Center for Teaching, Research and Learning
Research Support Group
American University, Washington, D.C.
Hurst Hall 203
rsg@american.edu
(202) 885-3862

Introduction to SPSS

This workshop is designed to introduce you to SPSS and provide tips and tools you can use to get started with statistical analysis. When we are finished you will:

- Understand the components of the SPSS interface
- Be able to enter, load, and describe data
- Be able to use the menus to analyze and graph data
- Generate new and recode existing variables
- Generate Descriptive Statistics
- Run Correlations and Linear Regressions
- Create Histograms, Bar Charts, and Scatter Plots

SCENARIO: We are interested in studying the relationship between the amount of corruption in a country and the quality of their economy.

DATA: For this question, we are lucky as data already exists. We will use the "World Development Indicators" which are compiled annually by The World Bank. We are using data from a past year, and we have simplified the data set, but you can find the full data set here: http://data.worldbank.org/data-catalog/world-development-indicators. We are also lucky, because the World Bank saves there data in a way that make it easy to open, often when you find a data set (for example, old Census data), it can take a lot of effort just to transform the data into a usable form. And, of course, there is never a guarantee that someone else has created the data set you want in advance; often the first step in a research project is spending a long time looking up values to create the data set you need.

Within the WDI data set, we will focus on the variables "corrup" and "gdppc." The first is an index of *perceived* corruption within a country's government. Citizens are surveyed, and asked questions like "Would you need to plan bribes to start a local business?" Note that corrupt is "reverse coded" so that high scores equal low corruption. The second variable is Gross Domestic Product per Capita, which indicates how well the economy is doing, scaled for the population size of a country (for example, it is not very interesting to note that Botswana produces less than Brazil overall, but it might be interesting to note that Botswana produced more *per citizen* than Brazil in 2014).

Note that, in using these variables from the WDI data set, we have given over the responsibility of determining what "corruption" and "quality of economy" means to the World Bank. We have also allowed

dynamic concepts to become simple numbers. However, this allows us to move quickly to the later stages of analysis, which has its advantages.

I. SPSS OVERVIEW

Before we begin, it is important to become familiar with SPSS, including becoming familiar with the menus, and the Data View, Variable View, and Output Window.

II. MANAGING DATA AND DESCRIPTIVE STATISTICS

The WDI data comes as an Excel file. After we look briefly at the file in Excel, we will import the data into SPSS and save it as an SPSS(.sav) file.

Importing Data

- **File, Open, Data**.
- Change Files/Type to —Excel.
- Find and select the Excel file. **J:\CLASSES\Workshops\cs.xlsx**
- The first row includes variable names, so select —read variable names from the first row

Saving Data

- **File, Save As**
- Choose save location, name data, choose .sav format
- Click Save [Note, you will need to save to a location to which you have permission to save]

Data Labels

- **Click the tab at the bottom left to go into "Variable View"**
- Leave the Names as they are, but change the Label of "gdppc" to "Gross Domestic Product Per Capita" and the label of "corrup" to "Corruption Perception Index"

Perform Math on a Continuous Variables

- **Transform, Compute Variable**.
- In the "Numeric Expression:" panel, write "10 – "
- Select the variable (**corrupt)** and arrow it over to the "Numeric Expression:" panel.
- In the "Target Variable:" panel write "Corruption2" (This will become the name of our new variable.)
- **OK**.

Recode Continuous Variables into Discrete Variable

- **Transform, Recode Into Different Variable**.
- Select this variable (**Corruption2**), Output Variable Name (**Corrupt_Categ**),
- Old and New Values,
- Range, Less than 3.3, Value = 1, **Add**; AND
- Range 3.31 through 6.6, Value = 2, **Add**; AND
- Range 6.61 to 10, Value = 3,
- **Add**; **Continue. Change. OK**.

In the "Variable View" window, add a Variable Label and Value Labels to this **Corrupt_Categ**.

III. ASKING QUESITONS ABOUT THE MEAN OF A VARIABLE

For most questions you might have about your data, you will want to look at descriptive statistics, then look at graphs, then run a statistical test. Let's say, for example, that you had data from the 1930s, indicating that the average country had a corruption level of "5" at that time. We will find descriptive statistics for our correption in our modern sample, we visualize the data using a Histogram, then we will run a statistical test to help determine whether the differences we think we see could be due to random variation. We hope that the test below reveals that the probability is less than 5% (less than .05) that the difference we see can be accounted for by chance.

Descriptives – Find the mean and standard deviation of a continuous variable

- **Analyze, Descriptive Statistics, Descriptives**
- Select variable (**Corruption2**) and move it to the variable box using the right arrow.
- **OK.**
- Note, the group mean in your sample is 5.9, which is higher than the 5.0 from the 1930's.

Histogram – A bar-chart with "count" as the Y axis

- **Graphs, Legacy Dialogs, Histogram**
- Select variable (**Corrupt_Categ**) and move it to the Variable box using the right arrow.
- **OK.**
- **Graphs, Legacy Dialogs, Histogram**
- Select variable (**Corruption2**) and move it to the Variable box using the right arrow.
- **OK.**

T-test – Determine if the mean of a variable differs from a set value

- **Analyze, Compare Means, One-Sample T Test**
- Move (**Corruption2**) to the "Test Variable(s):" window
- Under "Test Value:" put 5.
- **OK**
- In the Output window, the "Sig. (2-tailed)" shows .000, which indicates the probability is less than .001 (less than 1 out of 1000 chance) that your data is so far above the 1930's value of 5, due to random chance. Results would typically be reported with the value of the t statistic itself, the degrees of freedom (df), and the probability. For example:

  > Based on a one-sample t-test ($t(156) = 5.57$, $p. < .05$), there *is* a statistically significant difference between the mean of the modern sample and the mean of the historic data.

IV. ASKING QUESITONS THAT COMPARE THE MEANS OF DIFFERENT GROUPS

Here we will try to determine whether gdppc differs based on the corruption categories that we created earlier. We will "explore" to determine the mean gdppc for each category, then create a bar chart, then run a t-test and an ANOVA to explore the difference.

Explore

- **Analyze, Descriptive Statistics, Explore**
- Select variable (**gdppc**) and move it to the "Dependent List"
- Select variable (**Corrupt_Categ**) and move it to "Factor List"
- **OK.**

Bar Chart

- **Graphs, Legacy Dialogs, Bar**
- Keep it "Simple" and "Summaries for groups of cases" and click **Define**.
- Select variable (**Corrupt_Categ**) and move it to the Category Axis box using the right arrow.
- In "Bars Represent", select "Other Statistics (e.g., mean)
- Select variable (**gdppc**) and move it into the now-available "Variable" box using the right arrow.
- **OK.**

T-test – Compares the means of two (and only two) groups

- A t-test is the most basic way to compare two groups, based on their scores on a single variable. While we divided the **Corruption2** into three groups, you can only compare two of those groups at a time using a t-test.
- **Analyze, Compare Means, Independent-Sample T Test**
- Select the test variable (**Gdppc**) and the grouping variable (**Corrupt_Categ**)
- **Define Groups** and enter Group 1 as "2" and Group 2 as "3".
- **Continue, OK**
- In the Output, look for the "t", which gives the value of the t-test, the "df", which tells you your degrees of freedom, and the "Sig. (2-tailed)", which tells you the probability of getting that result. Remember that, in general, you are looking for a probability value less than .05.

One-way ANOVA - F-test – Compares the means of more than two groups

- **Analyze, Compare means, One-way ANOVA**
- Select (**Gdppc**) variable and click right arrow to move it into the **Dependent List** section.
- Select (**Corrupt_Categ**) variable and click right arrow to move it into the **Factor** section. **OK.**
- This output is more complicated. The "Corrected model" line gives you your F-value and its significance, but you care about the degrees of freedom for both that value and the error term. The output indicates a significant difference between the groups, and you report this as

  $F(2,148) = 219.11, p<.05.$

V. QUESTIONS ABOUT THE RELATIONSHIP BETWEEN CONTINUOUS VARIABLES

Here we will try to determine the relationship between gdppc and corruption, leaving corruption on its original scale from 0 to 10. The most basic graph for two continuous variables is a scatter plot. The scatter plot will show us that the two variables appear to be related. We will follow this visual intuition up with two tests: First, a correlation tells us how tight the relationship is between the two variables. A correlation close to 0 indicates that that there is no relationship, a correlation close to 1 (or negative 1) indicates that the relationship is very close to a perfectly straight line. Second, a regression tells us the nature of the line in question. A line is defined by a slope and an intercept, and the regression output will tell you if either of those is significantly different than zero.

Scatter Plot

- **Graphs, Legacy Dialogs, Scatter/Dot**
- Keep it "Simple Scatter" and click **Define**.
- Select variable (**gdppc**) and move it into the "Y Axis" box using the right arrow.
- Select variable (**Corruption2**) and move it to the "X Axis" box using the right arrow.
- **OK.**
- **Double click to edit graph, and click icon to "add fit line at total".**

Correlation

- The correlation coefficient is a measure of association between two variables. The sign indicates inverse versus directly proportional relationships. 0 would be a non-relationship, while a value of 1 would be a perfect positive relationship, and -1 would be a perfect inverse relationship.
- **Analyze, Correlate, Bivariate**.
- Select your two variables (**Gdppc, Corruption2**) and move them to the **—**Variables box.
- Leave all default selections **—**as-is (Pearson, Two-tailed, and Flag significant correlations).
- **OK.**
- Your output will be a 2x2 table, but only one cell is relevant. Use either cell that shows the relation between your two variables (there are two such identical cells). The first value in that cell is the correlation coefficient.

Regression

- Regression provides additional information about the relationship between two variables. Specifically, it allows you to guess the value of one variable given information about another variable.
- We will regress **Gdppc** (our dependent variable) over **Corruption2** (our independent variable).
- **Analyze, Regression, Linear**.
- Put variable (**Gdppc**) in the **—**Dependent box.
- Put variable (**Corruption2**) in the **—**Independent box.
- **OK**.
- This output can be used to make a Linear equation:
    - **Gdppc** = b0 + b1* **Corruption2**