

We notice several things. First we notice how closely related the two SAT scores are. Secondly, we notice a trend for higher scores as teacher salaries increase. Of course, a number of explanations could be explored to understand these relationships.

Stem and Leaf Plot

The stem and leaf plot is one of the earlier ways to graphically represent a distribution of scores for a variable. It essentially reduces the data to the two most significant digits of each value, creates a “stem” for the first (leftmost) digit and “leaves” for the second digit. If there are a large number of “leaves” for a given stem, the representation of each leaf digit may have a “depth” of more than 1 value. This prevents the plot of the individual leaf values from spilling over the right edge of your output form. The stem and leaf does give a quick view of the distribution of many variables. The example we will use is from the SchoolData.LAZ file which contains 135 cases. We will create stem and leaf plots for three of the variables in this file. The dialog form is shown below:

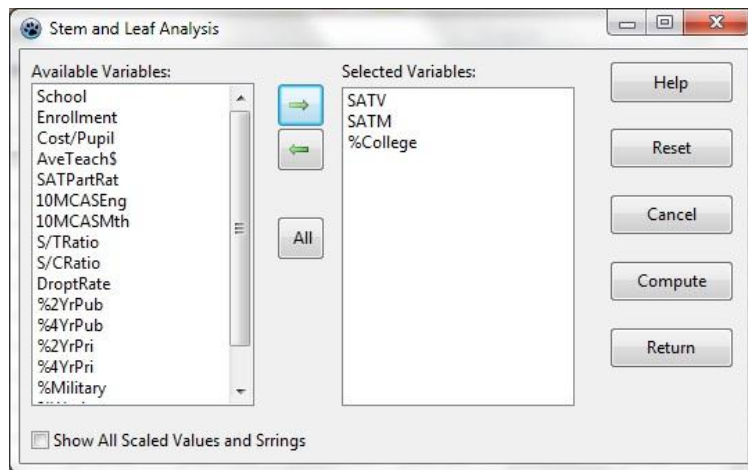


Figure 23 Stem and Leaf Plot Dialog

When we click the Compute button, we obtain:

STEM AND LEAF PLOTS

Stem and Leaf Plot for variable: SATV

Frequency	Stem	Leaf
2	3	89
10	4	223444
31	4	55667777888999999
68	5	0000000011111111222222233333334444
23	5	555566777889
1	6	0

Stem width = 100.00, max. leaf depth = 2
 Min. value = 387.000, Max. value = 609.000
 No. of good cases = 135

Stem and Leaf Plot for variable: SATM

Frequency	Stem	Leaf
1	3	8
5	4	334
40	4	555666777788888899999
57	5	00001111111111222222333333444
28	5	555666777889999
4	6	011

Stem width = 100.00, max. leaf depth = 2
 Min. value = 381.000, Max. value = 617.000
 No. of good cases = 135

Stem and Leaf Plot for variable: %College

Frequency	Stem	Leaf
1	5	9
2	6	34
5	6	58899
12	7	122223334444
25	7	555566667778888888999999
29	8	00001111111111222223333344444
30	8	555555666777777888889999999
27	9	000000011111222333444444444
3	9	556
1	10	0

Stem width = 10.00, max. leaf depth = 1
 Min. value = 59.000, Max. value = 100.000
 No. of good cases = 135

If we examine this last variable, we note that the stem width is 10. Now look at the top stem (5) and the leaf value 9. These are the two leftmost digits. We multiply the stem by the stem width to obtain the value 50 and then replace the second digit behind the first with the leaf value to obtain 59.. Now examine the previous plot for the SATM variable. The stem width is 100 so the first values counted are those with digits of 380. This we get by multiplying the stem width of 100 times the stem of 3 and entering the second digit of 8 behind the 3. We also note that the leaf depth is 1 in the last plot but is 2 in the previous plot. This indicates that each leaf digit in the last plot represents one value while in the previous plot each leaf represents one or two values. You might also note that the stems are “broken” into a lower half and upper

half. That is, if the second digit is 0 to 4 it is plotted in the lower half of the stem value and if 5 to 9 it is plotted in the upper half of the digits for that stem.

Multiple Group X Versus Y Plot

When you have obtained data on multiple groups that includes variables possibly related, you have several choices for viewing the data graphically. One would be to plot the two variables (e.g. X and Y) against each other in a traditional X vs. Y scatter plot. This would be repeated by first selecting one group at a time. Another option would be to concurrently plot X vs. Y for all of the groups. This procedure provides this last alternative. Our example uses the BubblePlot2.LAZ file. There are eight schools that have been sampled and we wish to plot the Student to Teacher ratio (our X variable) against the Achievement variable (our Y variable.) The dialog for specifying this analysis is shown below:

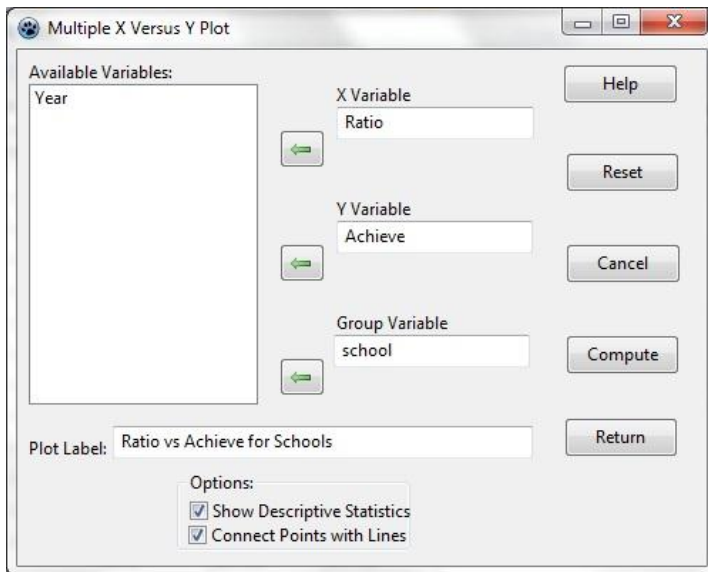


Figure 24 The Multiple Group X vs. Y Plot Dialog

When we click the OK button we obtain:

X VERSUS Y FOR GROUPS PLOT

VARIABLE	MEAN	STANDARED DEVIATION
X	23.125	4.268
Y	18.925	3.675

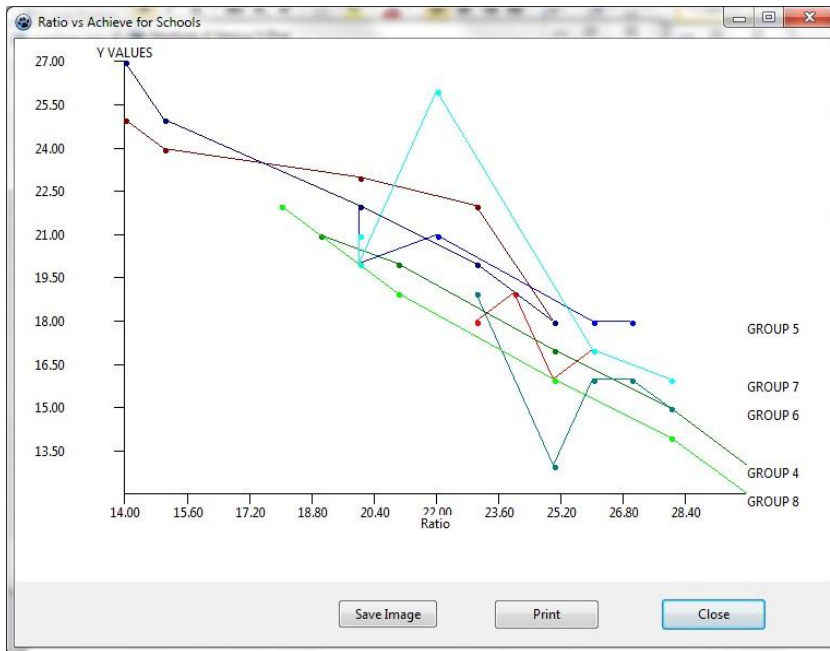


Figure 25 An X vs. Y Plot for Multiple Groups

We note the common relationship among all groups that as the Student to Teacher ratio increases, the achievement of students in the schools decreases. The trend is stronger for some schools than others and this suggests we may want to complete a further analysis such as a discriminant function analysis to determine whether or not the school differences are significant.

Resistant Line for Bivariate Data

Tukey (1970, Chapter 10) proposed the three point resistant line as an data analysis tool for quickly fitting a straight line to bivariate data (x and y paired data.) The data are divided into three groups of approximately equal size and sorted on the x variable. The median points of the upper and lower groups are fitted to the middle group to form two slope lines. The resulting slope line is resistant to the effects of extreme scores of either x or y values and provides a quick exploratory tool for investigating the linearity of the data. The ratio of the two slope lines from the upper and lower group medians to the middle group median provides a quick estimate of the linearity which should be approximately 1.0 for linearity.

To demonstrate, a file labeled “Sickness.LAZ” will be analyzed. The initial dialogue is shown below and the results follow that figure.

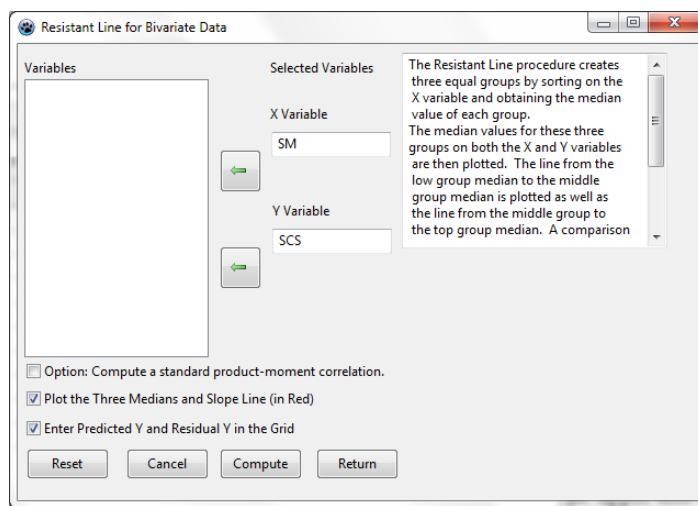


Figure 26 Resistant Line Dialogue Form

Group	X Median	Y Median	Size
1	109.500	189.600	3
2	122.000	216.250	4
3	132.700	228.200	3

Half Slopes = 2.132 and 1.117

Slope = 1.664

Ratio of half slopes = 0.524

Equation: $y = 1.664 * X + (-9.366)$

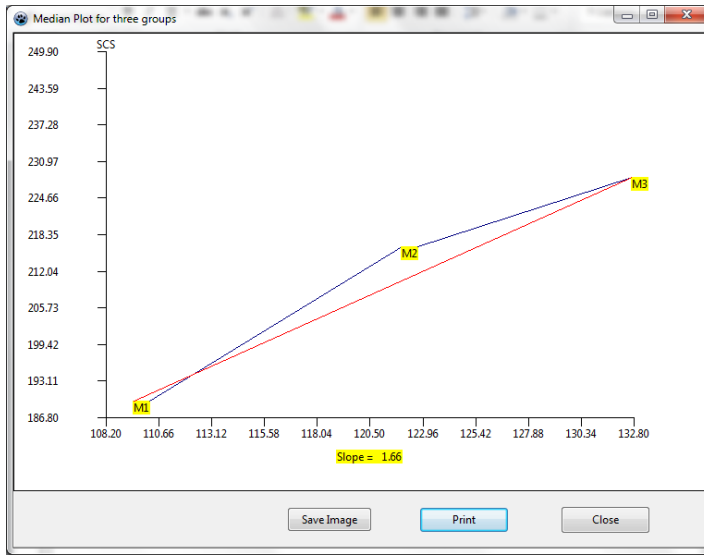


Figure 27 Plot of Resistant Line Slopes