**POS 3713: Assignment 4**
**Assigned: Monday, 3/13/2000**
**Due date: In class, Monday, 3/20/2000**
**Tutorial Session: Thursday, 3/16/2000, 9am-10:45am; and Friday, 3/17/2000, 9am–10:45am**
**Professors Mitchell and Lubell**

**Instructions: Everyone must answer the questions in Part A. You can then choose to answer the questions in Part B OR Part C. For Dr. Mitchell's class, we recommend answering the questions in Part B, and for Dr. Lubell's class, we recommend answering the questions in Part C. You have the option to do all three sections; if you do, you will receive extra credit points. You must type your responses to each question, showing all relevant work.**

The purpose of this assignment is to introduce you to hypothesis testing. For the first time, we will be using both the 1996 National Election Study and the 1995 World Survey. Begin this assignment by opening the file called "World95.sav" in SPSS. You can find the World95.sav file at the same Internet address used to access the NES data. To launch SPSS, click on the Start button, select Programs, and then click on SPSS Windows 9.0. When you are finished with this assignment, be sure to save your working data file by clicking on "File", and then "Save As". You will want to save the working file for the 1995 World Survey under a new name, so that you will not overwrite the changes you have made to the NES data set.

## Part A: Hypothesis Testing, Mean

Hypothesis tests for the mean involve the comparison of a sample to a larger population. We usually want to test if the sample was drawn from the population in question, or some other population. For example, suppose we wanted to determine if elderly citizens are more likely to be victims of criminal behavior than all citizens. Suppose we know that the mean number of times people are victimized per year in the U.S. is 1.3 ($\mu = 1.3$). We collect a random sample of elderly citizens and discover that their (sample) mean victimization rate is 4.3, while the sample standard deviation is 1.4. To determine if elderly citizens are victimized more frequently than the population at large, we set up the following hypothesis test:

$$H_o: \mu = 1.3$$
$$H_1: \mu > 1.3$$

If there is no difference between elderly victimization rates and general victimization rates in the U.S., then the elderly group will be drawn from the general population, and hence will have a population mean of 1.3. If the elderly group is different, then it will be drawn from some other population that has a mean higher than 1.3. If we reject the null hypothesis, we could conclude that elderly citizens are more likely to be victims of crime.

The goal of this exercise is to determine if there is a significant difference in literacy rates based on the level of urbanization in a country. Assume that the World Survey collected in 1995 contains data for the population, thus any statistics calculated using all of the countries gives you the population values (such as the mean). We are going to compare various sub-groups, which we will assume are randomly selected. Begin by calculating the mean literacy level for the population (all countries).

- Select “Analyze”, “Descriptive statistics”, then “Descriptives” to open the “Descriptives” window.
- Move *Literacy (People who Read %)* into the variables text box, and click OK.

You are going to compare the literacy rate of the population (what you just calculated above) to the literacy rates of countries that are less urbanized. We will be looking at the group of countries who are below the mean level of urbanization for the population ($\mu = 56.5\%$). You will be testing the hypothesis that these countries have a *lower* literacy rate than the population at large. To test this hypothesis, you must tell SPSS to select only those countries with less than the mean level of urbanization:

- Select “Data”, and then “Select Cases”. This will open up a "Select Cases" window, where you have several options. You should click on the box next to "If condition is satisfied", and then click on the "If" button below that option. This will open up another window.
- Move *Urban (People living in Cities %)* into the box on the right. Click on the < key in the keypad below the box, and then enter 56.5 (either on your keyboard or using the numbers on the keypad). Click Continue, and then OK. Make sure that the option for "Unselected Cases Are" is filtered rather than deleted (SPSS uses filtered as the default). You will notice that SPSS places a hash mark through the cases that do not meet these criteria.
- Now re-run your descriptive analysis for the literacy variable. Select “Analyze”, “Descriptive statistics”, then “Descriptives” to open the “Descriptives” window. The variable, *Literacy (People who Read %)*, should already be in the variables text box; click OK.
- When you are finished, select "Data", "Select Cases", click on "All Cases", and then "OK". This will de-select the cases that have lower levels of urbanization (if you did this correctly, you should not see any hash marks across the case numbers in the data window).

**Exercise:** Use the information calculated above to answer the following questions:

1) Produce a descriptive statistics table for a) the entire population, or all countries and b) the subset of countries that are less urbanized.
2) Identify the null and alternative hypotheses (remember that you are testing the hypothesis that countries with lower levels of urbanization will have lower literacy rates, thus the test is one-tailed).
3) Conduct a hypothesis test assuming that $\alpha = .05$ (95% confidence). Since we are assuming that the population consists of data for all countries, you should use the standard deviation for this group in your hypothesis test (i.e., $\sigma = 22.88$). What can you conclude based on this test? Do nations with low levels of urbanization have lower literacy rates than the general population of countries? Is this what you would have expected based on your knowledge about urbanization and literacy rates?

Note: If you are answering the questions in Part B, then save the World95.sav with a new working file name (e.g., w95work.sav). Now open your saved working file for the 1996 NES data set. If you are answering the questions in Part C, leave the World95.sav file open in SPSS.

**Part B: Hypothesis Testing, Proportions**

Hypothesis tests for proportions are similar to hypothesis tests for the means, but they are typically used for nominal level variables. The central question is still "Does the population from which the sample was drawn have a certain characteristic?" In the exercise below, we want to determine the if the 1996 NES sample is representative of the larger population, based on certain known characteristics of the population.

We want to determine if the data collected in the 1996 NES survey is representative of the general population based on the party identification of the respondents. Suppose we know from Census data that party identification in the United States is distributed as follows (this is data for the population):

| Party | Proportion | Percentage |
|---|---|---|
| Democrats | 0.40 | 40% |
| Independents | 0.30 | 30% |
| Republicans | 0.30 | 30% |

**Exercise:** Using the procedures described below, answer the following questions.

1) Calculate the 95% confidence interval for the proportion of Democrats, Independents, and Republicans using formula 7.3 in Healey (page 162). Use the sample proportions obtained in SPSS (include your frequency table for the "Partyid" variable in your output). How do you interpret these confidence intervals? Is the actual population proportion value listed in the table above contained in these intervals?
2) Using the proportion of *Democrats*, test the hypothesis that the NES sample comes from the general US population (under the null hypothesis) assuming that $\alpha = .05$ (95% confidence); see formula 8.3 in Healey (page 195). Identify the null and alternative hypotheses being tested. What can you conclude based on this test? Is the proportion of Democrats surveyed in the 1996 National Election Study significantly different than the proportion of Democrats in the general U.S. population? Is this consistent with what you found in question #1?

- Produce a frequency distribution of the variable "Partyid" (click on "Analyze", "Descriptive Statistics", "Frequencies", and then move this variable into the box and click OK). You should have already renamed this variable (V960417) in a previous assignment, and recoded the other party/don't know responses as missing. Thus your frequency distribution should produce only three valid responses (Democrat, Independent, and Republican). If any other responses appear in the frequency table, then *you must* re-code these values are system-missing.
- Calculate the proportion of Democrats, Republicans, and Independents by dividing the frequency for each category by the total number of valid cases (do not include those cases that are defined as missing). Use these calculated proportions to answer question #1.
- Use the total number of valid cases (N = 1580) for your hypothesis testing calculation in question #2.

**Part C: Hypothesis Testing with Sample Means, Matched Samples**

The purpose of hypothesis testing with sample means is to compare the mean level of a variable between two groups. In the case of matched samples, we cannot assume that the two groups we have randomly selected are independent. For example, we might want to determine if men and women who are married have different attitudes about gun control. We cannot assume that their attitudes are independent because a husband's views on gun control may influence his wife's views (and vice versa). Suppose that we collect the following data for 5 married couples (assume that we measure each respondent's attitudes about gun control on an 0-100 interval scale).

| Couple # | Husband | Wife | Difference (Husband Support - Wife support) |
|---|---|---|---|
| 1 | 75 | 60 | 15 |
| 2 | 90 | 85 | 5 |
| 3 | 60 | 65 | -5 |
| 4 | 30 | 20 | 10 |
| 5 | 45 | 42 | 3 |

We can see that in most cases, women score lower than their husband's on the support for gun control scale. The hypothesis test will tell us if this difference is large enough to justify the conclusion that it did not occur by random chance alone, but rather reflects an actual difference between husbands and their wives on this issue. The null and alternative hypotheses for this test are:

$$H_o: \mu_D = 0$$
$$H_1: \mu_D \neq 0$$

Rejecting the null hypothesis would lead us to conclude that there is a difference between husbands and wives with respect to their attitudes about gun control.

We are going to conduct a similar test using the 1995 World survey data (World95.sav). We want to determine if there is any difference between men and women in terms of their life expectancy (using the variables LIFEEXPF and LIFEEXPM). This is a matched or paired samples test because life expectancy rates for men and women in particular countries are not independent (we would expect nations that have higher life expectancy rates for men to also have higher life expectancy rates for women).

**Exercise:** Using the procedures described below, answer the following questions.

1) Compare the mean life expectancy rates between men and women. Do they seem different just by looking at the descriptive statistics?
2) Test the difference between male and female life expectancy (use a matched or paired samples t-test). State the null and alternative hypotheses in words and mathematically. What can you conclude based on this test? Do men and women have significantly different life expectancy rates? What could account for this difference?

- Before you begin, make sure all cases are selected (there are no hash marks through any of the cases from the procedure in Part A). If not, go back up to the instructions at the end of Part A to re-select all cases for analysis.
- Select "Analyze", "Compare Means", "Paired-Samples T-Test".
- Click on "Average Male Life Expectancy" and "Average Female Life Expectancy" in the box on the left. Next, click on the right arrow. If you have

done this correctly, you should see "lifeexpf - lifeexpm" in the Paired Variables Box. Click on Ok to continue.

- The first box in the output window contains the descriptive statistics for each variable. You can ignore the second box (we will learn more about correlation in a few weeks). The third box contains the calculated t-statistic (which is the value for t(obtained) in Healey, on page 212), in addition to the degrees of freedom (df) for the test. Use this degrees of freedom value to find the critical value of t in your book (Appendix B).