Using SPSS for Windows

by Dr. Richard Wielkiewicz

College of Saint Benedict/Saint John's University

A Review of Correlation	*Starting Up SPSS*	*The SPSS Program*	*Data Input for SPSS*
Advanced Data Entry and File Handling	*Computing the Pearson Correlation*	*A Review of the t-test*	*The t-test For Independent Groups on SPSS*
The t-test For Dependent Groups on SPSS	*Analysis of Variance with SPSS*	*The One-Way ANOVA with SPSS*	*Factorial ANOVA with SPSS*
Chi-Square with SPSS	*Chi-Square Test for Goodness of Fit*	*Chi-Square Test for Independence*	*Transformations*
Exploratory Data Analysis	*Help Features*	*Reliability Analysis*	*Moving Output to Other Applications*

SPSS for Windows is the Windows version of the Statistical Package for the Social Sciences. It is one of the most useful, popular, and easy-to-use software packages for performing statistical analyses. Familiarity with SPSS may be an important step in your professional or educational advancement. The purpose of this site is to explain the basics of using the program beginning with computing a correlation between two variables and continuing with t-tests, ANOVAs, and chi-square.

This section has a dual purpose. One purpose is to review the basics of computing and interpreting a correlation coefficient using SPSS. The second purpose is to explain the basics of entering data into the SPSS program.

A Review of Correlation

Remember that a correlation coefficient provides a measure of the degree of linear relationship between two variables. Generally, correlations are computed between two *different* variables that have each been measured on the same group of people. Each person in the sample provides a score on each of the two variables. For example, a researcher might be interested in the relationship between current college GPA and the number of hours the student studies in an average week in the middle of the school year. Hopefully, the result would be a positive correlation with higher GPAs associated with more hours of study time. A data from designed to summarize data for such a study might look like this:

Participant	College GPA	Weekly Study Time
Participant #01	1.8	15 hrs
Participant #02	3.9	38 hrs
Participant #03	2.1	10 hrs
Participant #04	2.8	24 hrs
Participant #05	3.3	36 hrs
Participant #06	3.1	15 hrs
Participant #07	4.0	45 hrs
Participant #08	3.4	28 hrs
Participant #09	3.3	35 hrs
Participant #10	2.2	10 hrs
Participant #11	2.5	6 hrs

Notice that there are eleven participants, each having a score on both the GPA variable and the Weekly Study Time variable. Participants in such studies are normally given a number as shown here in order to protect their confidentiality. SPSS also uses the spreadsheet format shown here. Each **row** of the spreadsheet is called a **CASE** which is almost always one of the participants in the study. Each **column** of the spreadsheet is used to store a particular bit of information about the participant, such as GPA and Weekly Study Hours, as shown here, or any other information relevant to the study. Thus, each column has a different variable with a value for each person or case in the study. Complex studies may have thousands of participants and hundreds of variables.

Back to the Top of the Page

Starting Up SPSS

SPSS is usually part of the general network available in the computer labs and residence halls of most college campuses. To activate SPSS, sign-on to the network with your username and password. Then click the **Start** icon in the lower left-hand corner of the screen followed by **Network Programs > SPSS for Windows** > **SPSS for Windows**. If SPSS is not found on the "Network Programs" group, it may be installed as a "local program" in which case the proper sequence is **Start > Programs > Local Programs > SPSS 8.0 for Windows > SPSS 8.0 for Windows**. Another possibility is that a shortcut already exists on the desktop, in which case double-clicking it will open the SPSS program.

The SPSS Program

After clicking the SPSS icon, there is a short wait and the SPSS program appears. SPSS for Windows begins with two windows. The top window offers several options which may be useful eventually, but the easiest thing to do is close the top window which then gives access to the main program. At this point one of the most sophisticated and popular data analysis programs is available. Thanks to a user-friendly interface, it is possible to do almost anything from the most simple descriptive statistics to complex multivariate analyses with just a few clicks of the mouse. The program is also quite "smart" in that it will not execute a procedure unless the necessary information has been provided. Although it can be frustrating when working with complex procedures, it saves a lot of time in the long run because the feedback is immediate and corrections can be made on the spot.

Back to the Top of the Page

Data Input for SPSS

SPSS appears on the screen looking like most other Windows programs. Two windows are initially available: the data input window and the output widow. When SPSS first comes up, it is ready to accept new data. To begin entering data, look at the menu options across the top of the screen:

File Edit View Data Transform Statistics Graphs Utilities Wndow Help

Clicking on one of these options opens a menu of related options, many of which will not be available until enough information has been provided to allow the procedure to run. To begin the process of a computing a correlation, click on the **Data** option, then click on **Define Variable**. This will open an input window that allows you to define the first variable by giving it a name and other information that will make it easier to use the variable in statistical analyses and interpret the output. When this window is opened the default name for the variable is displayed and highlighted. Just type a name for the first variable that uses less than eight characters. For example, the first variable in the above example could be called **colgpa**, a name that is less than eight characters and gives a good indication of the nature of the variable (college GPA). It is also useful to have more information about the variable and this can be done by clicking on the **Labels...** button which appears at the bottom of the window. This button opens another window which allows you to add more information about the variable, including an extended label, such as **College GPA for 1999**. You can also add what are called **value labels** using this same window. Value labels allow you to give names to particular values of a nominal or categorical variable. For example, most studies have a variable called Sex that can take on two values, 1 = Female or 2 = Male. The value labels option allows you to have these labels attached to all the output from statistical analyses which simplifies interpretation and reporting. Entering value labels also means you don't need to remember how the variable was coded (i.e., whether males were coded 1 or 2) when you view the output. After entering a variable name and value labels for the first variable, close the **Labels...** input window by clicking the **Continue** button. Then click the **OK** button on the Define Variable window. The next step is to use the mouse and left mouse button or the arrow keys to reposition the cross to the first cell in the second column of the data input spreadsheet. Then define the second variable using the same process. Continue defining variables until all the variables have been defined.

Once the variables have been defined, the data can be entered into the spreadsheet. (These tasks can be done in the opposite order, as well.) This requires working with the Newdata or spreadsheet window. To begin, make sure the cursor is flashing at the top of the spreadsheet window and that the upper left cell of the spreadsheet is highlighted. To highlight a cell use the mouse to move the cross to the desired cell of the spreadsheet and click the left mouse button. The arrow keys also work well to navigate around the spreadsheet. Now begin entering data by typing the first piece of data. In the above example the first entry would be **1.8**. This number will appear at the top of the spreadsheet. Hit to move the data into the correct cell. Notice that after hitting the second cell in the first column is now highlighted. The next piece of data (**3.9**) can be entered using the same procedure. Thus, data is automatically entered vertically.

Continue until all the data for the first column have been entered. After entering all the data for the first column, use the mouse or arrow keys to highlight the first cell in the second column and begin entering the second column of data using the same technique. If a piece of data is missing (e.g., the participant did not answer one or more of the questions on a survey), simply hit when the input cell at the top of the spreadsheet is empty. This will cause a dot to appear in the spreadsheet cell which is interpreted by SPSS as missing data. SPSS has very flexible options for handling missing data. Usually, the default or standard option is the best one to use.

In larger studies with a lot of variables, it may be more convenient to go across or horizontally, entering all the data for the first participant followed by all the data for the second participant, continuing until all the data have been entered. In order to do this it will be necessary to make more frequent use of the mouse and left button or the arrow keys to highlight the next cell going across. When data for a large study is being entered, it is best to work with a partner. One person can read the data and the other can type. This greatly increases speed and accuracy.

Back to the Top of the Page

Advanced Data Entry and File Handling

Sometimes a researcher begins with an ASCII file created manually or by optical scanning. An ASCII data file should have lines of no more than 80 columns with all the data for the first participant followed by all the data for the next participant and so on until all the data has been entered. Each case or participant should begin with a new line and each variable should be in the exact same location for each participant. For example, biological sex may be coded in the first line, fifth column for **each** participant. There are other ways to format data for SPSS input but this is the most common and probably the most useful.

To convert the ASCII file to an SPSS file, click **File > Read ASCII Data...** and select the appropriate file. Then click the Define button and a window will open that will allow you to specify the name and exact location of each variable. "Record" refers to the line of data. After completing all four boxes click **ADD** and go to the next variable, continuing until all variables have been defined. Then click **OK** and the window closes showing the spread-sheet with all the data read into it. Whether you have started with an ASCII file or entered data directly into the spreadsheet, the data file can be saved as an SPSS file that can be recalled at any time. The save operation should be repeated each time the file is permanently changed. I suggest that you maintain at least two backup copies, one which travels with you and a second which stays in a secure location.

Back to the Top of the Page

Computing the Pearson Correlation

After entering the data, the next step is to order the program to actually compute the correlation coefficient for you. Use the mouse to go to the top of the screen and click on the following sequence: **Statistics > Correlate > Bivariate**. This will open another input window. You will see two boxes with the one on the left containing the complete list of variables for the study. [Note: The variables will appear in alphabetical order which is the default variable display. However, it can often be more convenient to display the variables in the same order as they appear on the spreadsheet or input window. The display order can be changed by clicking **Edit > Preferences**. Then change "Alphabetical" to "File" by clicking the empty circle next to "File" under "Display Order for Variable Lists." Unfortunately, SPSS will need to be exited and then reloaded before this option will take effect.] The box on the right will be empty. In between the boxes is a right-pointing arrow. The sequence for computing a correlation is to highlight variables from the list on the left and then use the mouse to click the right-pointing arrow. This will cause each highlighted variable to jump to the box on the right. Each variable in the box on the right will be included in the correlation matrix computed by SPSS. Thus, in order to compute the correlation between COLGPA and STUDYHRS, move both variables over to the box on the right. A variable can be removed from the box on the right by highlighting it and clicking the arrow in the middle which will now face in the opposite direction. Once the variables you want to correlate are in the right-hand box, the **OK** button could be clicked which would cause the correlations to be computed and appear in an Output window. However, there are a couple of additional points worth considering.

First, it can be extremely helpful to click the **Options** button which appears at the bottom of the input window. This will cause another input window to appear. Generally, all options can be left on their default settings. However, one option allows you to print means and standard deviations for each variable in the analysis by just clicking the box. This is worth doing. The other options should be left alone unless you have a specific reason for changing one. At this point you must click the **Continue** button in order to close this box and move on with your task. The next step is simply to click the **OK** button. After a short delay, an Output window will appear with the results of your analysis. The information in the output file can be viewed or saved to a disk using standard Windows conventions. Additional analyses can be performed and their results will be appended to the end of the current output window so the results of a complex series of analyses can be contained in one output window. Be sure to give this file a name that will remind you of its contents. The results for the example are shown below:

	Mean	Std. Deviation	N
COLGPA	2.9455	.7285	11
STUDYHRS	23.8182	13.4001	11

Correlations

		COLGPA	STUDYHRS
COLGPA	Pearson Correlation	1.0000	.868**
	Sig. (2-tailed)	.	.001
	N	11	11
STUDYHRS	Pearson Correlation	.868**	1.0000
	Sig. (2-tailed)	.	.001
	N	11	11

. Correlation is significant at the 0.01 level (2-tailed). To interpret the output, look at the table labeled Correlations. This is a correlation matrix with three numbers for each correlation. The top number is the actual Pearson correlation coefficient which will range from -1.00 to +1.00. The further away the correlation is from zero, the stronger the relationship. The correlation between study hours and college GPA in this fictional study was .868 which represents an extremely strong relationship. The next number is the probability. Remember, you are looking for probabilities **less than .05 in order to reject the null hypothesis and conclude that the correlation differs significantly from a correlation of zero. The third number is the sample size, in this case 11. Correlation coefficients that can not be computed will be represented as a dot.

Another nice thing to do when computing a correlation is to look at the scatter diagram. To produce a scatter-plot, click **Graphs > Scatter > Define >**. Use the same technique as before to transfer variables to the x-axis and y-axis boxes. Then click **OK** and the graph will appear in the Chart Carousel window. To insert the plot in another document, click on **File > Copy Chart**, open your word processing document, and **Paste** it into the document.

Saving Output and Data Files

If you attempt to close either the data input or data output windows of SPSS, the program will respond with another window prompting you to save the file with either a user-supplied name or a generic name. Output files are given the extension, .spo, and data files are given the extension, .sav. The usual Windows conventions with respect to saving and reopening files apply using commands under the **File** menu.

Back to the Top of the Page

The t-test with SPSS

A Review of the t-test

The t-test is used for testing differences between two means. In order to use a t-test, the **same variable** must be measured in different groups, at different times, or in comparison to a known population mean. Comparing a sample mean to a known population is an unusual test that appears in statistics books as a transitional step in learning about the t-test. The more common applications of the t-test are testing the difference between independent groups or testing the difference between dependent groups.

A t-test for independent groups is useful when the same variable has been measured in two independent groups and the researcher wants to know whether the difference between group means is statistically significant. "Independent groups" means that the groups have different people in them and that the people in the different groups have not been matched or paired in any way. A t-test for related samples or a t-test for dependent means is the appropriate test when the same people have been measured or tested under two different conditions or when people are put into pairs by matching them on some other variable and then placing each member of the pair into one of two groups.

Back to the Top of the Page

The t-test For Independent Groups on SPSS

A t-test for independent groups is useful when the researcher's goal is to compare the difference between means of two groups on the same variable. Groups may be formed in two different ways. First, a preexisting characteristic of the participants may be used to divide them into groups. For example, the researcher may wish to compare college GPAs of men and women. In this case, the **grouping variable** is biological sex and the two groups would consist of men versus women. Other preexisting characteristics that could be used as grouping variables include age (under 21 years vs. 21 years and older or some other reasonable division into two groups), athlete (plays collegiate varsity sport vs. does not play), type of student (undergraduate vs. graduate student), type of faculty member (tenured vs. nontenured), or any other variable for which it makes sense to have two categories. Another way to form groups is to randomly assign participants to one of two experimental conditions such as a group that listens to music versus a group that experiences a control condition. Regardless of how the groups are determined, one of the variables in the SPSS data file must contain the information needed to divide participants into the appropriate groups. SPSS has very flexible features for accomplishing this task.

Like all other statistical tests using SPSS, the process begins with data. Consider the fictional data on college GPA and weekly hours of studying used in the correlation example. First, let's add information about the biological sex of each participant to the data base. This requires a numerical code. For this example, let a "1" designate a female and a "2" designate a male. With the new variable added, the data would look like this:

Participant	Current GPA	Weekly Study Time	Sex
Participant #01	1.8	15 hrs	2
Participant #02	3.9	38 hrs	1
Participant #03	2.1	10 hrs	2
Participant #04	2.8	24 hrs	1
Participant #05	3.3	36 hrs	.
Participant #06	3.1	15 hrs	2
Participant #07	4.0	45 hrs	1
Participant #08	3.4	28 hrs	1
Participant #09	3.3	35 hrs	1
Participant #10	2.2	10 hrs	2
Participant #11	2.5	6 hrs	2

With this information added to the file, two methods of dividing participants into groups can be illustrated. Note that Participant #05 has just a single dot in the column for sex. This is the standard way that SPSS indicates missing data. This is a common occurrence, especially in survey data, and SPSS has flexible options for handling this situation. Begin the analysis by entering the new data for sex. Use the arrow keys or mouse to move to the empty third column on the spreadsheet. Use the same technique as previously to enter the new data. When data is missing (such as Participant #5 in this example), hit the key when there is no data in the top line (you will need to the previous entry) and a single dot will appear in the variable column. Once

the data is entered, click **Data > Define Variable** and type in the name of the variable, "Sex." Then go to "value" And type a "1" in the box. For "Value Label," type "Female." Then click on **ADD**. Repeat the sequence, typing "2" and "male" in the appropriate boxes. Then click **ADD** again. Finally, click **CONTINUE >OK** and you will be back to the main SPSS menu.

To request the t-test, click **Statistics** > **Compare Means** > **Independent SamplesT Test**. Use the right-pointing arrow to transfer COLGPA to the "Test Variable(s)" box. Then highlight Sex in the left box and click the bottom arrow (pointing right) to transfer sex to the "Grouping Variable" box. Then click Define Groups. Type "1" in the Group 1 box and type "2" in the Group 2 box. Then click **Continue**. Click **Options** and you will see the confidence interval or the method of handling missing data can be changed. Since the default options are just fine, click **Continue > OK** and the results will quickly appear in the output window. Results for the example are shown below:

T-Test

Group Statistics

	Variable	N	Mean	Std. Deviation	Std. Error Mean
SEX	1.00 Female	5		.487	.218
	2.00 Male	5		.493	.220

Independent Samples Test

		Levene's Test for Equality of Variances	
		F	Sig.
SEX	Equal variances assumed	.002	.962
	Equal Variances not assumed		

		t-test for Equality of Means			
		t	df	Sig. (2-tailed)	Mean Difference
SEX	Equal variances assumed	3.68	8	.021	.1750
	Equal variances not assumed	3.68	8.00	.025	.1750

The output begins with the means and standard deviations for the two variables which is key information that will need to be included in any related research report. The "Mean Difference" statistic indicates the magnitude of the difference between means. When combined with the confidence interval for the difference, this information can make a valuable contribution to explaining the importance of the results. "Levene's Test for Equality of Variances" is a test of the homogeneity of variance assumption. When the value for *F* is large and the *P*-value is less than .05, this indicates that the variances are heterogeneous which violates a key assumption of

the t-test. The next section of the output provides the actual t-test results in two formats. The first format for "Equal" variances is the standard t-test taught in introductory statistics. This is the test result that should be reported in a research report under most circumstances. The second format reports a t-test for "Unequal" variances. This is an alternative way of computing the t-test that accounts for heterogeneous variances and provides an accurate result even when the homogeneity assumption has been violated (as indicated by the Levene test). It is rare that one needs to consider using the "Unequal" variances format because, under most circumstances, even when the homogeneity assumption is violated, the results are practically indistinguishable. When the "Equal" variances and "Unequal" variances formats lead to different conclusions, seek consultation. The output for both formats shows the degrees of freedom (df) and probability (2-tailed significance). As in all statistical tests, the basic criterion for statistical significance is a "2-tailed significance" less than .05. The .006 probability in this example is clearly less than .05 so the difference is statistically significant.

A second method of performing an independent groups t-test with SPSS is to use a noncategorical variable to divide the test variable (college GPA in this example) into groups. For example, the group of participants could be divided into two groups by placing those with a high number of study hours per week in one group and a low number of study hours in the second group. Note that this approach would begin with exactly the same information that was used in the correlation example. However, converting the Studyhrs data to a categorical variable would cause some detailed information to be lost. For this reason, caution (and consultation) is needed before using this method. To request the analysis, click **Statistics > Compare Means > Independent Samples T Test...**. Colgpa will remain the "Test Variable(s)" so it can be left where it is. Alternately, other variables can be moved into this box. Click "Sex(1,2)" to highlight it and remove it from the "Grouping Variable" box by clicking the bottom arrow which now faces left because a variable in the box has been highlighted. Next, highlight "Studyhrs" and move it into the "Grouping Variable" box. Now click **Define Groups...** and click the **Cut point** button. Enter a value (20 in this case) into the box. All participants with values less than the cutpoint will be in one group and participants with values greater than or equal to the cutpoint will form the other group. Click **Continue** > **OK** and the output will quickly appear. The results from the example are shown below:

Group Statistics

	Studyhours	N	Mean	Std.Deviation	Std. Error Mean
COLGPA College GPA for Fall 1997	Studyhours >= 20.00	6	3.4500	.4416	.1803
	Studyhours < 20.00	5	2.3400	.4930	.2205

The "Group Statistics" table provides the means and standard deviations along with precise information regarding the formation of the groups. This can be very useful as a check to ensure that the cutpoint was selected properly and resulted in reasonably similar sample sizes for both groups. The remainder of the output is virtually the same as the previous example.

Back to the Top of the Page

The t-test For Dependent Groups on SPSS

The t-test for **dependent** groups requires a different way of approaching the data. For this type of test, each case is assumed to have two measures of the same variable taken at different times. Each "Case" would therefore consist of a **single person**. This would be what is called a repeated measures design. Alternately, each case could contain the same information about two **different** individuals who have been paired or matched on a variable. In the repeated measures situation, one might collect GPA information at two different points in the careers of a group of students. The table below shows how this situation might appear in the fictional example. In this case, GPA data have been collected at the end of each participant's first year (Colgpa1) and senior year (Colgpa2).

Participant	Colgpa1	Weekly Study Time	Sex	Colgpa2
Participant #01	1.8	15 hrs	2	.
Participant #02	3.9	38 hrs	1	3.88
Participant #03	2.1	10 hrs	2	2.80
Participant #04	2.8	24 hrs	1	3.20
Participant #05	3.3	36 hrs	.	3.60
Participant #06	3.1	15 hrs	2	3.57
Participant #07	4.0	45 hrs	1	4.00
Participant #08	3.4	28 hrs	1	3.35
Participant #09	3.3	35 hrs	1	3.66
Participant #10	2.2	10 hrs	2	2.55
Participant #11	2.5	6 hrs	2	2.67

One thing to note about the new data is that the GPA of the first participant is missing. Given the 1.8 GPA at the first assessment, it seemed reasonable that this person might not remain in college for the entire four years. This is a common hazard of repeated measures designs and the implication of such missing data needs to be considered before interpreting the results.

To request the analysis, click **Statistics > Compare Means > Paired-Samples T Test ....** A window will appear with a list of variables on the left and a box labeled "Paired Variables" on the right. Highlight **two** variables (Colgpa and Colgpa2, in this example) and transfer them to the "Paired Variables" box by clicking the right-pointing arrow between the boxes. Several pairs of variables can be entered at this time. The **Options...** button opens a window that allows control of the confidence interval and missing data options. Click **Continue** (if you opened the **Options...** window) **> OK** to complete the analysis. The output will appear in an Output window. Results for the example problem are shown below:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Colgpa1	3.0600	10	.6552	.2072
	Colgpa2	3.3280	10	.5091	.1610

Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 Colgpa1 - Colgpa2	10	.944	.000

Paired Samples Test

	Paired Differences					t
				95% Confidence Interval of the Difference		
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	
Pair 1 Colgpa1 - Colgpa2	-.2680	.2419	7.649E-02	.4410	-9.50E-02	-3.504

Paired Samples Test

	df	Sig. (2-tailed)
Pair 1 Colgpa1 - Colgpa2	9	.007

The output is similar to the independent groups t-test. The first table of the output shows the means and standard deviations for the two groups and the second table shows the correlation between the paired variables. The next table shows the mean of the differences, standard deviation of the differences, standard error of the mean, the confidence interval for the difference, and the obtained value for t. The 2-tailed Sig[nificance] which is stated as a probability is shown in the last table. As usual, probabilities **less than** .05 indicate that the null hypothesis should be rejected. In this case, the interpretation would be that GPA increased significantly from firstyear to senior year, $t(9) = 3.50$, $p = .007$.

Back to the Top of the Page

Analysis of Variance with SPSS

The analysis of variance (ANOVA) is a flexible statistical procedure that can be used when the researcher wishes to compare differences among more than two means. Two different ANOVA models will be described in this handout: the simple one-way ANOVA and the two-way factorial ANOVA. The one-way ANOVA is analogous to the t-test except that more than two means can be tested for differences simultaneously. For example, to investigate GPA in college students, a researcher may wish to conduct a t-test between mean GPAs for first-year and senior students. However, why restrict the data to only two levels of class membership? It would make more sense to look at average GPAs for first-year, sophomore, junior, and senior students. Since more than two means are being tested, a one-way analysis of variance would be the appropriate test. The end result of an ANOVA is an *F*-ratio which can be interpreted in the same way as the *t*-ratio. However, a significant *F*-ratio only indicates that some difference exists among the tested means. In order to determine what mean, means, or combination of means differs, it is necessary to employ subsequent tests which can either be planned ahead of time (*a priori*) or after the results have been seen (*post hoc*). The main issue in selecting exactly which test to use is to prevent Type I errors that would result if a number of dependent tests were conducted without adjusting the alpha level. SPSS has several flexible options for selecting a subsequent test.

Back to the Top of the Page

The One-Way ANOVA with SPSS

The table below shows fictional data for a study of college GPA and class membership. Class refers to whether the individual reports being a first-year, sophomore, junior, or senior student. The analysis would begin by entering the data into the SPSS-Win spreadsheet as described previously. The variables can then be named and labeled as appropriate.

Participant	Current GPA	Class
Participant #01	1.9	1
Participant #02	3.8	4
Participant #03	2.2	2
Participant #04	2.8	3
Participant #05	3.3	4
Participant #06	3.1	1
Participant #07	4.0	3
Participant #08	3.5	2
Participant #09	3.4	3
Participant #10	2.1	2
Participant #11	2.6	1
Participant #12	3.6	4

Once the data have been entered, click **Statistics > Compare Means > One-way ANOVA....** A window will appear with a list of variables on the left and boxes labeled "Dependent List" and "Factor" on the right. Transfer the dependent variable (the variable for which means are to be computed) into the "Dependent List" box and the independent variable (the variable used as the grouping variable) into the "Factor" box. Next, click **Define Range...** and enter the minimum and maximum for the grouping variable. This may seem like an unnecessary step but it allows the researcher to exclude extreme values of the grouping variable. This could be desirable because too few participants were in one of the categories or a category had a label such as "Does Not Apply" or "Don't Know" in which the researcher was not interested. Once the range has been defined, click **Continue**. I also recommend clicking the **Options...** button and requesting descriptive statistics and a test for homogeneity of variance by clicking on the appropriate boxes. It may be interesting to take a look at what is available under the Contrasts... and Post Hoc... buttons which allow the researcher to select subsequent tests which may be needed to ascertain exactly where differences between and among the means may be found. Clicking **Cancel** takes one back to the main window without any subsequent tests being requested. Finally, click **OK** to cause the analysis to be computed.

Results are shown below. It begins with descriptive statistics and the results of the test for homogeneity of variance, ending with the familiar ANOVA summary table. The key information is the F-ratio and associated probability (F Prob.). In this example, the difference in GPAs among the four classes was not statistically significant as shown by the probability which

is considerably more than .05. The Levene test tests the assumption that the group variances are homogenous. When the results of this test are significant, that is, the "2-tail Sig." is less than .05, the assumption has been violated. Procedures for dealing with this situation are discussed in most advanced statistics books. The solution involves adjusting the degrees of freedom for finding the critical value of *F*.

Descriptives

					95 Pct Conf Int for Mean			
Group	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
firstyr	3	2.5333	.6028	.3480	1.0360	4.0307	1.90	3.10
soph	3	2.6000	.7810	.4509	.6598	4.5402	2.10	3.50
junior	3	3.4000	.6000	.3464	1.9095	4.8905	2.80	4.00
senior	3	3.5667	.2517	.1453	2.9415	4.1918	3.30	3.80
Total	12	3.0250	.6982	.2016	2.5814	3.4686	1.90	4.00

Test of Homogeneity of Variances

Levene Statistic	dfl	df2	Sig.
1.1910	3	8	.373

[This nonsignificant result is good because it shows that the homogeneity of variance assumption was not violated. A "Sig." value below .05 would be a cause for concern.]

ANOVA

COLGPA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.56692	3	.8564	2.4527	.138
Within Groups	2.7933	8	.3492		
Total	5.3625	11			

[The value for "Sig." is greater than .05, therefore the result is NOT significant.]

Back to the Top of the Page

Factorial ANOVA with SPSS

Factorial analysis of variance is an extension of the one-way analysis. The difference is that a factorial analysis has more than one independent (grouping) variable. For example, a study could be designed to simultaneously assess the relationship of sex (male vs. female) and class (firstyear, sophomore, junior, senior) to current college GPA. College GPA would be the dependent variable. Sex and class would be the two independent or grouping variables. Factorial ANOVA procedures can become very problematic due to the complexity of assumptions involved and the variety of methods available for computing the analysis. As a general rule, it is recommended that beginning researchers try to have the same number of participants in each of the cells of the design. When the cells have unequal numbers of participants, the danger of assumptions being violated becomes greatest and the available options for computing the analysis are the most variable. However, even when the cells contain unequal numbers of participants, the default options (those automatically available from SPSS) will provide good solutions.

As an example, consider the following data which represents the results of a fictional study of the relationship of sex (male vs. female) and class (firstyear, sophomore, junior, senior) to cumulative college GPA. Although these data are fictional, the means are representative of prior studies. The minimum information needed to analyze such data is three pieces of information about each participant, their GPA, sex, and class. In preparing the data for analysis by hand, it might appear as shown below. Note that each cell contains the data for two individuals which also makes the analysis very straightforward and easy to understand. This presentation makes it easy to see to which group each participant's GPA belongs, but it is not the way the data need to be placed into the SPSS data input window.

Class				
	First-year (1)	**Sophomore (2)**	**Junior (3)**	**Senior (4)**
Female (1)	3.2; 3.1	3.3; 3.4	3.2; 3.3	3.3; 3.2
Male (2)	2.8; 2.9	3.3; 3.0	3.1; 3.2	3.2; 3.1

In order to prepare these data for analysis with SPSS, they need to be arranged differently. The information about each participant (GPA, sex, class) needs to be arranged horizontally. The result will be twelve rows of data, one row for each participant. Reserving the first column for a participant identification number may also be helpful, especially when it is necessary to look up

data on the original data forms. The data below have been arranged in the same order in which they appear in the above table so it is easy to see how the coding was done. This is not necessary because SPSS (or any statistical package for that matter) will place each participant in the correct group based upon coding for the Sex and Class variables. For example, Participant #1 will be classified as a first-year female based upon the values for Sex (1) and Class (1). Similarly, Participant #4 will be classified as a sophomore female based upon the values for Sex (1) and Class (2). In more advanced analyses, each classification could be based upon several independent variables instead of only two.

To begin the analysis with SPSS, the data would be entered into a Newdata spreadsheet in the manner shown below. Adding variable names and variable labels will make the output much easier to interpret. These tasks can be accomplished using the procedures described at the beginning of the handout. The next task is to request that the analysis of variance be performed.

Participant #	GPA	Sex	Class
#1	3.2	1	1
#2	3.1	1	1
#3	3.3	1	2
#4	3.4	1	2
#5	3.2	1	3
#6	3.3	1	3
#7	3.3	1	4
#8	3.2	1	4
#9	2.8	2	1
#10	2.9	2	1
#11	3.3	2	2
#12	3.0	2	2
#13	3.1	2	3
#14	3.2	2	3
#15	3.2	2	4
#16	3.1	2	4

To do the analysis, select **Statistics > General Linear Model > GLM - General Factorial**. An input window will appear. Highlight the dependent variable (GPA, in this case) and transfer it to the "Dependent" box and highlight the independent or grouping variables and transfer them to the "Fixed Factor(s)" box. At this point the **Okay** button will be available. Clicking it will produce the desired analysis using the default settings that the SPSS program provides.

At this point, an interesting choice needs to be made. Clicking the **Options...** button reveals that three **Methods** (Type I, Type II, Type III, and Type IV) are available for calculating the sums of squares for the analysis of variance. The default option is Type III which should not be changed. The **Enter Covariates** and **Maximum Interactions** boxes can be ignored, although these are useful options for complex studies. The **Options...** window may also be used to request means and frequency counts. As you become more skilled in data analysis, the **Help** button or a click of the right mouse button can be used to obtain more information about a procedure and what it means. Annotated results for the above example are shown below.

Back to the Top of the Page

Univariate Analysis Of Variance

Between-Subjects Factors

		N
Sex	1.00	8
	2.00	8
Class	1.00	4
	2.00	4
	3.00	4
	4.00	4

The above table simply shows the number of individuals in each of the conditions.

Descriptive Statistics

SEX	CLASS	Mean	Std. Dev.	N
1.00	1.00	3.1500	7.071E-02	2
	2.00	3.3500	7.071E-02	2
	3.00	3.2500	7.071E-02	2
	4.00	3.2500	7.071E-02	2
	Total	3.2500	9.258E-02	8
2.00	1.00	2.8500	7.071E-02	2
	2.00	3.1500	.2121	2
	3.00	3.1500	7.071E-02	2
	4.00	3.1500	7.071E-02	2
	Total	3.0750	.1669	8
Total	1.00	3.0000	.1826	4
	2.00	3.2500	.1732	4
	3.00	3.2000	8.165E-02	4
	4.00	3.2000	8.165E-02	4
	Total	3.1625	.1586	16

Although this table may appear a bit complicated at first, it is really easy to understand. The two columns on the left indicate the condition or group for each row of data. For example, the first mean in the table (3.15) is the mean for first-year females because the data were coded by having a "1" for Sex indicate females and a "1" for Class indicate first-year students. Of course, the researcher must remain aware of how the data were coded in order to interpret the table unless variable labels are used. The second mean (3.35) is for female sophomores. The third mean (3.25) is for female juniors and the fourth mean (3.25) is for female seniors. The fifth or Total mean (3.25) is the mean for all females in the study. The same interpretation applies to the five following means except that they are for males. The designation "Total" in the column labeled "Sex" is the means for all individuals in each Class. For example, the mean for Sex =

"Total" and Class = "1.00" is the mean of all four first-year students. Finally, the mean for Sex = "Total" and Class = "Total" is the mean of all 16 individuals in the study.

The standard deviations are interpreted in the same way as the means. The unusual notation for some standard deviation values is standard scientific notation. The "E-02" that follows some values indicates that the decimal point should be shifted two places to the left to read the number. For example, the number 8.165E-02 stands for .08165.

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	DF	Mean Square	F	Sig of F	Eta Squared
Corrected Model	.298	7	4.250E-02	4.250	.030	.788
Intercept	160.063	1	160.023	16002.250	.000	1.000
SEX	.122	1	.122	12.250	.008	.605
CLASS	.147	3	4.917E-02	4.917	.032	.648
SEX * CLASS	2.750E-02	3	9.167E-02	.917	.475	.256
Error	8.000E-02	8	1.000E-02			
Total	160.400	16	.010			
Corrected Total	.377	15	.025			

The Source, Type III Sum of Squares, DF, Mean Square, F, and Sig[nificance] of F provide information that can be interpreted as described in your textbook. Eta Squared is a measure of the effect size or magnitude of the effect. It is a squared measure of association and has an interpretation similar to a squared correlation coefficient. It describes the degree of association between the independent and dependent variable.

When the number of individuals per cell or condition is equal (also called a "balanced" design), as in this example, the Type III Sum of Squares for SEX, CLASS, the SEX * CLASS interaction, and the Corrected Total will correspond to the "classic" computational method described in most introductory textbooks. When the cells or conditions contain different numbers of individuals, the Type III sum of squares will differ from the "classic" computations. However, these differences provide necessary adjustments that result from the unbalanced nature of the design. Another issue that results from an unbalanced design (unequal numbers of participants in the cells) is that even the various main effect (Total) or marginal means may be distorted unless adjustments are made. The adjusted marginal means may be requested by clicking the **Options** button and checking the appropriate boxes. These issues will be covered in advanced statistics courses.

Back to the Top of the Page

Chi-Square Test

The chi-square goodness of fit test and test for independence are both available on SPSS. Recall that chi-square is useful for analyzing whether a frequency distribution for a categorical or nominal variable is consistent with expectations (a goodness of fit test), or whether two categorical or nominal variables are related or associated with each other (a test for independence). Categorical or nominal variables assign values by virtue of being a member of a category. Sex is a nominal variable. It can take on two values, male and female, which are usually coded numerically as 1 or 2. These numerical codes do not give any information about how much of some characteristic the individual possesses. Instead, the numbers merely provide information about the category to which the individual belongs. Other examples of nominal or categorical variables include hair color, race, diagnosis (e.g., ADHD vs. anxiety vs. depression vs. chemically dependent), and type of treatment (e.g., medication vs. behavior management vs. none). Note that these are the same type of variables that can be used as independent variables in a t-test or ANOVA. In the latter analyses, the researcher is interested in the means of another variable measured on a interval or ratio scale. In chi-square, the interest is in the frequency with which individuals fall in the category or combination of categories.

Back to the Top of the Page

Chi-Square Test for Goodness of Fit

A chi-square test for goodness of fit can be requested by clicking **Statistics > Nonparametric Tests > Chi-square**. This opens up a window very similar to other tests. Enter the variable to be tested into the **Test Variable** box. Then a decision about the expected values against which the actual frequencies are to be tested needs to be made. The most common choice is "All categories equal." However, it is also possible to enter specific expected values by checking the other circle and entering expected values in order. The expected values used in computing the chi-square will be proportional to these values. The **Options...** button provides access to missing value options and descriptive statistics for each variable. To submit the analysis click the **OK** button. Results for a goodness of fit chi-square are shown below.

NPar Tests
Chi-Square Test
Frequencies

Class

	Observed N	Expected N	Residual
first-year	3	3.0	.0
Sophomore	3	3.0	.0
Junior	3	3.0	.0
Senior	3	3.0	.0
Total	12		

[The data were taken from the previous ANOVA example.]

[The "residual" is just the difference between the observed and expected frequency.]

[Warning: Using the Chi-Square statistic is questionable here because all four cells have expected frequencies less than 5. See your statistics textbook for advice if you are in this situation.]

	Class
Chi-Square	.000
df	3
Asymp. Sig.	1.000

a. 4 cells (100.0%) have expected frequencies less than 5. The minimum expected cell frequency is 3.0.

The value under "Chi-Square" is .0000 because the cell frequencies were all equal. As usual, statistically significance results are indicated by "Asymp. Sig.[nificance]" values below .05. Obviously, this example is **NOT** statistically significant. In words, these results indicate that the obtained frequencies do not differ significantly from those that would be expected if all cell frequencies were equal in the population.

Back to the Top of the Page

Chi-Square Test for Independence

The chi-square test for independence is a test of whether **two** categorical variables are associated with each other. For example, imagine that a survey of approximately 200 individuals has been conducted and that 120 of these people are females and 80 are males. Now, assume that the survey includes information about each individual's major in college. To keep the example simple, assume that each person is either a psychology or a biology major. It might be asked whether males and females tend to choose these two majors at about the same rate or does one of the majors have a different proportion of one sex than the other major. The table below shows the case where males and females tend to be about equally represented in the two majors. In this case college major is **independent** of sex. Note that the percentage of females in psychology and biology is 59.8 and 60.2, respectively. Another way to characterize these data is to say that sex and major are independent of each other because the proportion of males and females remains the same for both majors.

	Psychology Majors	**Biology Majors**
Females	58	62
Males	39	41

The next example shows the same problem with a different result. In this example, the proportion of males and females **depends** upon the major. Females compose 79.6 percent of psychology majors and only 39.2 percent of biology majors. Clearly, the proportion of each sex is different for each major. Another way to state this is to say that choice of major is strongly related to sex, assuming that the example represents a statistically significant finding. It is possible to represent the strength of this relationship with a coefficient of association such as the contingency coefficient or Phi. These coefficients are similar to the Pearson correlation and interpreted in roughly the same way.

	Psychology Majors	**Biology Majors**
Females	82	38
Males	21	59

The method for obtaining a chi-square test for independence is a little tricky. Begin by clicking **Statistics > Summarize > Crosstabs...**. Transfer the variables to be analyzed to the **Row(s)** and **Column(s)** boxes. Then go to the **Statistics...** button and check the Chi-square box and anything that looks interesting in the **Nominal Data** box, followed by the **Continue** button. Next, click the **Cells...** button and check any needed descriptive information. **Percentages** are particularly useful for interpreting the data. Finally, click **OK** and the output will quickly appear.

Sample results are shown below. These data are from the ANOVA example so the number of observations in each cell is only two. This is a problematic situation for chi-square analysis and,

should this be encountered in an actual analysis, consulting a textbook is recommended. Furthermore, the results are far from significant because the distribution of sex across class remains constant.

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEX *	16	100.0%	0	.0%	16	100.0%

The "Case Processing Summary" provides some basic information about the analysis. In studies with large numbers of participants, this information can be very useful.

Sex * Class Crosstabulation

			Class				Total
			1.00	2.00	3.00	4.00	
Sex	1.00	Count	2	2	2	2	8
		% within Sex	25.0%	25.0%	25.0%	25.0%	100.0%
		% within Class	50.0%	50.0%	50.0%	50.0%	50.0%
		% of Total	12.5%	12.5%	12.5%	12.5%	12.5%
	2.00	Count	2	2	2	2	8
		% within Sex	25.0%	25.0%	25.0%	25.0%	100.0%
		% within Class	50.0%	50.0%	50.0%	50.0%	50.0%
		% of Total	12.5%	12.5%	12.5%	12.5%	50.0%
Total		Count	4	4	4	4	16
		% within Sex	25.0%	25.0%	25.0%	25.0%	100.0%
		% within Class	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	25.0%	25.0%	25.0%	25.0%	100.0%

Note: The above results can be obtained by requesting all the available percentages in the cross-tabulation. In this simple example, the percentages are not very useful. However, when large numbers of participants are in the design, the percentages help greatly in understanding the pattern of the results. Also, when the analysis is presented in a research report, the percentages within one of the variables will help the reader interpret the results.

Chi-Square Tests

Chi-Square	Value	df	Asymp. Sig (2-sided)
Pearson [Standard computation]	.00000	3	1.000
Likelihood Ratio	.00000	3	1.000
Linear-by-Linear Association	.00000	1	1.000
N of Valid Cases	16		

a. 8 cells (100.0% have expected count less than 5. The minimum expected count is 2.00.

The values for "Sig" are probabilities. A statistically significant result has a probability of less than .05.

Back to the Top of the Page

Other Helpful Features of SPSS

There are a number of additional features available in SPSS that can be extremely helpful for the beginning researcher. These features will be described briefly.

Transformations

Two particularly valuable features are available from the **Transformations** menu: **Recode**, and **Compute**. The purpose of a recode is very simple. Imagine a variable that is coded from 1 to 5. Sometimes extreme values are not selected by very many individuals. Thus, it may be desirable to combine individuals who responded with either a 4 or a 5 into a single category such as 4. The recode feature is the way to do this. Another situation that often calls for a recode is when a variable is part of a scale but the scoring needs to be reversed before it can be added to other items to make a total score. This is also accomplished with the recode command. To do a recode, click **Transformations** > **Recode** > **Into Same Variables...** or **Into Different Variables...** and enter the required information. The choice of recoding into the Same or Different Variables is a question of whether it is desirable to preserve the old data. By doing the recode into a Different variable, the old data can be preserved in case a mistake is made or another recoding procedure is tried.

The **Compute...** command is also under the Transformations menu. This command allows the researcher to construct an equation for changing the scale of a variable. The main usefulness of the procedure is for remedying the situation where the raw data do not meet statistical assumptions. A transformation using the **Compute...** command can often bring the data back into conformity with statistical assumptions. Most statistics books have a discussion of the various common types of transformations and their potential benefits.

Exploratory Data Analysis

Exploratory data analysis is a process of carefully examining data prior to performing inferential statistical tests. Access to exploratory data analysis techniques can be obtained by clicking **Statistics > Summarize > Explore..**. which leads to plots (boxplots; stem and leaf) and descriptive statistics that can help greatly in the early stages of data analysis. Distributions can also be tested for normality.

Help Features

The Help menu provides access to information about specific **Topics**, a **Tutorial**, a **Statistics Coach**, and other useful features. It is also possible to click the right mouse button while pointing to a term of interest which will result in a display of the definition of that term. The dialog or input boxes also have context-specific **Help** buttons.

Reliability Analysis

Reliability is one of the most important characteristics of good psychological measures. To compute the standard measure of internal consistency, coefficient alpha, click **Statistics > Scale > Reliability Analysis...**. The variables that make up the scale to be analyzed should be transferred to the **Items** box. Then click the **Statistics...** button and request all the descriptive statistics plus the inter-item correlations.

Moving Output to Other Applications

Often, it will be desirable to move output to another application such as a word-processing file. This operation will prove especially useful in research methods courses. To do this, copy the table, chart, or plot that you wish to move using the **Edit** menu. Then open up the target application (for example a word processing file into which you would like to copy the item) and select **Paste Special...** from the **Edit** menu. From the list of options, choose **Picture**. This is the simplest method of including SPSS output in another file. There are other methods which enable you to update the table or chart with SPSS. You can learn more about these processes by searching the Help files in SPSS. Transferring information from one program to another is often one of the most difficult tasks to accomplish with modern technology so it is wise to seek consultation when difficulties are encountered.

Conclusion

These directions are meant describe some basic analyses and point the way toward more advanced procedures. As you become more skilled you will find that very complex analyses can be easily performed. The best way to learn the advanced features of SPSS for Windows is to explore the program using data from an original study. At this stage it will often be necessary to learn by trial-and-error. Learning may be slow and frustrating but keep in mind that SPSS is much faster, convenient, and accurate than computing the analysis by hand. Allow a lot of time for exploring your options and interpreting the output at each stage. Although one correlation can be computed in a few seconds, analysis of an entire study may take weeks or even months. I suggest that every data analysis begin with the "Frequencies" procedure which allows easy identification of outliers and out-of-range data (e.g., a value that is beyond what is allowable). The Frequencies output also allows identification of imbalances in the sample such as too few males or too many first-year students. Once the frequency distributions have been examined, you must return to the original questions that inspired the research to develop your ideas about appropriate analyses

Back to the Top of the Page