

Chapter Five

Statistical Analysis

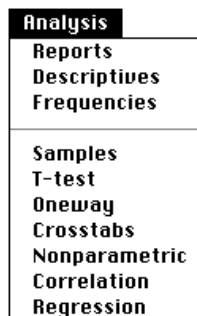
Overview

The statistical procedures and data manipulation utilities in FLO•STAT were chosen largely because they are the most frequently needed statistical procedures in the conduct of day-to-day business--selecting, sorting, generating small lists, calculating univariate summary measures of central tendency and dispersion, determining if there are differences among two or more groups of respondents, seeing if two variables are associated, and whether or not one variable can be used to predict another.

This manual is intended to provide the basic information for using FLO•STAT quickly and efficiently. It is not intended to be a statistical textbook and should not be used as such. There are many excellent statistical textbooks available on the market check with your local bookstore and you will find a wide assortment in such areas as applied statistics, biology, business, economics, mathematics, medicine, psychology, and sociology.

Statistical procedures

There are ten basic procedures in FLO•STAT, each of which can be accessed from either the icon menu bar or the **Analysis** menu.



Reports generates simple column reports on as many as four variables, providing univariate measures for each numeric variable in the list.



Descriptives calculates summary univariate measures (mean, standard deviation, variance and minimum and maximum score) and displays the results in a compact report format.



Frequencies generates frequency, percent, valid percent and cumulative percent distributions and univariate summary statistics.



Samples is a sampling exercise module for investigating the properties of sampling distributions drawn from populations with known distributions.



T-Test calculates a paired sample, independent sample, or one group T-test.



Oneway performs a one-way analysis of variance (ANOVA).



Crosstabs examines the joint distribution of two categorical variables, calculating the degree of association and a statistical test of independence, chi-square, Gamma, Cramer's V, Contingency Coefficient, and Z-score test.



Nonparametric contains a number of often used distribution free statistical measures.



Correlation measures the degree of association between two interval-level variables summarizing the strength and significance of association using the Pearson product-moment correlation, r .



Regression calculates the simple linear regression between the dependent variable and up to 25 independent variables, one at a time.

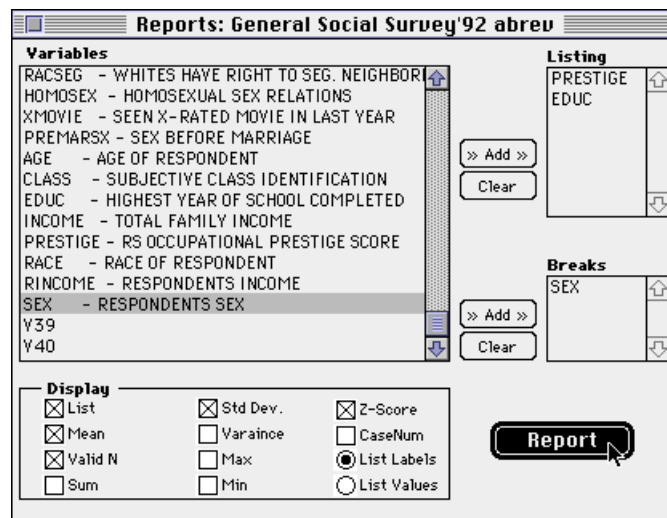
ANALYSIS	VARIABLE TYPE REQUIRED	
Reports	Numeric or Character	
Descriptives	Numeric	
Frequencies	Numeric or Character	
Samples	Numeric	
T-Test	<u>Group Variable</u>	<u>Dependent Var.</u>
Independent	Numeric or Character	Numeric
Paired	Not applicable	Numeric
One Group	Not applicable	Numeric
Oneway	<u>Independent Var.</u>	<u>Dependent Var.</u>
Crosstabs	Numeric or Character	Numeric
Nonparametric	Numeric or Character	Numeric or Character
Correlation	Numeric	Numeric
Regression	Numeric	Numeric

Reports

The **Reports** procedure provides a powerful, yet quick and simple way of obtaining column listing of cases. The procedure can be used in conjunction with the **Sort Cases** and **Case Selection** utilities to generate a variety of useful reports.

Select **Reports** from the **Analysis** menu, choose those variables to be included in the report and any break variables, if desired. Both numeric and character variables can be included in a report.. Cases will be listed within categories of break variables.

Display options include various univariate descriptive statistics, z-score values for all cases and variables selected, case number (i.e., data matrix row number), and a choice of displaying either the actual data values or the value labels. In the event value labels are not present, data values are automatically displayed.



Variables are displayed in their order of appearance in the data matrix. If a break variable is included in the report, as shown below, each case's variables are presented in their order of appearance within in categories of the break variable(s).

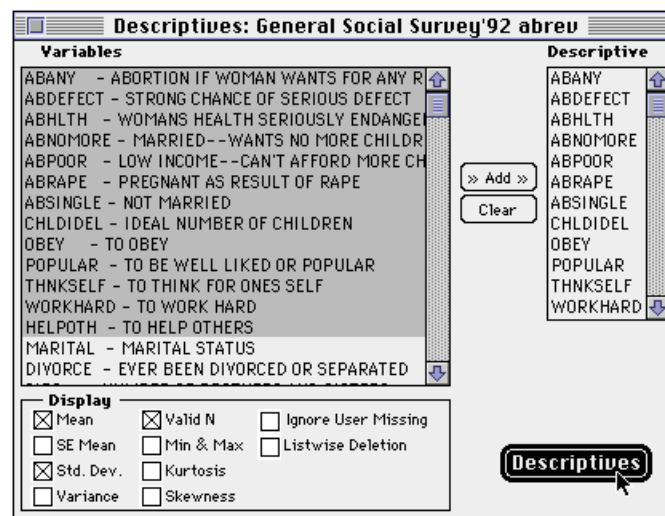
	PRESTIGE	Z-Score	EDUC	Z-Score
	56	0.848	10	-0.460
	45	0.227	15	0.073
	48	0.397	13	-0.140
Mean	40.969		14.312	
Valid N	96		96	
Std Dev.	17.732		9.380	
2.000=FEMALE	55	1.093	16	1.061
	36	-0.027	14	0.338
	25	-0.675	8	-1.834
	43	0.386	19	2.147
	36	-0.027	14	0.338
	46	0.562	12	-0.386

Descriptives

The **Descriptives** procedure is ideal for obtaining an overview of the distributional properties of the numeric variables in a data set. Summary univariate measures on as many as memory permits are arrayed in a compact column format.

Select **Descriptives** from the **Analysis** menu, add variables to the descriptives list, set the display options and click the Descriptives button.

Along with standard univariate descriptive measures and variable labels, user defined missing values can be ignored and those cases included in the calculation of the summary statistics or cases can be deleted in a listwise fashion (listwise deletion excludes an entire case if one variable in the list contains a missing value.)



Variable name, mean, standard deviation, valid number of cases, and variable label are the default display.

Variable	Mean	Std Dev	Valid N
ABANY	1.529	0.501	138
ABDEFECT	1.191	0.395	136
ABHLTH	1.065	0.247	139
ABNOMORE	1.457	0.500	140
ABPOOR	1.434	0.497	136
ABRAPE	1.078	0.269	141
ABSINGLE	1.474	0.501	135
CHLDIDEL	2.860	1.670	121
OBEY	3.317	1.378	126
POPULAR	4.421	0.861	126
THINKSELF	2.056	1.352	126
WORKHARD	2.397	1.028	126
HELPOTH	2.810	1.056	126

Frequencies

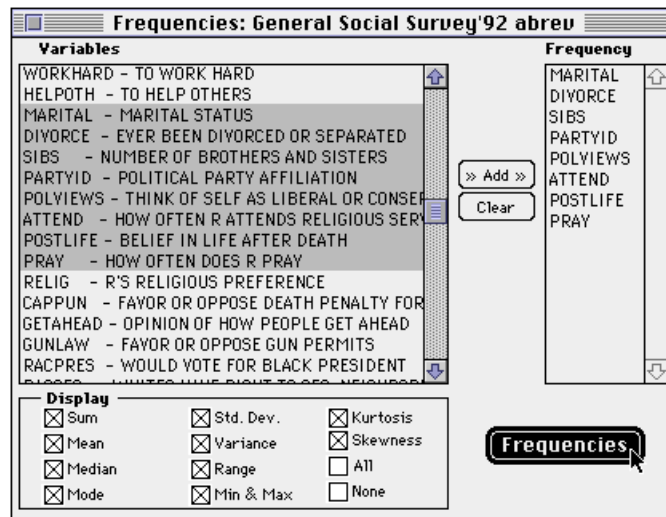
The **Frequencies** procedure provides additional information about the nature of each variable's distribution.

All values are shown for each variable, as well as each value's label (if available), frequency, percent, valid percent and cumulative percent.

Summary statistics includes: valid number of cases, number of missing cases, sum of all values across all cases, mean, median, mode, standard deviation, variance, range, kurtosis and skewness.

To run, select **Frequencies** from the **Analysis** menu. Select and add variables from the variables list on the left-hand side to the Frequency listing on the right-hand side of the Frequencies window.

Choose the desired univariate statistical measures from the display area at the bottom of the window. Click the **Frequencies** button when ready.



Variables are assigned separate pages in the output window. To view the table for each variable, click either arrow in the lower left-hand corner of the tabular output window, hit the left or right arrows on the keyboard, or select a specific variable from the **Table** menu, as shown below.

The screenshot shows the 'Tabular Output: General Social Survey'92 abbrev' window. The variable being analyzed is 'THINK OF SELF AS LIBERAL OR CONSERVATIVE'. The table displays the following data:

Value Label	Value	Frequency	Percent	Valid Percent
EXTREMELY LIBERAL	1	6	2.9	3.1
LIBERAL	2	39	19.1	20.0
SLIGHTLY LIBERAL	3	20	9.8	10.3
MODERATE	4	76	37.3	39.0
SLIGHTLY CONSERVATIVE	5	33	16.2	16.9
CONSERVATIVE	6	17	8.3	8.7
EXTRMELY CONSERVATIVE	7	4	2.0	2.1
NAP	0	5	2.5	Missing
		4	2.0	S. Missing
	Total	204	100.0	100.0

Summary statistics shown below the table:

- Valid cases: 195
- Missing cases: 9
- Sum: 743.000
- Mean: 3.810
- Median: 4.000
- Mode: 4.000
- Std. Dev.: 1.362
- Variance: 1.856
- Range: 6.000
- Min: 1.000
- Max: 7.000

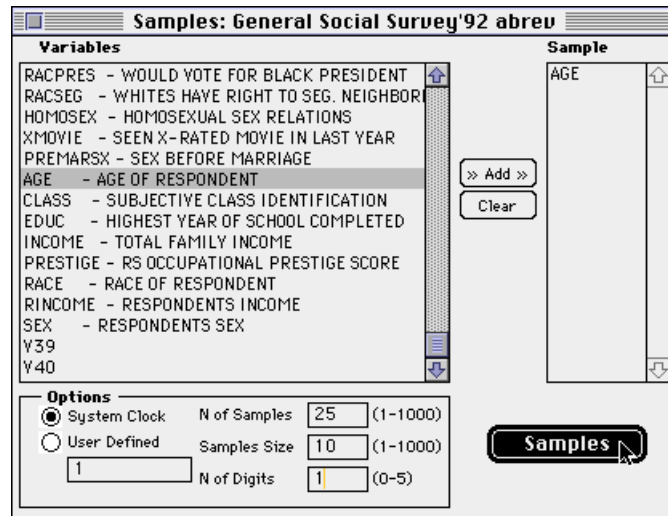
The 'Table' menu is open, showing a list of variables: SIBS, PARTYID, POLVIEWS (checked), ATTEND, POSTLIFE, and PRAY. The 'Table' menu is located in the upper right corner of the window.

Samples

Samples generates a series of random samples of N-size drawn from a population of the user's choice. Each sample mean is reported in a frequency distribution along with summary statistics (e.g., mean of means and standard deviation of means).

The **Samples** procedure is particularly useful when first learning about sampling theory and exploring the impact of sample size on characteristics of random samples. In an applied setting it provides a simple method for evaluating and exploring the implications of certain sampling decisions for a research project.

Select **Samples** from the **Analysis** menu., specify the N of samples, the size of each sample, and the precision to be used in reporting sample means. The pseudo-random number generator seed (i.e., a large number used to start the generation of random numbers) defaults to the computer's system clock. To generate a replicatable series of random numbers, enter a large integer, such as 123456789, instead. Variables selected from the list, in turn, serve as the population from which random sample are drawn.



Twenty five samples, each consisting of 10 cases, were drawn from a data set consisting of 200 cases. The system clock was used as the seed for random number generator.

The output includes the frequency, cumulative frequency, percent and cumulative percent distributions of the 25 sample means as well as the N of samples, sample size, sample mean, standard deviation of the distribution of sample means (standard error), the sum of means, and the population's mean and standard deviation.

SAMPLED VARIABLE= AGE

AGE OF RESPONDENT

	Value	Frequency	Cum. Freq	Percent	Cum Per
	51.1	1	20	4.0	
	51.4	1	21	4.0	
	52.5	1	22	4.0	
	53.3	1	23	4.0	
	54.5	1	24	4.0	
	56.4	1	25	4.0	

N of Samples 25
 Sample Size 10
 Sample Mean 45.904
 Sample Std 5.679
 Sample Sum 1147.6
 Var: Mean 47.040
 Var: Std 17.938

T-Test

The T-Test procedure computes Student's t used in determining whether the difference between two sample means is statistically different.

Three types of t-tests are available:

test for independent samples--e.g., test the difference in men and women's occupational prestige scores. (In the event the criterion variable contains more than three or more categories, two categories can be specified and entered in the Group 1 and Group 2 fields.)

test for paired samples--e.g., test whether a group of students, given the same exam twice, obtained scores on the second exam which were significantly different than the first. (No criteria variable is used in this test. T-tests are computed among all combination of variables added to the T-Test vars list.)

one group test--e.g., test the difference between a national mean score and the score obtain from a local experimental group. (No criteria variable is used for this test, instead, a one group mean score is compared with each variable in the T-test vars list.)

T-test: General Social Survey'92 abrev

Variables

- RELIG - R'S RELIGIOUS PREFERENCE
- CAPPUN - FAVOR OR OPPOSE DEATH PENALTY FOR
- GETAHEAD - OPINION OF HOW PEOPLE GET AHEAD
- GUNLAW - FAVOR OR OPPOSE GUN PERMITS
- RACPRES - WOULD VOTE FOR BLACK PRESIDENT
- RACSEG - WHITES HAVE RIGHT TO SEG. NEIGHBOR
- HOMOSEX - HOMOSEXUAL SEX RELATIONS
- XMOVIE - SEEN X-RATED MOVIE IN LAST YEAR
- PREMARSX - SEX BEFORE MARRIAGE
- AGE - AGE OF RESPONDENT
- CLASS - SUBJECTIVE CLASS IDENTIFICATION
- EDUC - HIGHEST YEAR OF SCHOOL COMPLETED
- INCOME - TOTAL FAMILY INCOME
- PRESTIGE - RS OCCUPATIONAL PRESTIGE SCORE
- RACE - RACE OF RESPONDENT
- RINCOME - RESPONDENTS INCOME
- SEX - RESPONDENTS SEX

T-Test Type

- Independent-Sample T-Test
- Paired-Sample T-Test
- One Group T-Test

T-Test Vars

- CHLDIDEL
- PRESTIGE
- INCOME

Criterion Values

- SEX

Groups

Group 1:

Group 2:

Use Group:

T-Test

At the top of the page, the results show the independent variable's group values and labels. Each group's number of cases, mean, standard deviation and standard error are shown along with the T-value, degrees of freedom, and both one and two tailed tests of significance.

Tabular Output: General Social Survey'92 abbrev

Group1 - SEX= 1.000 MALE
Group2 - SEX= 2.000 FEMALE

Variable	N	Mean	Std.	St. Error
PRESTIGE				
Group1	92	42.750	15.851	1.653
Group2	94	40.330	12.694	1.309
T-Value	Degrees of Freedom	1-Tail Test	2-Tail Test	
1.151	184	0.1257	0.2514	

Oneway

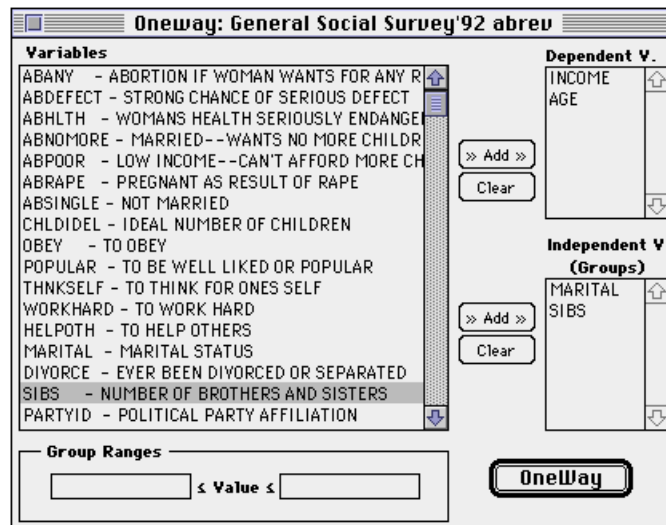
The ONEWAY procedure permits testing for differences among means from more than two independent samples. Rather than testing for differences between just two groups, say runners and non-runners, it may be theoretically advantageous to compare runners, swimmers, dancers and "couch potatoes" in terms of some performance test score to see if they differ.

The ONEWAY procedure simultaneously compares more than two sample means. The procedure is known as analysis of variance or ANOVA. The test is actually based on a comparison of group variances.

Select **Oneway** from the **Analysis** menu, add the dependent and independent variables to their appropriate lists and click the Oneway button.

To restrict the independent variable's range of values, enter the desired lower and upper limits of the categories in the Group Ranges fields.

Independent variable can be numeric or character.



Output from the analysis of variance provides the number of cases (count), mean, standard deviation, standard error, and the minimum and maximum score for each group as well as for the total sample.

The summary table contains the degrees of freedom and sum of squares for the between-groups, within-groups, and total sample. The within-groups and between-groups variances (**Mean SQR**) are provided along with the computed F ratio.

The screenshot shows a window titled "Tabular Output: General Social Survey'92 abbrev". The main content is a table with the following data:

Group	Count	Mean	Std.	Std. Error
1 MARRIED	87	11.690	0.906	0.097
2 WIDOWED	18	9.111	2.610	0.615
3 DIVORCED	16	9.438	3.444	0.861
4 SEPARATED	10	9.700	2.983	0.943
5 NEVER MARRIED	39	10.179	2.564	0.411
Total	170	10.741	2.270	0.174

Source	D.F.	Sum SQR	Mean SQR	F Ratio
Between Groups	4	176.43	44.11	10.4841
Within Groups	165	694.18	4.21	
Total	169	870.61		

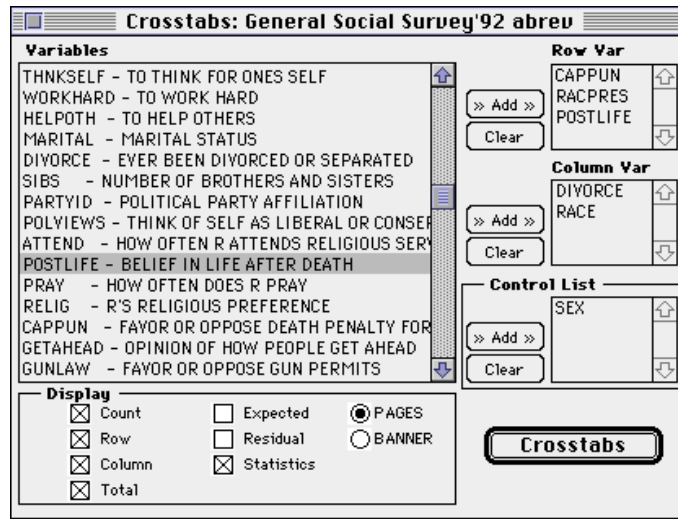
The bottom of the window shows a status bar with "1/4" and various navigation icons.

Crosstabs

Crosstabs is one of the most frequently used analytic procedures in science and business. It produces a bivariate (two variables) frequency distribution table which can be analyzed using tests of statistical significance and summarized with various measures of association.

Select **Crosstabs** from the **Analysis** menu.

Select the desired dependent, independent and control variables. Choose the tabular display options, including whether the tables should be displayed in single tables or together in a banner format.



By default each table includes cell count, row, column, and total percent. At the bottom of the page are several measures of statistical significance and association (see below).

CAPPUN\DIVORCE		1.000 YES	2.000 NO	Row Total
1.000	Count	10	36	46
FAVOR	Row %	21.74	78.26	83.64
	Col %	90.91	81.82	
	Tot %	18.18	65.45	
2.000	Count	1	8	9
OPPOSE	Row %	11.11	88.89	16.36
	Col %	9.09	18.18	
	Tot %	1.82	14.55	
Total Count		11	44	55
Total %		20.00	80.00	100.0
Chi-Square		0.531		
D.F.		1		
Sig.		0.466		
Gamma		0.379		
Cramer's V		0.0983		
Cont. Coeff.		0.0978		
Zscore		0.5954		
Missing		41		

Chi-square (χ^2) is used to test the null hypothesis that the two variables in question are not associated, in other words, they are independent. It is a test of statistical significance and does not measure strength of a relationship. Chi square should be used in conjunction with one of the accompanying measures of association. To the right of Chi square are its degrees of freedom and significance level.

Gamma is a symmetrical measure of association appropriate for two ordinal variables, ranging in value from -1.0 to +1.0, with 0 indicating no association. Gamma is a proportionate reduction in error (PRE) measure.

$$G = \frac{N_s - N_d}{N_s + N_d}$$

Where, N_s = the number of pairs of cases ranked in the same way on both variables, N_d = the number of pairs ranked in the opposite way on the two variables.

Crammer's V is a measure of association for nominal scale data and ranges in value from 0 to 1.0; with the larger the number the greater the association. Crammer's V is not a proportionate reduction in error measure.

Contingency Coefficient or Pearson's C is a measure of association developed primarily for square tables (e.g., 4x4) having more than two rows and columns. The coefficient ranges in value from 0 to less than 1. It is difficult to compare results from tables of different size since its value is determined in part by the number of rows and columns in the table.

Zscore (Test of Significance for Gamma [g]) is used to determine whether the population correlation of two ordinal level variables is different from zero.

When G is used as an estimate of the corresponding g parameter, it is necessary to test the null hypothesis, $g=0$, to evaluate the possibility that the computed G is merely due to sampling error.... Goodman and Kruskal have worked out a normal approximation of the sampling distribution of G which makes test of the null hypothesis possible (Goodman and Kruskal, 1963). They give the following formula for converting G to a standard score:

$$z = \left(\frac{G}{\sqrt{\frac{N_s + N_d}{N(1 - G^2)}}} \right)$$

Assuming that the null hypothesis is true, the g in the formula will be 0.

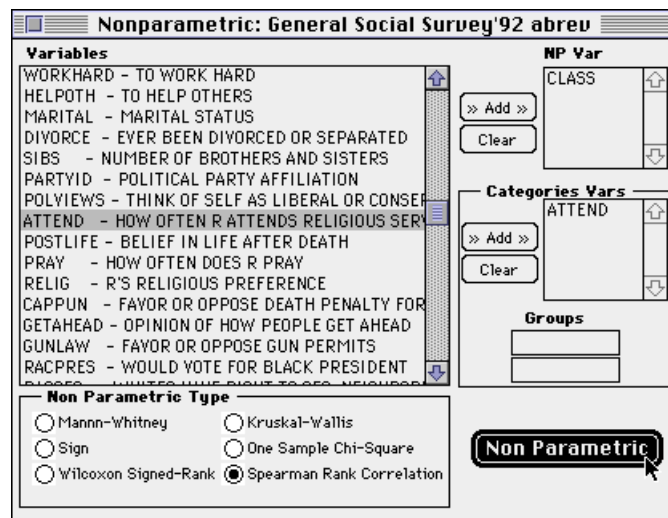
Substitute the necessary sample data into the formula, solve for z and then determine if the z score exceeds the z score that bounds the critical region of the sampling distribution at the .05 level. If the z score exceeds the critical level then the null hypothesis (i.e., $g = 0$) can be rejected.

Nonparametric

The procedures found in Nonparametric include a collection of tests which make few assumptions about the distributions of the data being examined. Most of these test are based on rankings and as such can require considerably more memory and computing time than many of Flo•Stat's other statistical procedures.

The measures available in Nonparametric include: the Mann-Whitney test, Sign test, Wilcoxon Signed-Ranks test, Kruskal-Wallis test, the One-Sample Chi Square test and the Spearman Rank Correlation.

As shown, once Nonparametric is selected from the Analysis menu, first set the desired nonparametric test type, enter the appropriate variables for that test and click the Nonparametric button.



Correlation

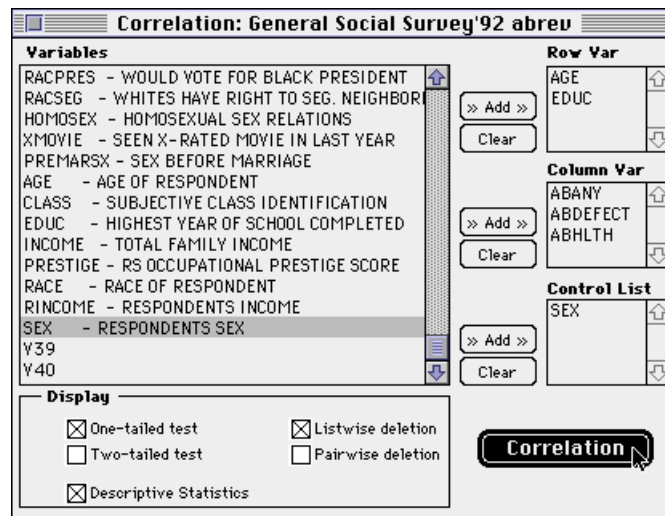
The Correlation procedure computes Pearson product-moment correlation coefficients (Pearson's r) plus partial correlations when control variables are specified.

Pearson's r is a measure of the linear relationship between two interval level variables. Pearson's r is a symmetrical measure of association ranging from -1.0 to $+1.0$ with 0 indicating no association. As a proportionate reduction in error measure, r^2 measures the proportion of variance in one variable explained by the other.

To run, select **Correlation** from the **Analysis** menu.

Analysis options include simple bivariate correlations when variables are entered in the Row and Column list, and partial correlations when control variables are added to Control list.

Display options include one or two-tailed test of significance, descriptive statistics, and listwise or pairwise case deletion.



The full tabular output includes separate tables for the zero-order correlation matrix, descriptive statistics, when requested, and first order and higher partial tables when control variables are specified.

Tabular Output: General Social Survey'92 abrev

ZERO-ORDER CORRELATIONS

Table: Zero-Order Table
 Descriptives Table
 First Order Table: SEX

	ABANY	ABDEFECT	ABHLTH	SEX
AGE	0.10679 (126) P=0.11512	0.24882 (126) P=0.00231	0.25007 (126) P=0.00221	0.08248 (126) P=0.17733
EDUC	-0.20843 (126) P=0.00911	-0.21669 (126) P=0.00701	-0.11097 (126) P=0.10620	-0.01684 (126) P=0.42517
SEX	-0.17321 (126) P=0.02528	-0.00892 (126) P=0.46020	-0.08085 (126) P=0.18216	1.00000 (126) P=0.50000

1/3

Tabular Output: General Social Survey'92 abrev

PARTIAL CORRELATIONS

Controlling For: SEX

Table:

	ABANY	ABDEFECT	ABHLTH
AGE	0.12335 (125) P=0.08354	0.25042 (125) P=0.00226	0.25846 (125) P=0.00167
EDUC	-0.21462 (125) P=0.00769	-0.21688 (125) P=0.00716	-0.11271 (125) P=0.10353

3/3

Regression

The basic goal of simple linear regression is to determine the one straight line that best describes or fits the relationship between variable X (independent variable) and variable Y (dependent variable).

The best fitting line is obtained by the method of least squares - when the differences between the values of Y and the predicted values of Y for each X are squared and summed, and that sum is at its lowest possible value.

To run, select **Regression** from the **Analysis** menu.

Select the dependent (Y) and independent (X) variables from the variables list at the left.. Specify whether the regression is to be a simple or multiple regression.

If the simple regression option is selected, a regression analysis will be completed on each pair of variables between in the dependent and independent variable lists. If the multiple regression option is selected a regression equation consisting of all variables contained in the independent variables list will be calculated for each dependent variable.

Regression: General Social Survey'92 abbrev

Variables

- ABANY - ABORTION IF WOMAN WANTS FOR ANY R
- ABDEFECT - STRONG CHANCE OF SERIOUS DEFECT
- ABHLTH - WOMANS HEALTH SERIOUSLY ENDANGE
- ABNOMORE - MARRIED--WANTS NO MORE CHILDR
- ABPOOR - LOW INCOME--CAN'T AFFORD MORE CH
- ABRAPE - PREGNANT AS RESULT OF RAPE
- ABSINGLE - NOT MARRIED
- CHLDIDEL - IDEAL NUMBER OF CHILDREN
- OBEY - TO OBEY
- POPULAR - TO BE WELL LIKED OR POPULAR
- THNKSELF - TO THINK FOR ONES SELF
- WORKHARD - TO WORK HARD
- HELPOTH - TO HELP OTHERS
- MARITAL - MARITAL STATUS
- DIYORCE - EYER BEEN DIVORCED OR SEPARATED
- SIBS - NUMBER OF BROTHERS AND SISTERS
- PARTYID - POLITICAL PARTY AFFILIATION

Dependent V.

- CHLDIDEL
- POPULAR

Independent V.

- AGE
- EDUC
- PRESTIGE
- RINCOME

Regression Type

Simple Regression

Multiple Regression

Regression

Output from the regression procedure includes the following:

Simple R is the Pearson product-moment coefficient, or simply r .

R Squared is the square of Pearson's r and indicates the proportion of variance in the dependent variable explained by the independent variable.

Standard Error is the standard error of the estimate, or the standard deviation of Y values from predicted Y' values. The standard error may be interpreted as the average error in predicting Y using the regression equation.

Slope coefficient indicates the expected change in the dependent variable with a change of one unit in the independent variable. More simply, it refers to the slope of the regression line.

Intercept refers to the point where the regression line crosses the Y axis.

		Anova	DF	Sum of Squa...	Sum of Squa...
Simple R	0.09014	Regression	1	0.75321	0.75321
R Square	0.00812	Residual	124	91.95313	0.74156
Standard Error	0.86114				
Slope	0.004194	Beta	0.090137	F	1.01572
Intercept	4.225167	Std ErrorB	0.004161	Sig F	0.3155

Regression Equation:
POPULAR=4.225167+0.004194 * AGE

Beta is the standardized regression coefficient indicating the number of standard deviation units change in the dependent variable when the independent variable changes one standard deviation unit.

Standard Error of B is the standard error of the standardized regression coefficient.

F the ratio of the regression to the residual mean square.

Sig. F the probability of obtaining F .

The **Regression Equation** used to predict values of the dependent variable given values of the independent variable.

