Chapter 8 - Large Sample Estimation

Points covered in this chapter.

1. Point/interval estimators.
2. Properties of sound estimators.
3. Typology of the research questions that use the statistical tools in Chapters 8 through 10.
4. Methods to estimate population parameters (e.g., to estimate the mean and variance of the population (for a normal distribution) and the proportion (for a binomial distribution) when these values are unknown).
5. Calculation of the margin of error and confidence intervals.
6. How to choose a sample size.

Types of estimators

1. **Point Estimator** - what is the best single value that can be used to estimate a population parameter (mean or proportion).

2. **Interval Estimator** - what is the best interval (refer to as a confidence interval) that contains the population estimate.

Properties of point estimators
1. **Unbiased** - average values of the estimated parameter equals the population parameter.

2. **Consistent** - Estimators from sample converge to the true value as the sample size increases.

3. **Efficient** - Estimator with smallest sampling variance.

**Measuring the error of the estimation - margin of error**
Way to describe degree of biasness.

1. General equation

   Margin of error → ±1.96*Standard error of the estimator

2. Standard error for \( \bar{x} \) as a point estimator.

   \[
   SE = \frac{\sigma}{\sqrt{n}}
   \]

   if \( \sigma \) is unknown and \( n \) is 30 or larger, the sample deviation value can be used for \( \sigma \).

3. Standard error for \( \hat{p} \) as a point estimator.

   \[
   SE = \sqrt{\frac{pq}{n}}
   \]

   **Assumption** np > 5 and nq > 5. p and q are replaced by \( \hat{p} \) and \( \hat{q} \)
Examples

Typology of the research questions

1. Testing the choice of a single parameter of interest.

2. Examining two parameters of interest and determining if they are truly different.

   There are two kinds of data sets used for bivariate analysis

   (a) Data sets are not paired (or data sets are independent from one another). There is no relationship between the two parameters differed (e.g., MCAT scores for Biochemistry and Biology Majors).

   (b) Data sets are paired (e.g., there is a relationship between the two data sets). Examples, comparing the differences in gas mileage when a car is first given one type and then another type of gasoline, test scores of trainees before and after viewing an instructional video.

Interval Estimation

Depending upon the research question, we are interested in the confidence interval around the value of interest (e.g., sample mean, proportion). In
some instances, we are interested in understanding the lower and upper bounds or limits. For some research questions, we are only interested in understanding one of the boundaries.

**General function - one tail test**

Point estimator $- z_\alpha \times$ Standard Error $-$ Left tail

Point estimator $+ z_\alpha \times$ Standard Error $-$ Right tail

$z_\alpha$ ?

1. A $z$ score, in this case, a $z$ score measuring the interval to the left or to the right of the mean. The probability values of either tail is $\alpha$.

**General function - two tail test**

Point estimator $\pm z_{\alpha/2} \times$ Standard Error

$z_{\alpha/2}$ ?

(a) A $z$ score, in this case, a $z$ score measuring the interval around the mean. The probability values of each tail is $\alpha/2$. The $z$ score is a percentage measurement of the interval around the mean (just as the Empirical Rule and Tchebysheff’s Theorem).
<table>
<thead>
<tr>
<th>Confidence</th>
<th>$\alpha$</th>
<th>$\frac{\alpha}{2}$</th>
<th>$z_{\frac{\alpha}{2}}$</th>
<th>$z_{\alpha}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.0%</td>
<td>0.010</td>
<td>0.0050</td>
<td>2.58</td>
<td>2.33</td>
</tr>
<tr>
<td>98.0%</td>
<td>0.020</td>
<td>0.0100</td>
<td>2.33</td>
<td>2.055</td>
</tr>
<tr>
<td>97.5%</td>
<td>0.025</td>
<td>0.0125</td>
<td>2.24</td>
<td>1.96</td>
</tr>
<tr>
<td>95.0%</td>
<td>0.050</td>
<td>0.0250</td>
<td>1.96</td>
<td>1.645</td>
</tr>
<tr>
<td>90.0%</td>
<td>0.100</td>
<td>0.0500</td>
<td>1.645</td>
<td>1.28</td>
</tr>
</tbody>
</table>

i. Population mean

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

When $n > 30$

$$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

ii. Population proportion

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Examples

Estimating the difference between two means

<table>
<thead>
<tr>
<th>Population 1</th>
<th>Population 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu_1$</td>
</tr>
<tr>
<td>Variance</td>
<td>$\sigma_1^2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\bar{x}_1$</td>
</tr>
<tr>
<td>Variance</td>
<td>$s_1^2$</td>
</tr>
<tr>
<td>Sample size</td>
<td>$n_1$</td>
</tr>
</tbody>
</table>

5
Properties of Sampling distribution of 
\( (\bar{x}_1 - \bar{x}_2) \) - not paired

(a) Mean and Standard error

\[
\mu(\bar{x}_1-\bar{x}_2) = \mu_1 - \mu_2
\]

\[
SE = \sigma(\bar{x}_1-\bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

(b) Margin of Error

\[
\pm 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

(c) Confidence interval (two-tail)

\[
(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

(a) If sampled populations are normally 
distributed, then the sampling distribution 
of \((\bar{x}_1 - \bar{x}_2)\) is normally distributed 
regardless of size.

(b) If the sampled populations are not normally 
distributed, then the sampling distribution 
of \((\bar{x}_1 - \bar{x}_2)\) is approximately normally 
distributed when \(n_1\) and \(n_2\) are large, due 
to the Central Limit Theorem.
(c) If $\sigma_1^2$ and $\sigma_2^2$ are unknown, but both $n_1$ and $n_2$ are greater than or equal to 30, you can substitute the sample variances for the population variances.

<table>
<thead>
<tr>
<th>Population 1</th>
<th>Population 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>$p_1$</td>
</tr>
<tr>
<td>Sample 1</td>
<td>$\hat{p}_1$</td>
</tr>
<tr>
<td>Sample size</td>
<td>$n_1$</td>
</tr>
</tbody>
</table>

Properties of Sampling distribution of $(\hat{p}_1 - \hat{p}_2)$-not paired

(a) Mean and Standard Error

\[
(\hat{p}_1 - \hat{p}_2) = \left( \frac{x_1}{n_1} - \frac{x_2}{n_2} \right)
\]

\[
\mu(\hat{p}_1 - \hat{p}_2) = (p_1 - p_2)
\]

\[
SE = \sigma(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}
\]

(b) Margin of Error

\[
\pm 1.96 \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}
\]

(c) Confidence Interval (two-tail)

\[
(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}
\]
(a) The sampling distribution of \((\hat{p}_1 - \hat{p}_2)\) is approximately normally distributed when \(n_1\) and \(n_2\) are large, due to the Central Limit Theorem.

(b) \(n_1\) and \(n_2\) must be sufficiently large so that the sampling distribution of \((\hat{p}_1 - \hat{p}_2)\) can be approximated by a normal distribution. \(n_1p_1, n_1q_1, n_2p_2\) and \(n_2q_2 > 5\).

**Properties of Sampling distribution of \((\bar{x}_1 - \bar{x}_2)\) - paired**

Mean and Standard error

\[
\mu(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2
\]

\[
SE = \sigma(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_D^2}{n}}
\]

where \(\sigma_D^2\) is the variance of the differenced data and \(n_1 = n_2 = n\).

Margin of Error

\[
\pm 1.96 \sqrt{\frac{\sigma_D^2}{n}}
\]

Confidence interval (two-tail)

\[
(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_D^2}{n}}
\]
Properties of Sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ - paired
Will not be covered in this class.

Sample Size
Choosing a sample size is an application of the point and interval estimation techniques.
Suppose you want to generate a sample such that the margin of error is equal to some value, let’s call it B. You also want a sample such that 95% of repeated sampling will given you a margin of error less than or equal to B.

For univariate and bivariate analyses, each margin of error function is a function of n. Here is the case for the population mean, univariate case.

\[ B \pm 1.96 \left( \frac{\sigma}{\sqrt{n}} \right) \]

If I rearrange the function above, one finds the function for computing the sample size.

\[ n \geq \left( \frac{1.96}{B} \right)^2 \sigma^2 \]

If \( \sigma \) is not known, the sample standard deviation can be used or a value based on the range of the values divided by 4.
<table>
<thead>
<tr>
<th>Analysis</th>
<th>Estimator</th>
<th>Minimum sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate</td>
<td>$\bar{x}$</td>
<td>$n \geq \frac{z_{\alpha/2}^2 \sigma^2}{B^2}$</td>
</tr>
<tr>
<td></td>
<td>$\hat{p}$</td>
<td>$n \geq \frac{z_{\alpha/2}^2 pq}{B^2}$</td>
</tr>
<tr>
<td>Bivariate - not paired</td>
<td>$\bar{x}_1 - \bar{x}_2$</td>
<td>$n \geq \frac{z_{\alpha/2}^2 (\sigma_1^2 + \sigma_2^2)}{B^2}$</td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_1 - \hat{p}_2$</td>
<td>$n \geq \frac{z_{\alpha/2}^2 (p_1 q_1 + p_2 q_2)}{B^2}$</td>
</tr>
</tbody>
</table>

For the Bivariate functions $n_1 = n_2 = n$. 