THE MEANING OF
STATISTICAL SIGNIFICANCE:
STANDARD ERRORS AND
CONFIDENCE INTERVALS
LOGISTICS

- Homework #3 will be due in class on Wednesday, May 28 (not May 21)

- Note: Monday, May 26 is a holiday
OUTLINE

1. Issues in Sampling (review)
2. Statistics for Regression Analysis
3. Central limit theorem
4. Distributions: Population, Sample, Sampling
5. Using the Normal Distribution
6. Establishing Confidence Intervals
Parameters and Statistics

A parameter is a number that describes the population. It is a fixed number, though we do not know its value.

A statistic is a number that describes a sample. We use statistics to estimate unknown parameters.

A goal of statistics: To estimate the probability that the sample statistic (or observed relationship) provides an accurate estimate for the population. Forms:

(a) Placing a confidence band that around a sample statistic, or
(b) Rejecting (or accepting) the null hypothesis on the basis of a satisfactory probability.
Problems in Sampling

<table>
<thead>
<tr>
<th>$H_o$ for Population</th>
<th>$H_o$ for Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepted</td>
<td>Rejected</td>
</tr>
</tbody>
</table>

$H_o = \text{null hypothesis}$
Population parameter = Sample statistic + Random sampling error

Random sampling error = (Variation component)/(Sample size component)

Sample size component = $1/ \sqrt{n}$

Random sampling error = $\sigma / \sqrt{n}$

where $\sigma$ = standard deviation in the population
SIGNIFICANCE MEASURES FOR REGRESSION ANALYSIS

1. Testing the null hypothesis:

\[ F = \frac{r^2(n-2)}{(1-r^2)} \]

2. Standard errors and confidence intervals:

Dependent on desired significance level

Bands around the regression line

95% confidence interval ±1.96 x SE
Central limit theorem:

If the N of each sample drawn is large, regardless of the shape of the population distribution, the sample means will (a) tend to distribute themselves normally around the population mean (b) with a standard error that will be inversely proportional to the square root of N.

Thus: the larger the N, the smaller the standard error (or variability of the sample statistics)
On Distributions:

1. Population (from which sample taken)
2. Sample (as drawn)
3. Sampling (of repeated samples)
**Population Distribution**

\[ \mu \]

**Sample Distribution**

\[ \bar{x} (\neq \mu) \]

**Sampling Distribution**

\[ \text{Standard Error} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \]
Figure 5-1 The Normal Distribution
Characteristics of the “Normal” Distribution

• Symmetrical

• Unimodal

• Bell-shaped

• Mode = mean = median

• Skewness = 0 = \[
\frac{3(\bar{X} - \text{md})}{s} = \frac{\bar{X} - \text{Mo}}{s}
\]

• Described by mean (center) and standard deviation (shape)

• Neither too flat (platykurtic) nor too peaked (leptokurtic)
Areas under the Normal Curve

**Key property:** known area (proportion of cases) at any given distance from the mean expressed in terms of standard deviation Units (AKA Z scores, or standard scores)

- 68% of observations fall within ± one standard deviation from the mean
- 95% of observations fall within ± two standard deviations from the mean (actually, ± 1.96 standard deviations)
- 99.7% of observations fall within ± three standard deviations from the mean
Putting This Insight to Use

Knowledge of a mean and standard deviation enables computation of a Z score, which $= (X_i - \bar{X})/s$

Knowledge of a Z scores enables a statement about the probability of an occurrence (i.e., $Z > \pm 1.96$ will occur only 5% of the time)
Random sampling error = standard error

Refers to how closely an observed sample statistic approximates the population parameter; in effect, it is a standard deviation for the sampling distribution

Since $\sigma$ is unknown, we use $s$ as an approximation, so

$$\text{Standard error} = \frac{s}{\sqrt{n}} = SE$$
Establishing boundaries at the 95 percent confidence interval:

Lower boundary = sample mean – 1.96 SE

Upper boundary = sample mean + 1.96 SE

Note: This applies to statistics other than means (e.g., percentages or regression coefficients).

Conclusion: 95 percent of all possible random samples of given size will yield sample means between the lower and upper boundaries
## Postscript: Confidence Intervals (for %)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.20</td>
</tr>
<tr>
<td>2000</td>
<td>±1.4</td>
</tr>
<tr>
<td>1000</td>
<td>2.0</td>
</tr>
<tr>
<td>500</td>
<td>2.9</td>
</tr>
<tr>
<td>50</td>
<td>9.1</td>
</tr>
</tbody>
</table>