DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS

Posc/Uapp 816

WINDOWS, MINITAB, AND INFERENCE

I. AGENDA:
   - A. An example with a simple (but) real data set to illustrate
     - 1. Windows
     - 2. The importance of saving data
     - 3. MINITAB
       - a. Entering and saving data
       - b. MINITAB statistics
   - B. Reading:
     - 1. The ***Student Edition of MINITAB for Windows***, Chapters 2, 4, 5.
       - a. Read the Tutorial as needed.

II. EXAMPLE:
   - A. Background: suppose someone asserts that welfare programs have made a huge difference in the quality of people's lives. In particular, they have improved health care. As an example, the infant mortality rate in 1992 (defined as the number of infant deaths per 1,000 live births) was 8.5.[1] He further asserts that this rate must have been higher in the years before 1990 but doesn't have any data to make the point. So we decide to investigate the situation ourselves.
     - 1. We'll have to make an educated guess about the tenability of this argument, since we don't have the time or resources to do a complete analysis.
   - B. We'll proceed as follows:
     - 1. First (the gods willing) we'll cruise the internet to collect some data pertaining to infant mortality;
     - 2. Then draw a small sample (to save time and labor);
     - 3. And finally, using MINITAB to analyze the sample data, we will test the hypothesis that the infant mortality rate was higher 1988.
   - C. First though, let's get familiar with Windows.

III. WINDOWS DEMONSTRATION:
   - A. This material repeats somewhat the information presented in class 1.
   - B. Some basics

[1] According to ***Health United States 1994*** Table 23, page 87.

1. **Note: there are many different ways to do the same thing.**
   a. What I demonstrate or talk about is usually just one of many ways to accomplish the same thing.
2. The **desktop**
3. Multiple windows
   a. Sizing, opening, closing, minimizing (hiding) windows
   b. Scrolling
4. Menu and task bars
   a. "Dialogue" boxes
5. Keeping track of windows: the title bar
6. "GUIs"
7. Mouse, pointer, cursor, insertion point
   a. Left and right buttons.

C. Moving information from one window to another.
1. Cutting and copying and pasting text on the clipboard.
2. Hold down and drag the left mouse button to highlight text.
3. After text has been highlighted, press right button to see list of options such as cutting or copying.
4. For "fine tuning" mover cursor to start of the text to be block; hold the shift key down; and move the cursor to the end of the text to be selected.
5. Note: most Windows programs (that is, programs that "run" under the Windows operating system) allow you to highlight (select) text and cut and paste to the clipboard, even if you can't move it in the application.
   a. Example: the University provides a communications package that has a telnet program for terminal "emulation," as discussed last time.
      i. When using, say, Pine, which is not a screen editor (that is does not permit the usual cut and paste operations with the mouse), you can still select text and place it in the clip board.
   b. Similarly, you can highlight text on internet pages and drop or paste it on the clipboard.

D. Help menus

IV. USING THE INTERNET TO COLLECT DATA:

A. Everyone is no doubt familiar with web surfing to find information.

B. But to expedite matters, lets go to the course web page, **www.udel.edu/htr/Statistics**, and then click on **Information, software, and data sources**.
1. Go to the Census Bureau Home page (**www.census.gov**).
2. Go to **Subjects A to Z**; then click on **C**, since we want to look for the **City and County Data book**.
3. Then scroll down to **Top City Rankings** and click on "Infant Death Rates,

Crime Rate"

a. The address for these data is: **www.census.gov/statab/ccdb/ccdb309.txt**

C. Now that we have some potentially useful data we need to move it to the desktop, store it, convert it to a format suitable for processing by MINITAB.

1. There are lots of ways of doing this.

a. You can use the cursor to drag and highlight text and then copy it to the clipboard. From there it can be put in Notepad or a word processing program such as Word.

b. Or you can click on **File** and then **Save as...** In the **Save as...** dialogue box pick a folder in your hard drive or, if you are using a public computer, the A: drive. Type a file name and **Ok**.

c. At worst you can print the page or if desperate enough you could copy the information.

D. Here is some more information about storing data; read it as needed.

V. KEEPING DATA FOR FUTURE USE:

A. It is important that you know how to save data and other information in proper formats.

B. Folders and files

1. As noted above files are stored in folders or even sub-folders under two-part names. The first describes the file's content, the second (the extension) its type.

C. MINITAB (and SPSS) file formats versus plain "text" data.

1. Both SPSS and MINITAB can read pure text and data files, which have normally have extensions **.txt** and **.dat** respectively.

2. But they cannot read each other's file types (e.g., **.mtw**). So make sure that you name the files accurately and remember them.

D. Public versus private desktop machines.

1. If you are using a public site, be sure to save your data and results on your own **diskette**.

2. Before starting make sure that you have a **formatted disk**.

E. To format a disk with Windows 95

1. Make certain write tab is not in **protected** position.

2. Insert a **blank** (unused) disk into (usually) the A: drive

3. Double click on "My Computer" (on desktop) or click Start button (at lower left of monitor) and then Windows Explorer.

4. Double click on **31/2 Floppy**

5. Follow the instructions: check "full" and type a label if you want one.

VI. CONVERTING OR REFORMATTING DATA:

A. Strictly speaking both SPSS and MINITAB read and process "alphanumeric" data, that is files containing both numbers and textual labels.

B. But it is much easier when dealing with small amounts of data to remove **all** alphanumeric characters before entering them into one of these programs.

C. Again, there are lots of ways to do this.

1. I use a word processor such as Word or Word Perfect.

a. Word Perfect allows you to select rectangular blocks of text, which can be deleted. This feature, which can be found in the **Edit, Select** menus, requires that you put the insertion point at the beginning of the text to be highlighted and then drag the cursor to the end. You will have to highlight more text than you want, but when you choose **Select rectangle** only block of text you want will be chosen. Then you can delete it or copy it to the clipboard.

b. Word has a similar feature that is perhaps easier. Just hold the **alt** key while highlighting the block of text.

2. I also use a the word processor to look for stray alphanumeric characters in the data.

a. For example, the data in the present example contain the abbreviation **NA** for cities not having data. To indicate to MINITAB that these are "missing" use the "Find and replace" tools to change them to *****, which in MINITAB stands for missing value. Or you can change the **NA** characters to, say, 99999 and then declare it to be a missing value.

b. **Note**: that the simple way of using MINITAB requires that all commas be removed. It cannot, in other words, correctly interpret 1,222 because of the comma. So again use the "find and replace" tools in your word processor to change all commas to nothing. (Example: the data will now appear as 1222.)

D. **After converting or reformatting the data make sure they are stored as DOS (ascii) text, not as word processor file.**

1. All of this may seem complicated but it only takes a few minutes.

VII. MINITAB:

A. Start the program by clicking Windows start button, moving cursor to program name, and clicking.

B. "Resize" the windows as needed.

C. Open data window first.

1. See Figures attached figures.

D. Click arrow in upper left corner to make it point down.

E. Enter numbers in columns starting with first row.

1. **In MINITAB and SPSS never enter percent signs or commas; only numbers and decimals.**

F. Use cursor or pointer to highlight box at top of column.

G. Type a short name, one having less than 8 characters.

H. Enter and name other data.

- I. Click file, then **Save worksheet as...**
  - 1. Under type of worksheet select **MINITAB**
  - 2. Click **Select File**
  - 3. Filling dialogue box.
    - a. Make sure that you save the file in the proper location.
- J. Now the data are stored in a format that can be used by MINITAB.
  - 1. Follow similar procedures when using SPSS.

VIII. STATISTICAL INFERENCE:

- A. Now lets briefly review some terms and basic concepts
- B. ESTIMATION:
  - 1. Terms:
    - a. Population parameters, usually denoted with Greek letters
      - i. Examples:
        - ○ **μ**: population <u>mean</u>
        - ○ **σ**: population standard deviation
        - ○ **Δ**: difference of means, $\mu_1 - \mu_2$
    - b. Sample statistics
      - i. These numbers are calculated from a sample that has been drawn **randomly** from a population.
      - ii. They are usually, but not always, denoted with a Greek letter with a caret "^" over them.
      - iii. Examples:
        - ○ $\overline{Y}$: sample mean (exception to rule)
        - ○ **σ̂:** Sample standard deviation
        - ○ $\hat{\ }$: sample estimate of difference of means
    - c. Sample statistics are the basis of inferences about population parameters.
- C. Estimation:
  - 1. **Expected value, E**: the "long run" value of a sample statistic over many independent samples.
  - 2. **Bias**:
    - a. If $E(\hat{\ }) = \ $, where $\hat{\ }$ is a sample estimator of Θ, then $\hat{\ }$ is called <u>unbiased</u>.
    - b. Example: since $E(\overline{Y}) = \mu$, then $\overline{Y}$, the sample mean is an unbiased estimator of the population mean; that is, the mean of the population from which the sample is drawn.
    - c. If $E(\hat{\ }) \neq \ $, then $\hat{\ }$ is called a biased estimator.
      - i. Example: the sample standard deviation calculated by dividing by N, instead of (N - 1) is a <u>biased</u> estimator of the population standard deviation, **σ**.
- D. Sampling distributions:

1. Consider this thought experiment: we are interested in comparing large and small cities in terms of some variable such as "the infant death rate." Suppose we take two independent samples from the population of counties in each state. (Denote the sample sizes N_L and N_S respectively. That is, we might draw a sample of 10 counties from the "population" of large cities and 15 from smaller ones. Then, $N_L = 10$ and $N_S = 15$.) After finding out the mean or average infant death rate in the two types, we then calculate:

$$\boxed{\hat{\Delta} = \bar{Y}_L - \bar{Y}_S}$$

2. Now suppose we repeat the process by drawing another sample of N_L cases from the large cities and N_S from the small ones and calculate the difference of means on the infant mortality rate. This estimated difference will probably not equal the first one because we have drawn new independent samples.
3. Let's continue in this manner, drawing samples of size 10 and 15 respectively from the populations of large and small cities an infinite number of times. Stated differently, suppose we some how obtain sample differences of means from all possible samples from these two kinds of cities. We will have a "pile" of $\hat{\ }'s$.
4. What will be the mean and standard deviation of this collection?
5. A sampling distribution is a theoretical distribution that shows the relationship between the possible values of a sample statistic ($\hat{\ }$, for example) and the probability or likelihood of observing these values, given that the samples are drawn independently from a population with a corresponding parameter (Δ here).

E. Instead of some population parameter of interest we could also describe the distribution of **test statistics** (such as the **z** or **t** statistics).

1. For example, suppose we draw a small sample from some population, say, American cities, in order to test a statistical hypothesis such as

$$\boxed{\begin{array}{c} H_0\text{: } \mu_{1988} = 8.98 \\ \textit{versus} \\ \mu_{1988} > 8.98 \end{array}}$$

2. For a small sample the test statistics is

$$\boxed{t = \frac{(\bar{Y} - \mu)}{\hat{\sigma}}}$$

3. Were we to draw repeated independent samples of size N from the population, these t's would have a t distribution with N - 1 degrees of freedom. (See below.)

F. Parameters of a sampling distribution.

1. The mean of the sampling distribution usually equals the expected value of the sample statistic.

a. The mean of the difference of means, for example, will equal Δ , the population difference of means. In other words,

$$\boxed{E(\hat{\Delta}) \; = \Delta \; = \mu_L \; - \mu_S}$$

2. The standard deviation of the sample statistics, that is the standard deviation of the sampling distribution, is called the standard error. In the case of the difference of means we would denote it $\sigma_{\hat{}}$.

G. The form of the sampling distribution.

1. If the sample sizes are large enough and other conditions are met, the sampling distribution of a sample statistic ($\hat{}$) will be **normal** with mean and standard deviation (called the standard error) of $\sigma_{\hat{}}$.

2. As just noted the distribution of sample t statistics is t with N - 1 degrees of freedom.

3. In the case of the difference of means statistic where the two sample sizes are relatively small, $\hat{}$ will have a so-called t distribution. If the population variances (standard deviations) are the same (e.g., σ_L equals σ_S), then the sampling distribution will be the t distribution with (in this case):

$$\boxed{df \; = N_L \qquad N_S \qquad 2}$$

IX. GENERAL STEPS IN TESTING HYPOTHESES:

A. For the t-test described next see Agresti and Finlay, ***Statistical Methods for the Social Sciences***, pages 180 to 183. For hypothesis testing in general see Chapter 6.

B. Problem: for a moment let us return to the original problem and suppose that we have only a (random!) sample of cities and we want to test the hypothesis that the infant mortality rate has declined since the 1980s.

1. We have data for American cities for 1988 so we can use that year to formulate a hypothesis.

2. Although we have data for the entire "population," we'll take a small sample, say 4 cities, for expository purposes.

3. Let's let μ represent the mean infant mortality rate in the population of

American cities and μ_i be the rate for year the year i where i is 1988 or 1992.

4. In effect, for this simple case we'll test the hypothesis that the mean infant death rate per 1,000 births was higher in 1988 than in 1992.
    a. There are all sorts of ways to express the hypothesis.

C. Hypotheses:

1. Research: H_A: $\mu_{1988} > \mu_{1992}$ (that is, µ, the infant death rate was higher in 1988 than in 1992.)
2. More specifically, in this weird case we know the population value (see page 1) so the research hypothesis is $\mu_{1988} > 8.5$
3. Null: H_0: $\mu = \mu_{1988} = \mu_{1992} = 8.5$
4. Perhaps stated more simply we are testing H_0: $\mu = 8.5$ versus H_A: $\mu > 8.5$

D. Sampling distribution:

1. Given certain conditions (independent random sampling, relatively small N's, etc.) and the null hypothesis, we expect a test statistic will have a certain distribution. If we know the properties of the distribution we can use it to gauge the likelihood of sample results. In this case the test statistic will be the t distribution, a symmetric roughly bell shaped distribution centered at 0 whose exact form is determined partly by the degrees of freedom (see below).
2. Critical region:
    a. Some possible sample results will be deemed so unlikely that should one of them occur we will reject the null hypothesis. Others will be considered possible and will cause us to accept the null hypothesis.
    b. Because we do not know the true situation we might make an error in rejecting or accepting the null hypothesis.
    c. Errors
        i. Type I error: falsely rejecting the null hypothesis. The probability of making a type I error is alpha (α) and is called the level of significance. Usually one chooses a level of significance (an alpha level) of .05, .01, or .001.
        ii. Type II error: falsely accepting the null hypothesis.
    d. Critical region: the choice of alpha determines the critical region, namely those sample outcomes that will lead to rejection of the null hypothesis. Thus, if a test statistic falls in the critical region, we reject the null hypothesis.
        i. In this case we'll use the .025 level and a one-tailed test.
    e. Critical value(s) marks off the critical region.
        i. We need a critical value that marks off the top 2.5 percent of the t distribution.
        ii. Since there are N - 1 = 4 - 1 = 3 degrees of freedom, we see (from Agresti and Finlay's table, page 669, or any t table) that the critical values is 3.182. Thus our observed t must equal or be greater than 3.182 in order for us to reject (not

accept) the null hypothesis that $\mu_{1988} = 8.5$.

3. Test statistic:

a. The appropriate test statistic is determined by the data, the nature of the problem, and the assumptions that are made.

b. For this example, we are interested in testing a hypothesis about a mean. The sample size is relatively small (N = 4), and we do not know much about the population (its standard deviation, for example). These considerations suggest using the t statistic, which is computed by the formula given below. The observed t has a t distribution with N - 1 degrees of freedom.

c. The test statistic is

$$\boxed{t_{obs} = \frac{(\bar{Y} \quad \mu)}{\hat{\sigma}_{\bar{Y}}}}$$

d. The sigma in the denominator, the **estimated standard error** of the mean, has to be estimated.

i. This is why we use the t-distribution.

ii. If **σ**, the population standard deviation were known, we could use the z statistic and the standard normal distribution.

e. The formula for the estimated standard error is

$$\boxed{\hat{\sigma}_{\bar{Y}} = \frac{\hat{\sigma}}{\sqrt{N}}}$$

i. Here $\hat{\sigma}$ is the sample standard deviation.

4. Decision:

a. Compare the observed statistic (here the t) with the critical value. If it is larger, then reject the null hypothesis; otherwise, do not reject H_0.

X. MINITAB:

A. Let’s use MINITAB to do the analysis. The data are stored in a worksheet called “infant-crime.”

B. Start MINITAB, go to **File**, then click **Open worksheet...**

1. In the dialogue box find the file, here “infant-crime.”

2. If you have “raw” data, open **Other files** and **Special files** data.

3. If using the student version, go to **File** and the **Import ascii**

C. Now let’s draw a sample of 4 cases.

1. Go to **Calc**, then **Random data**, then **Sample from columns**.

2. In the dialogue box enter 4 in the sample window, choose c1 (infant) from

the variables, and store the sample in c5.

3. Press **Ok**.

4. You can rename c5 to sample.

D. Click **Stat**, then **Basic Statistics**

1. Choose **1-sample t**

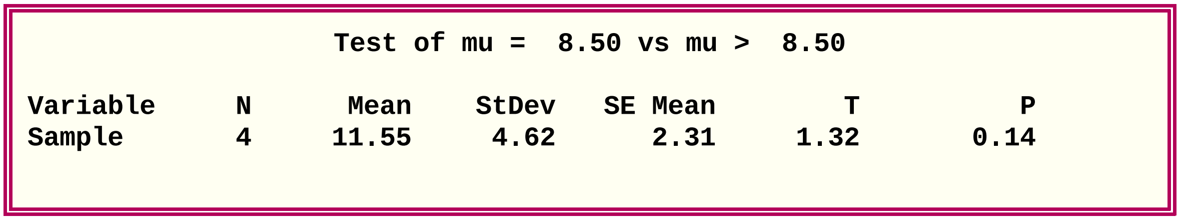
2. Highlight variable name infant by clicking on it and then clicking **Select**

3. The check **Test mean**.

4. In the window type 8.5.

5. Then select from the **Alternative** box **greater than**.

6. Here are the results:



```
                    Test of mu =  8.50 vs mu >  8.50

Variable      N      Mean     StDev   SE Mean        T          P
Sample        4     11.55      4.62      2.31     1.32       0.14
```

a. The sample mean, $\bar{\boldsymbol{Y}}$**,** (based on 4 cases) is 11.55, with a standard deviation of $\hat{\sigma} = \mathbf{4.62}$

b. The observed t is 1.31, which does not fall in the critical region, suggests that the null hypothesis should continue to be accepted.

c. But of course this test involved only four cases.

7. We can of course observe the 1988 “population” mean for the cities, which is 12.039 and hence considerably larger than the 1992 average.

a. So assuming that this parameter corresponds to the 1992 figure we can see that there has in fact been a decline in infant death rates, but our small sample did not “pick it up.”

E. Note that of course all of this has been a numerical exercise since the data used to compute our 1988 sample are not at all comparable to the 1992 data used in calculating the average.

XI. NEXT TIME:

A. Further examples of MINITAB

B. Difference of means test.

C. Confidence intervals.