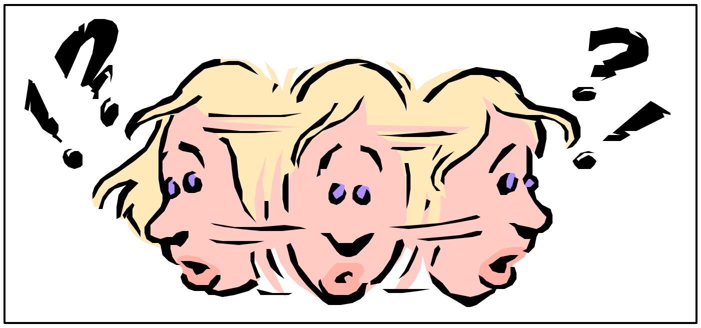
The Dummy's Guide to Data Analysis Using SPSS



Univariate Statistics

Scripps College

Amy Gamble

April, 2001

Table of Contents

PAGE

Creating a Data File

Creating New Variables

- First click on the **Variable View** tab on the lower left-hand corner of your screen. This is where you enter in your new variables.
- Under **Name** enter in the name of your variable. The name may need to be abbreviated since SPSS only allows 8 letters (e.g., "type of schooling" may need to be entered as "school"). It is usually best to start with an ID or name variable so that you can keep track of your cases.
- **For the following columns, clicking on the box will cause a small gray box to appear on the right-hand side. You will need to click on the gray box to change the settings. After entering in the name and clicking on tab, the default settings for the other columns will pop up.**
- Under **Type** be sure that **Numeric** is selected if your data is numeric or ordinal, or if you will be using codes (e.g. 1 = male, 2 = female). If you want to be able to type in letters instead of numbers, you need to click on **String**, then click on **Ok**.
- Under **Width** be sure the column will be wide enough to enter in the given values (the default setting usually is wide enough so you can ignore this).
- Under **Decimals** be sure there is the right number of places past the decimal. If you want things to show up as whole numbers, change the **Decimals** to zero, if you want it to the one's digit, set it to one, etc.
- Under **Labels** enter in the entire name of the variable so that you know what it is. (e.g. type in "Type of schooling" instead of just "school").
- **Values** is where you enter in your codes. Click on the gray box. Next to **Value** type in the numerical code (e.g. 1). Next to **Value Label** type in what that number represents (e.g. Male). Then click **Add**. Once you have entered in all of the codes, click on **Ok**.
- You don't really need to do anything with the next three columns (**Missing, Columns, and Align).**
- For **Measure** make sure the correct level of measurement is selected (usually the default setting will make it correct).

Entering the Data

- When all of the variables have been entered, click on the **Data View** tab in the lower left-hand corner of your screen.
- Your new variables will appear horizontally, across the top of the data sheet instead of vertically.
- Enter in the numerical values.

Determining Which Test to Use

Step 1: Which type of data do you have?

- When you are determining which type of data you have, remember that you are looking at the **Dependent Variable**, or in other words, the variable that measures the difference (or the relationship for correlations). For example, if you are looking at the difference between men and women in how well they scored on the SAT's, their score on the SAT is the dependent variable because it is the one that you are measuring in order to determine if there is a difference.

- Nominal Data: Numbers or words that are used merely as labels. Some examples of nominal data are types of religion (Christian, Catholic, Jewish, etc.), political affiliation (Republican, Democrat, etc.), and ethnicity (Hawaiian, Japanese, etc.). Note that these cannot be ranked in any logical order.

- Ordinal Data: Numbers (or words, but usually numbers) that can be rank ordered or scaled. Some examples of ordinal data are the place finished in a race (1st, 2nd, 3rd, etc.), degrees of happiness (sad, neutral, happy), and level of school (elementary, middle school, high school, college, etc.). Note however, that these do not express a degree of magnitude or interval. For example, you could not say that someone who felt neutral had twice as much "happiness" than a person who was sad. You could not say that the amount of time between the first place and second place runner was the same as the amount of time between the second and third place runner.

- Numeric Data: Numbers that can be rank ordered or scaled, that do express a degree of magnitude, that have consistent intervals. Some examples of numeric data are test scores, time, and height. Note that with the numeric data of time, 40 minutes is twice as much as 20 minutes, and that the amount of time between 20 to 40 minutes is the same amount of time as the amount of time between 40 and 60 minutes.

Step 2: What are you looking for? (This is pretty obvious!)

- Differences: If you are trying to determine if one group is different, greater than, or less than another group. For example, are men taller than women?

- Correlations: If you are trying to determine if there is a relationship between one variable and another. For example, does alcohol consumption increase as unemployment rates increase? If you are looking for a correlation, you do not need to complete steps 3 and 4.

- Regressions: If you are trying to determine if one variable can accurately predict another variable. For example, can the amount of rainfall predict the amount of mud slides? If you are looking for a regression, the variables must be numeric, and you do not need to complete steps 3 and 4.

Step 3: Between Groups or Within Groups? (aka Independent Groups or Correlated Groups)

- If you have a one group test (see step 4 if you do not know what that is), you do not need to complete step 3.

- Between (aka Independent) Groups: When each participant is in only one of the groups. For example, when comparing men and women, a man would not be in both groups. Thus, you are testing two (or more) different groups of people.

- Within (aka Correlated) Groups: When each participant is in all of the groups. For example, giving exams to a group of people on three separate occasions and comparing their scores on the three different exams (here the three "groups" would be the three test scores). Thus, you are comparing the same group of people at three different times.

Step 4: One Group, Two Groups, or More than Two Groups?

- Remember that you are looking at the **Independent Variable** when you are determining how many groups your variable has.

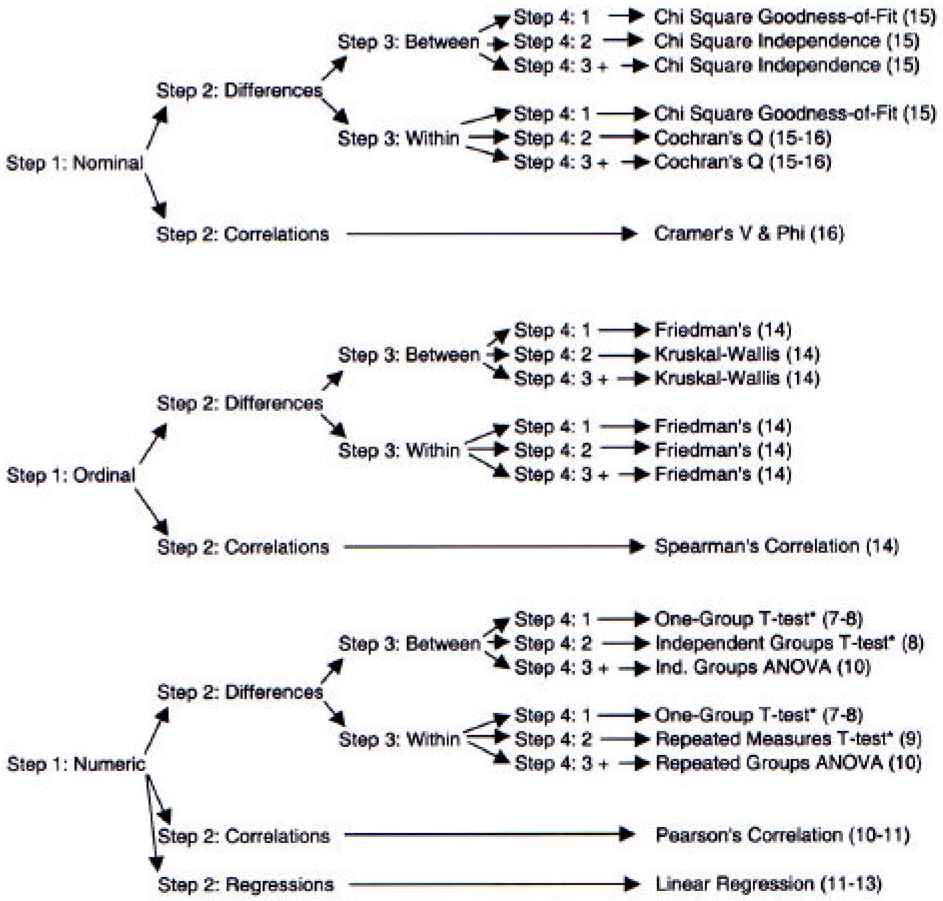
- One Group: If you are comparing the distribution of one group of people to a hypothetical or actual population distribution. For example, comparing the ethnic distribution of the Claremont Colleges with the ethnic distribution of the USA.

- Two Groups: For between (aka independent) groups: If you are comparing one group of people to another. For example, comparing men and women. For within (aka correlated) groups: If you are comparing the same group of people's test scores on two separate occasions.

- More than Two Groups: For between (aka independent) groups: If you are comparing more than two groups. For example, comparing students from Scripps, CMC, HMC, Pitzer, and Pomona. For within (aka correlated) groups: If you are comparing the same group of people's test scores on more than two occasions.

Step 5: Selecting the Test

- See next page



* If the population mean and the population standard deviation are known, then you would use a z-score; however, it is highly unlikely that you will know the population mean and standard deviation (unless you are testing variables that are asked about in the census), thus the t-test would be the most appropriate test to use.

Hints For All Tests

- Remember that the Significance (or Asymp. Sig. in some cases) needs to be **less than** 0.05 to be significant.
- The Independent Variable is always the variable that you are predicting something about (i.e. what your Ha predicts differences between, as long as your Ha is correct). The Dependent Variable is what you are measuring in order to tell if the groups (or conditions for repeated measures tests) are different. For correlations and for Chi-Square, it does not matter which one is the Independent or Dependent variable.
- H_a always predicts a difference (for correlations, it predicts that r is different from zero, but another way of saying this is that there is a significant correlation) and H_o always predicts no difference. If your H_a was directional, and you find that it was predicted in the wrong direction (i.e. you predicted A was greater than B and it turns out that B is significantly greater than A) you should still accept H_o, even though H_o predicts no difference, and you found a difference in the opposite direction.
- If there is a WARNING box on your Output File, it is usually because you used the wrong test, or the wrong variables. Go back and double check.

Tests For Numeric Data

Z-Scores (Compared to Data)

Analyze → Descriptive Statistics → Descriptives

- Click over the variable you would like z-scores for
- Click on the box that says **Save Standardized Values as Variables**. This is located right below the box that displays all of the variables.
- If means and standard deviations are needed, click on **Options** and click on the boxes that will give you the means and standard deviations.
- The z-scores will not be on the Output File!!!
- They are saved as variables on the Data File. They should be saved in the variable that is to the far right of the data screen. Normally it is called z, and then the name of the variable (e.g. ZSLEEP)
- Compare the z-scores to the critical value to determine which z-scores are significant. Remember, if your hypothesis is directional (i.e. one-tailed), the critical value is + or – 1.645. If your hypothesis is non-directional (i.e. two-tailed), the critical value is + or – 1.96.

Z-Scores Compared to a Population Mean and Standard Deviation:

- The methodology is the same except you need to tell SPSS what the population mean and standard deviation is (In the previous test, SPSS calculated it for you from the data it was given. Since SPSS cannot calculate the population mean and standard deviation from the class data, you need to plug these numbers into a formula).
- Remember the formula for a z-score is:

$$z = \frac{\overline{X} - m}{s}$$

- You are going to transform the data you got into a z-score that is compared to the population by telling SPSS to minus the population mean from each piece of data, and then dividing that number by the population standard deviation. To do so, go to the DATA screen, then:

Transform → Compute

  - Name the new variable you are creating in the **Target Variable** box (ZUSPOP is a good one if you can't think of anything).
  - Click the variable you want z-scores for into the **Numeric Expression** box. Now type in the z-score formula so that SPSS will transform the data to a US population z-score. For example, if I am working with a variable called Sleep, and I am told the US population mean is 8.25 and that the US population standard deviation is .50, then my **Numeric Expression** box should look like this:

(SLEEP – 8.25)/.50

  - Compare for significance in the same way as above.

For All T-Tests

- ***The significance that is given in the Output File is a two-tailed significance. Remember to divide the significance by 2 if you only have a one-tailed test!***

For One Group T-Tests

Analyze → Compare Means → One-Sample T Test

- The Dependent variable goes into the **Test Variables** box.
- The hypothetical mean or population mean goes into the **Test Value** box. **Be Careful!!!** The test value should be written in the same way the data was entered for the dependent variable. For example, my dependent variable is "Percent Correct on a Test" and my population mean is 78%. If the data for

the “Percent Correct on a Test” variable were entered as 0.80, 0.75, etc., then the vest value should be entered as 0.78. If the data were entered as 90, 75, etc., then the test value should be entered as 78. In order to know how the data were entered, click on the Data File screen and look at the data for the dependent variable.

For Independent Groups T-Tests

Analyze → Compare Means → Independent-Samples T Test

- The Dependent Variable goes in the **Test Variable** box and the Independent variable goes in the **Grouping Variable** box.
- Click on **Define Groups** and define them. In order to know how to define them, click on **Utilities → Variables**. Click on the independent variable you are defining and see what numbers are under the value labels (i.e. usually it’s either 0 and 1 or 1 and 2). If there are more than two numbers in the value labels then you cannot do a t-test unless you are using a specified cut-point (i.e. if there are four groups: 1 = old women, 2 = young women, 3 = old men, 4 = young men, and you simply wanted to look at the differences between men and women, you could set a cut point at 2). If there are no numbers, you should be using a specified cut-point. If you have more than two numbers in the value labels, or if you have no numbers in the value labels, and no cut-point has been specified on the final exam you are doing the wrong kind of test!!!!
- On the Output File: Remember, this is a t-test, so ignore the F value and the first significance value (Levene’s Test). Also, ignore the **equal variances not assumed** row.
- Before accepting Ha, be sure to look at the means!!! If I predict that boys had a higher average correct on a test than girls and my t-test value is significant I may say, yes, boys got more correct than girls. However, this t-test could be significant because girls got significantly more correct than boys!! (Therefore, Ha was predicted in the wrong direction!) In order to know which group got significantly more correct than the other, I need to look at the means and see which one is bigger!!

For Repeated Measures (a.k.a Correlated Groups, Paired Samples) T- Test

Analyze → Compare Means → Paired-Samples T Test

- Click on one variable and then click on a second variable and then click on the arrow that moves the pair of variables into the **Paired Variables** box. *[In order to make things easier on yourself, click your variables in, in order of your hypotheses. For example, if you are predicting that A is greater than B, click on A first. This way, you should expect that your t value will be positive. If the test is significant, but your t value is negative, it means that B was significantly greater than A!! (So H_a was predicted in the wrong direction and you should accept H_o). If you are predicting that A is less than B, then click on A first. This way, you* *should expect that the t value will be negative. If the test is significant but the t value is positive, you know that it means that B was significantly less than A (so H_a was predicted in the wrong direction, and you should accept H_o).]*

For Independent Groups ANOVA

Analyze → Compare Means → One-Way ANOVA

- Put the Dependent variable into the **Dependent List** box. Put the Independent variable into the **Factor** box.
- Click on the **Post Hoc** box. Click on the **Tukey** box. Click **Continue**.
- If means and standard deviations are needed, click on the **Options** box. Then click on **Descriptive**.

For Repeated Measures (a.k.a. Correlated Groups, Paired Samples) ANOVA

Analyze → General Linear Model → Repeated Measures

- Type in the within-subject factor name in the **Within-Subject Factor Name** box. This cannot be the name of a pre-existing variable. You will have to make one up.
- Type in the number of levels in the **Number of levels** box. How do you know how many levels there are? If my within-subject factor was “Tests” and I have a variable called “Test 1” a variable called “Test 2” and a variable called “Test 3,” then I would have 3 levels. In other words, the number of levels equals the number of variables (that you are examining) that correspond to the within-subject factor.
- Click on **Define**.
- Put the variables you want to test into the **Variables** box. Preferably, put them in the right order (if there is an order to them). This will keep you from getting confused. For example, I should put in my “Test 1” variable in first, my “Test 2” variable in second, etc.
- For post hoc tests, click on Options, highlight the variable, move it into **Display Means For** box, click on **Compare Main Effects**, change **Confidence Interval Adjustment** to **Bonferonni** (the closest we can get to Tukey’s Test). You may also want to click on **Estimates of Effect Size** for eta^2.
- Remember to look at the **Tests of Within-Subjects Effects** box for your ANOVA results.

Pearson’s Correlation Coefficient (r)

Analyze → Correlate → Bivariate

- Make sure that the **Pearson** box (and only the **Pearson** box) has a check in it.
- Put the variables in the **Variables** box. Their order is not important.

- Select your tailed significance (one or two-tailed) depending on your hypotheses. Remember directional hypotheses are one-tailed and non-directional hypotheses are two-tailed.
- If means and standard deviations are needed, click on **Options** and then click on **Means and Standard Deviations**.

Linear Regression

Graphs → Scatter → Simple

- This will let you know if there is a linear relationship or not.
- Click on **Define**.
- The Dependent variable (criterion) should be put in the **Y-axis** box and the Independent variable (predictor) should be put in the **X-axis** box and hit **OK**.
- In your Output: Double click on the scatterplot. Go to **Chart → Options**. Click on **Total** in the **Fit Line** box, then **OK**.
- Make sure it is more-or-less linear. The next step is to check for normality.

Graphs → Q-Q

- Put both variables into the **Variable** box. Hit **OK**.
- Look at the **Normal Q-Q Plot of ….**, not the **Detrended Normal Q-Q of …** box. If the points are a "smiley face" they are negatively-skewed. This means you will have to raise them to a power greater than one. If the points make a "frown face" then they are positively-skewed. This means you will have to raise them to a power less than one (**but greater than zero**). To do this, go to the DATA screen then:

Transform → Compute

- Give the new variable a name in the **Target Variable** box. Since you will be doing many of these (because it is a guess and check) it may be easiest to name it the old variable and then the power you raised it to. For example, SLEEP.2 if I raised it to the .2 power, or SLEEP3 if I raised it to the third power.
- Click the old variable (that you want to change) into the **Numeric Expressions** box. Type in the exponent function (**), and then the power you want to raise it to. Hit **OK**.
- Redo the Q-Q plot with the NEW variable (i.e. SLEEP.2, and not SLEEP).
- Repeat until you have the best fit data. The variable that you created with the best-fit data (i.e. SLEEP.2) will be the variable that you will use for the REST OF THE REGRESSION (no more SLEEP).

- The next step is to remove Outliers. To do so, run a regression:

Analyze → Regression → Linear

- The Independent variable is the predictor (the variable from which you want to predict). The Dependent variable is the Criterion (the variable you want to predict).
- Click on **Save** and then click on **Cook's Distance**. Like the z-scores, these values will NOT be in your Output file. They will be on your Data file, saved as variables (**COO-1**).
- Do a Boxplot:

Graphs → Boxplot

  - Click on **Simple** and **Summaries of Separate Variables**.
  - Put Cook's Distance (COO_1) into the **Boxes Represent** box. Put a variable that will label the cases (usually ID or name) into the **Label Cases by** box.
  - Double click on the Boxplot in the Output file in order to enlarge it and see which cases need to be removed (those lying past the black lines). Keep track of those cases. (Some people prefer to take out only extreme outliers, those marked by a *.)
  - Go back to the Data file. Find a case you need to erase. If you highlight the number to the far left (in the gray), it will highlight all of the data from that case. Then go to **Edit → Clear**. Repeat this for all of the outliers. ***ERASE FROM THE BOTTOM OF THE FILE, WORKING UP; IF YOU DON'T THE ID NUMBERS WILL CHANGE AND YOU'LL ERASE THE WRONG ONES!*** DO NOT RE-RUN THE BOXPLOT LOOKING FOR MORE OUTLIERS! Once you have cleared the outliers from the first (and hopefully only) boxplot, you can continue.
  - Now re-run the regression.
  - Remember: R tells you the correlation. R^2 tells you the proportion of variance in the criterion variable (Dependent variable) that is explained by the predictor variable (Independent variable). The F tells you the significance of the correlation. The predicted equation is: (B constant) + (Variable constant)(Variable). Where B constant is the number in the column labeled B, in the row labeled (constant), and where Variable constant is the number in the column labeled B and in the row that has the same name as the variable [under the (constant) row].

Tests For Ordinal Data

- Remember, since this is ordinal data, you should not be predicting anything about means in your Ha and Ho. Also, you should not be reporting any means or standard deviations in your results paragraphs.
- Therefore, if you need to report medians and/or ranges go to **Analyze → Descriptive Statistics → Frequencies**. Click on **Statistics** and then click on the boxes for **median** and for **range**.

Kruskal-Wallis

Analyze → Nonparametric Tests → K Independent Samples

- Make sure the Kruskal-Wallis H box (and this box only) has a check mark in it.
- Put the Dependent variable in the **Test Variable List** box and put the Independent variable in the **Grouping Variable** box.
- Click on **Define Range**. Type in the min and max values (if you do not know what they are you will have to go back to **Utilities → Variables** to find out and then come back to this screen to add them in).

Friedman's Test

Analyze → Nonparametric Tests → K Related Samples

- Make sure that the **Friedman** box (and only that box) has a check mark in it.
- Put the variables into the **Test Variables** box.
- In the Output File: Even though it says Chi-Square, don't worry, you did a Friedman's test (as long as you had it clicked on).

Spearman's Correlation (r_s)

Analyze → Correlate → Bivariate

- Click on the **Pearson** box in order to turn it OFF! Click on the **Spearman** box in order to turn it ON!
- Choose a one or two-tailed significance depending on your hypotheses.
- Put variables into the **Variable** box.

Tests for Nominal Data

- Remember, since this is for nominal data, your hypotheses should not be predicting anything about means. In addition, you should not be reporting any means, standard deviations, medians, or ranges in your results paragraphs.
- In order to determine the mode go to **Analyze → Descriptive Statistics → Frequencies**. Click on the **Statistics** box and then click on the **mode** box.

Chi-Square Goodness-of-Fit

Analyze → Nonparametric Tests → Chi Square

- Put the variable into the **Test Variable List**.
- If all of the levels in your variable are predicted to be equal, click on **All Categories Equal**.
- If there is a predicted proportion that is not all equal (i.e. 45% of the US population is male and 55% is female) then you need to click on the **Values** box. In order to know which value you enter in first (i.e. 45% or 55%) you need to look at which one (male or female) was coded first. To do this, go to **Utilities → Variables** and click on the variable you are interested in. The one with the smallest number (i.e., if male was coded as 1 and female was coded as 2, males would have the smaller number) is the one whose predicted percentage gets entered first. In this case, since male was coded as one, the first value I would enter into the **Values** box would be 45%. The second would be 55%.

Chi-Square Independence

Analyze → Descriptive Statistics → Crosstabs

- Put one variable into the **Rows** box and one into the **Columns** box. It does not matter which variable is put into which box.
- Click on **Statistics**. Click on **Chi-Square**.
- In the Output File: Look at one of the two boxes that compare variable A to variable B. Ignore the box that compares variable A to variable A, and the box that compares variable B to variable B (you will know which ones these are because they have a chi-square value of 1.00 – a perfect correlation because it correlated with itself).

Cochran's Q

Analyze → Nonparametric Tests → K Related Samples

- Click on **Friedman** to turn it OFF! Click on **Cochran's Q** to turn it ON!
- If there are more than two groups, and you need to tell which groups are significantly different from each other, the only way you can do this is by doing

many Cochran’s Q tests using only two groups. Therefore, you have to test every combination. For example, if there are three groups, you have to do one Cochran’s Q test between group 1 and 2, one test between group 1 and 3, and one test between group 2 and 3 (so let’s hope he doesn’t ask you to do that!!).

Cramer’s V or Phi (Correlation for Nominal Data)

- Do a Chi-Square Independence except you need also to click on **Statistics** and then click on the **Phi and Cramer’s V** box, and the **Contingency Coefficient** box.