POS 3713: Homework Assignment #3
Spring 2001
Due on Friday, March 9th

Instructions: **Type** your answers to the following questions. You are permitted to do any calculations by hand on paper and attach the work to your typed responses. You should, however, report the final results of your calculations in the typed portion of your assignment.

## 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. We will be using data from the 1995 World Survey. Suppose we know that the mean literacy rate in the population of all countries is 78.34 (over ¾of all people are literate on average), or $\mu = 78.34$.

You are going to compare the literacy rate of the population (given 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. You will be testing the hypothesis that these countries below the mean level of urbanization have a *lower* literacy rate than the population at large. The table below contains the sample information on literacy rates for countries with low levels of urbanization (this is a sample that you are comparing to the population of all countries)

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| People who read (%) | 49 | 18 | 99 | 64.24 | 24.51 |
| Valid N (listwise) | 49 | | | | |

Question A1: 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).

Question A2: 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, which is $\sigma = 22.88$. 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?

**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 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% |

The frequency table below reports the same information for our 1996 NES sample. Refer to the valid percent column (the percentage of democratic respondents is 42% or p=.42).

**PARTY ID**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | 1. Democrat | 663 | 38.7 | 42.0 | 42.0 |
| | 2. Republican | 471 | 27.5 | 29.8 | 71.8 |
| | 3. Independent | 446 | 26.0 | 28.2 | 100.0 |
| | Total | 1580 | 92.2 | 100.0 | |
| Missing | System | 134 | 7.8 | | |
| Total | | 1714 | 100.0 | | |

Question B1: 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 in the second table above. Note that the sample size (N) is 1580 (in other words, ignore the missing values). How do you interpret these confidence intervals? Is the actual population proportion value listed in the first table above (from the Census) contained in these intervals?

Question B2: 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?

**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 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. We want to determine if there is any difference between men and women in terms of their life expectancy. 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).

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | Average female life expectancy | 70.16 | 109 | 10.57 | 1.01 |
| | Average male life expectancy | 64.92 | 109 | 9.27 | .89 |

Question C1: Compare the mean life expectancy rates between men and women referring to the paired sample statistics table above. Do they seem different just by looking at the descriptive statistics (the mean and standard deviation)?

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference Lower | Upper | | | |
| Pair 1 | Average female life expectancy - Average male life expectancy | 5.24 | 2.27 | .22 | 4.81 | 5.67 | 24.109 | 108 | .000 |

Question C2: Test the difference between male and female life expectancy (use a matched or paired samples t-test) referring to the paired samples test table above. The reported t-value (24.109) is the value for t(obtained) in Healey on page 212; in other words it is already calculated for you. 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?