DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS

Posc/Uapp 816

CONFIDENCE INTERVALS AND TWO SAMPLE TESTS

I. AGENDA:
   A. Confidence intervals
      1. MINITAB example
   B. Two-sample means tests
   C. Reading:
      1. Agresti and Finlay, ***Statistical Methods for the Social Sciences***, 3rd edition:
         i. Basic ideas of statistical tests: pages 155-167.
         ii. "Z test" (pages 159-162): use z when the population standard deviation is known or when sample size is large (> 40).
         iii. Statistical decisions and errors pages 173-177.
         iv. Comparison of tests and confidence intervals: pages 177-180.
         v. Small sample t test: pages 180-187.
         vi. Two-sample tests: Chapter 7.

II. CONFIDENCE INTERVALS:
   A. This section contains notes reprinted from Class 3.
   B. General idea: William Hays, a psychologist and statistician, likens confidence intervals to tossing rings at a fixed post. The post symbolizes an (unknown) population parameter. The rings stand for confidence intervals. Their size represents what we think we know about the parameter: the smaller the ring, the more we know. Large rings mean that we know very little. Now, the goal is to throw the rings at the post in hopes that they will slide down the shaft (that is, cover the post). If the rings have a small diameter, it will be hard to hit the post; if they are large, it will be relatively easy. Thus, there's an inverse relationship between our ability to hit the post and our level of knowledge. For example, tossing a hula hoop would allow us to hit the post almost every time, but it wouldn't be much of a challenge (that is, it wouldn't convey much information). If the rings were much smaller and we managed to hit the shaft, we could feel proud (confident) of our ability, which in this example means we would have considerable knowledge about the post.
      1. Agresti and Finlay (page 126 of the 3rd edition of their book) explain confidence intervals this way: "A **confidence interval** for a parameter is a range of numbers within which the parameter is believed to fall."
      2. Some intuitive idea of confidence intervals can be seen by referring to

figures 1 and 2 attached to the last set of notes.

C. To illustrate the idea of confidence interval in more detain let's first estimate a population mean, **μ,** using a large sample. The so called “central limit theorem” indicates that the sampling distribution of $\bar{Y}$, the estimator of μ, will (under certain conditions) be normal with $E(\bar{y}) = \mu$ and standard error (deviation) $\sigma_{\bar{Y}}$.

1. Usually, $\sigma_{\bar{Y}}$. has to be estimated by:

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

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

D. Recall that 95 percent of the normal distribution lies within two 1.96 standard deviations of the mean; 99 percent falls within 2.56 standard deviations.

E. Under these (and other) circumstances, we can say that over all samples of size N, the probability is approximately .95 that

$$\boxed{-1.96\hat{\sigma}_{\bar{Y}} \leq \mu - \bar{Y} \leq +1.96\hat{\sigma}_{\bar{Y}}}$$

1. That is, very nearly 95 percent of all possible means calculate from samples from the population with mean **μ** will lie within 1.96 standard deviations of the true mean **μ**,. ($1.96\hat{\sigma}_{\bar{Y}}$ determines the "size" of the ring.)

F. The inequality can be restated by

$$\boxed{\bar{Y} - 1.96\hat{\sigma}_{\bar{Y}} \leq \mu \leq \bar{Y} + 1.96\hat{\sigma}_{\bar{Y}}}$$

1. That is, over all possible samples the probability is about .95 that the range between $\bar{Y} - 1.96\hat{\sigma}_{\bar{Y}}$ and $\bar{Y} + 1.96\hat{\sigma}_{\bar{Y}}$ will include the true mean. Stated more, crudely if we took, say, 10,000 independent samples of size N from a population with mean μ and standard deviation **σ**, about 95% of the calculated intervals would include **μ** somewhere between the upper and lower limits of the interval.

G. 99 percent confidence intervals are given by

$$\bar{Y} - 2.58\hat{\sigma}_{\bar{Y}} \le \mu \le \bar{Y} + 2.58\hat{\sigma}_{\bar{Y}}$$

H. In general, (1 - α)% confidence intervals for μ can be constructed by

$$\bar{Y} \pm z\hat{\sigma}_{\bar{Y}} = \bar{Y} \pm z\left(\frac{\hat{\sigma}}{\sqrt{N}}\right)$$

1. z is selected from the standard normal table in such a fashion that we can assert with probability of (1 - α) that the random variable $(\bar{Y} - \mu)/\hat{\sigma}_{\bar{Y}})$ will lie between -z and +z.

I. Given random samples from a population having a parameter θ and assuming certain conditions are met the general form of (1 - a) percent confidence intervals is:

$$(\hat{\ } - \hat{\sigma}) \le \quad \le (\hat{\ } + \hat{\sigma})$$

1. where is θ the population parameter, $\hat{\ }$ is an estimator of that parameter, $\hat{\sigma}$ is the (estimated) standard error of the appropriate distribution, and ζ, which depends on the desired degree or level of confidence, is chosen so as to define or mark off (1 - α) percent of the appropriate sampling distribution.

III. MINITAB:

A. Although the hand calculations for this problem are trivial, let's use MINITAB to construct confidence intervals for our sample mean of infant mortality.

1. Later we'll do them by hand.

B. Go to **Stat** and **Basic statistics**.

C. Choose **1-sample t**.

D. Make sure the column containing the sample data is selected and check the confidence interval box.

1. The default is 95 percent, so click **Ok**.

E. The results are:

```
                         Confidence Intervals
Variable      N      Mean     StDev  SE Mean        95.0 % CI
Sample        4    12.625     1.195    0.598  (   10.723,   14.527)
```

1. Interpretation: these intervals have been constructed so that we can be 95 percent confident that the true mean is between 10.723 and 14.527.

F. You can see the effect of increasing and decreasing confidence on the size of the intervals.

1. 90 percent intervals are rather narrow, but of course we can't be too confident that they contain the true value:

```
                           Confidence Intervals
Variable      N       Mean     StDev  SE Mean        90.0 % CI
Sample        4     12.625     1.195    0.598  (  11.218,  14.032)
```

i. Note the interval covers only 14.032 - 11.218 = 2.814 infant deaths.

2. 99 percent intervals are relatively wide:

```
                           Confidence Intervals
Variable      N       Mean     StDev  SE Mean        99.0 % CI
Sample        4     12.625     1.195    0.598  (   9.134,  16.116)
```

i. If we need to be 99 percent confident, then we need wide intervals.

G. There is thus a trade off: more confidence, wider intervals (less "precise" estimates); less confidence, narrower intervals.

IV. DIFFERENCE OF MEANS TEST:

A. Again this section largely reprints the notes from Class 3.

B. The problem is to test the hypothesis that two population means differ. Hence we have a population with mean = μ_1 and standard deviation = σ_1 and another population with mean = μ_2 and standard deviation = σ_2.

1. Note that we may or may not assume that the standard deviations equal one another. You will always be told which assumption holds.

C. Hypotheses:

1. Null: H_0 $\mu_1 = \mu_2 = \mu$

i. Stated equivalently: H_0: $\mu_1 - \mu_2 = 0$

2. Alternative: μ_1 not equal to μ_2; or $\mu_1 > \mu_2$; or $\mu_1 < \mu_2$

D. Notation;

1. Let's denote the **population** difference of means: $= \mu_1 - \mu_2$

2. Then the null hypothesis can be stated as

$$\boxed{H_0: \quad = \mu_1 - \mu_2 = 0}$$

E. Example: consider this thought experiment: we are interested in comparing

California and Alabama in terms of some variable such as "infant mortality rates." Suppose we take two independent samples from the population of counties in each state. (Denote the sample sizes N_{CA} and N_{AL} respectively. That is, we might draw a sample of 12 counties from California and 13 from Alabama. Then, N_{CA} = 12 and N_{AL} = 13.) After finding out the mean or average infant death rate in the two states, we then calculate:

$$\hat{\ } = \bar{Y}_{CA} - \bar{Y}_{AL}$$

F. Now suppose we repeat the process by drawing another sample of N_{CA} cases from California and N_{AL} from Alabama and calculate the difference of means on the tenn pregnancy rate. This estimated difference will probably not equal the first one because we have drawn new independent samples.

G. Let's continue in this manner, drawing samples of size 12 and 13 respectively from California and Alabama an infinite number of times. Stated differently, suppose we some how obtain sample differences of means from all possible samples from these two states. We will have a "pile" of $\hat{\ }'s$.

H. What will be the mean, standard deviation, and shape of this collection?

1. Theory tells us that if we are have small samples and unknown population standard deviations, the appropriate distribution will the t-distribution.

2. Parameters of a sampling distribution.

   i. 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,

$$E(\hat{\ }) = \quad = \mu_{CA} - \mu_{AL}$$

   ii. 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 $\sigma_{\hat{\ }}$

3. The form of the sampling distribution.

   i. 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 Δ and standard deviation (called the standard error) of $\sigma_{\hat{\ }}$.

   ii. 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_{AR} equals σ_{VA}), then the sampling distribution will be the t distribution with (in this case):

$$\boxed{df = N_{CA} + N_{AL} - 2}$$

I. Critical value and regions:
  1. For this test use the t-distribution with df indicated above.

J. Calculating the test statistic involves three of steps:
  1. We have samples from two populations having the same standard deviations; that is, assume for now $\sigma_1 = \sigma_2 = \sigma$.
  2. Use the two sample standard deviations to calculated a **pooled** estimate of the common standard deviation.
  3. Because the standard deviations are usually unknown we obtained a **pooled** estimator of the common standard deviation, σ, with

$$\boxed{\hat{\sigma} = \sqrt{\frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 + 2}}}$$

  4. This pooled estimator is used to calculate the **standard error** of the difference of means

$$\boxed{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{\frac{\hat{\sigma}^2}{N_1} + \frac{\hat{\sigma}^2}{N_2}} = \hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

  5. The standard error in turn is divided into the difference of sample means minus the hypothesized difference of population means to form the observed t statistic.

$$\boxed{t_{obs} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}}}$$

K. Decisi on:
  1. The observed t is compared with the critical: if it equals or exceeds the critical (absolute) value, reject the null hypothesis; otherwise, do not reject.

L. Interpretation:

1. Use the data, especially the estimated difference, to say something substantive.

V. LARGE SAMPLES

A. If N_1 and N_2 are both greater than or equal to 20, use the standard normal distribution and the z statistic.

1. Replace t in the above with z.
2. That is, for a two-tailed test at the .05 level use $z_{critical} = 1.96$.
3. Calculate a z statistic using the above formulas.

VI. EXAMPLE:

A. Let's compare Alabama and California's infant mortality rates (and perhaps some other variables as well).

1. It would be easy enough to obtain population parameters but instead let's assume that we are sampling from populations of infinite size.
2. Here the units are counties, so we assume that each state has an infinite number and we are sampling without replacement.
3. The variables[1] are:
   i. Infant mortality rate: deaths per 1,000 live births, 1990-1991
   ii. Violent crimes known to police per 100,000 population, 1994.
   iii. Practicing physicians per 100,000 population, 1993.
4. The sample data are (see next page):

[1]Source: ***1996 County and City Extra*** (Brenan Press).

California				Alabama			
County	Infant	Crime	Doctors	County	Infant	Crime	Doctors
Butte	8.1	5371	185	Autauga	9.9	4105	46
Colusa	3.4	3906	53	Calhoun	15.5	5582	119
Imperial	6.4	8370	62	Chambers	16.7	4141	67
Kern	10.2	6514	125	Colbert	9.0	3208	137
Mariposa	3.6	5499	97	Conecuh	17.0	1766	35
Merced	8.0	5532	106	Elmore	14.0	3599	46
Placer	5.8	4905	212	Escambia	10.9	888	72
Riverside	8.5	7291	121	Houston	8.2	3068	261
San Luis Obispo	6.9	4098	217	Jackson	10.3	2276	63
Santa Barbara	7.0	4336	246	Macon	18.8	6211	98
Sonoma	5.3	5079	241	Madison	8.8	7051	185
Sutter	5.6	6259	151	Perry	6.3	3539	32
Yuba	10.4	7108	96	Pickens	19.5	1191	67
				Walker	13.5	2553	80

B. Calculations

1. First find the sample means and standard deviations.
    i. MINITAB gives

```
Variable        N      Mean    Median  Tr Mean     StDev  SE Mean
CA-death       13     6.862     6.900    6.855     2.183    0.605
AL-death       14     12.74     12.20    12.72      4.25     1.14
```

    ii. We need the sample standard deviations in order to calculate the pooled estimate (estimated based on both samples) of the common standard deviation. Remember: we are assuming that the population standard deviations equal.

2. The pooled standard deviation is:

$$\hat{\sigma} = \sqrt{\frac{12(2.183)^2 = 13(4.25)^2}{13 + 14 - 2}}$$

$$= \sqrt{\frac{291.9984}{25}} = 3.4147 \approx 3.42$$

3. Now use this to obtain the **estimated standard error**:

$$\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \mathbf{3.4176}\sqrt{\frac{\mathbf{1}}{\mathbf{13}} + \frac{\mathbf{1}}{\mathbf{14}}}$$

$$= \mathbf{3.4176}\sqrt{\mathbf{.14835}}$$

$$= \mathbf{1.31633} \approx \mathbf{1.32}$$

4. Finally, the observed t is

$$t_{obs} = \frac{\mathbf{(6.862 - 12.74)}}{\mathbf{1.31633}}$$

$$= \mathbf{-4.46453}$$

5. The critical t with 25 degrees of freedom at the .05 level, two-tailed test (use the .025 column since .025 + .025 = .05) is 2.060.
6. The observed t exceeds this by quite a bit so reject the null hypothesis that Alabama has the same mean infant mortality rate as California.

VII. MINITAB:

A. Use **Basic statistics**, **2-sample t**.

1. Be sure to check the box indicating equal standard deviations.

B. Results for the infant mortality problem:

```
                Two sample T for CA-death vs AL-death
                        N      Mean     StDev   SE Mean
             CA-death  13      6.86      2.18      0.61
             AL-death  14     12.74      4.25       1.1

        95% CI for mu CA-death - mu AL-death: ( -8.58,  -3.2)
T-Test mu CA-death = mu AL-death (vs not =): T= -4.57  P=0.0002  DF=  19
```

1. Note that the observed t is -4.57. The probability of its occurrence under the hypothesis and assuming equal population standard deviations is .0002.
2. Note that the degrees of freedom is not the same as given by the formula. That's because the program uses (a more accurate) algorithm for calculating it.
3. Note also that the program automatically calculates confidence intervals for the difference of means.
   i. Since these 95 percent intervals do not include 0, we can as done above, reject the null hypothesis that the population means are equal.

VIII. NEXT TIME:
   A. Confidence intervals for the difference of means.
   B. Cross-tabulations

Go to Statistics page.