Introduction

• Populations are described by their probability distributions and parameters.
  – For quantitative populations, the location and shape are described by \( \mu \) and \( \sigma \).
  – For a binomial populations, the location and shape are determined by \( p \).
• If the values of parameters are unknown, we make inferences about them using sample information.

Types of Inference

• Estimation:
  – Estimating or predicting the value of the parameter
  – “What is (are) the most likely values of \( \mu \) or \( p \)?”
• Hypothesis Testing:
  – Deciding about the value of a parameter based on some preconceived idea.
  – “Did the sample come from a population with \( \mu = 5 \) or \( p = .2 \)?”

Types of Inference

• Examples:
  – A consumer wants to estimate the average price of similar homes in her city before putting her home on the market.
    \[\text{Estimation: Estimate } \mu, \text{ the average home price.}\]
  – A manufacturer wants to know if a new type of steel is more resistant to high temperatures than an old type was.
    \[\text{Hypothesis test: Is the new average resistance, } \mu_N, \text{ equal to the old average resistance, } \mu_O?\]
Definitions

- An estimator is a rule, usually a formula, that tells you how to calculate the estimate based on the sample.
  - **Point estimation**: A single number is calculated to estimate the parameter.
  - **Interval estimation**: Two numbers are calculated to create an interval within which the parameter is expected to lie.

Properties of Point Estimators

- Since an estimator is calculated from sample values, it varies from sample to sample according to its **sampling distribution**.
- An estimator is unbiased if the mean of its sampling distribution equals the parameter of interest.
  - It does not systematically overestimate or underestimate the target parameter.

Properties of Point Estimators

- Of all the unbiased estimators, we prefer the estimator whose sampling distribution has the **smallest spread** or variability.

Measuring the Goodness of an Estimator

- The distance between an estimate and the true value of the parameter is the **error of estimation**.
- In this chapter, the sample sizes are large, so that our unbiased estimators will have normal distributions. Because of the Central Limit Theorem.

The Margin of Error

- For unbiased estimators with normal sampling distributions, 95% of all point estimates will lie within 1.96 standard deviations of the parameter of interest.
- **Margin of error**: The maximum error of estimation, calculated as

\[ \text{Margin of error} = 1.96 \times \text{std error of the estimator} \]

Estimating Means and Proportions

- For a quantitative population,
  \[ \text{Point estimator of population mean } \mu : \hat{x} \]
  \[ \text{Margin of error} (n \geq 30) : \pm 1.96 \frac{s}{\sqrt{n}} \]
- For a binomial population,
  \[ \text{Point estimator of population proportion } p : \hat{p} = x/n \]
  \[ \text{Margin of error} (n \geq 30) : \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]
Example

A homeowner randomly samples 64 homes similar to her own and finds that the average selling price is $252,000 with a standard deviation of $15,000. Estimate the average selling price for all similar homes in the city.

Point estimator of $\mu = 252,000$
Margin of error: $\pm 1.96 \frac{s}{\sqrt{n}} = \pm 1.96 \frac{15,000}{\sqrt{64}} = \pm 3675$

Example

A quality control technician wants to estimate the proportion of soda cans that are underfilled. He randomly samples 200 cans of soda and finds 10 underfilled cans.

Estimator $\hat{p} = \frac{10}{200} = .05$
Margin of error: $\pm 1.96 \sqrt{\frac{.05(.95)}{200}} = \pm .03$

Interval Estimation

Create an interval $(a, b)$ so that you are fairly sure that the parameter lies between these two values.

“Fairly sure” means “with high probability”, measured using the confidence coefficient, $1 - \alpha$.

Usually, $1 - \alpha = .90, .95, .98, .99$

- Suppose $1 - \alpha = .95$ and that the estimator has a normal distribution.

To Change the Confidence Level

To change to a general confidence level, $1 - \alpha$, pick a value of $z$ that puts area $1 - \alpha$ in the center of the $z$ distribution.

<table>
<thead>
<tr>
<th>$z \alpha$</th>
<th>Tail area</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>.1</td>
<td>1.645</td>
</tr>
<tr>
<td>.025</td>
<td>.025</td>
<td>1.96</td>
</tr>
<tr>
<td>.01</td>
<td>.01</td>
<td>2.33</td>
</tr>
<tr>
<td>.005</td>
<td>.005</td>
<td>2.58</td>
</tr>
</tbody>
</table>

100(1-\alpha)\% Confidence Interval: Estimator $\pm z_{\alpha/2}SE$

Confidence Intervals for Means and Proportions

- For a quantitative population,
  
  $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

- For a binomial population,
  
  $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Example

• A random sample of \( n = 50 \) males showed a mean average daily intake of dairy products equal to 756 grams with a standard deviation of 35 grams. Find a 95% confidence interval for the population average \( \mu \).

\[
\bar{x} \pm 1.96 \frac{s}{\sqrt{n}} = 756 \pm 1.96 \frac{35}{\sqrt{50}} = 756 \pm 9.70
\]

or \( 746.30 < \mu < 765.70 \) grams.

Example

• Find a 99% confidence interval for \( \mu \), the population average daily intake of dairy products for men.

\[
\bar{x} \pm 2.58 \frac{s}{\sqrt{n}} = 756 \pm 2.58 \frac{35}{\sqrt{50}} = 756 \pm 12.70
\]

or \( 743.23 < \mu < 768.77 \) grams.

The interval must be wider to provide for the increased confidence that is does indeed enclose the true value of \( \mu \).

Example

• Of a random sample of \( n = 150 \) college students, 104 of the students said that they had played on a soccer team during their K-12 years. Estimate the proportion of college students who played soccer in their youth with a 98% confidence interval.

\[
\hat{p} \pm 2.33 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.69 \pm 2.33 \sqrt{\frac{0.69(1-0.69)}{150}}
\]

\( 0.60 < \hat{p} < 0.78 \).

Estimating the Difference between Two Means

• Sometimes we are interested in comparing the means of two populations.
  • The average growth of plants fed using two different nutrients.
  • The average scores for students taught with two different teaching methods.

• To make this comparison, we compare the two averages by making inferences about \( \mu_1 - \mu_2 \), the difference in the two population averages.

\[
\bar{x}_1 - \bar{x}_2
\]

The best estimate of \( \mu_1 - \mu_2 \) is the difference in the two sample means, \( \bar{x}_1 - \bar{x}_2 \).

The Sampling Distribution of \( \bar{x}_1 - \bar{x}_2 \)

1. The mean of \( \bar{x}_1 - \bar{x}_2 \) is \( \mu_1 - \mu_2 \), the difference in the population means.

2. The standard deviation of \( \bar{x}_1 - \bar{x}_2 \) is \( SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \).

3. If the sample sizes are large, the sampling distribution of \( \bar{x}_1 - \bar{x}_2 \) is approximately normal, and \( SE \) can be estimated as \( SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \).
Estimating $\mu_1 - \mu_2$

- For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal ($z$) distribution.

Point estimate for $\mu_1 - \mu_2$: $\hat{x}_1 - \hat{x}_2$

Margin of Error: $1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Confidence interval for $\mu_1 - \mu_2$:

$$(\hat{x}_1 - \hat{x}_2) \pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example

<table>
<thead>
<tr>
<th>Avg Daily Intakes</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Sample mean</td>
<td>756</td>
<td>762</td>
</tr>
<tr>
<td>Sample Std Dev</td>
<td>35</td>
<td>30</td>
</tr>
</tbody>
</table>

- Compare the average daily intake of dairy products of men and women using a 95% confidence interval.

$$(\hat{x}_1 - \hat{x}_2) \pm 1.96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example, continued

$-18.78 < \mu_1 - \mu_2 < 6.78$

- Could you conclude, based on this confidence interval, that there is a difference in the average daily intake of dairy products for men and women?
- The confidence interval contains the value $\mu_1 - \mu_2 = 0$. Therefore, it is possible that $\mu_1 = \mu_2$. You would not want to conclude that there is a difference in average daily intake of dairy products for men and women.

Estimating the Difference between Two Proportions

- Sometimes we are interested in comparing the proportion of "successes" in two binomial populations.
  - The germination rates of untreated seeds and seeds treated with a fungicide.
  - The proportion of male and female voters who favor a particular candidate for governor.

A random sample of size $n_1$ drawn from binomial population 1 and $n_2$ drawn from binomial population 2 with parameter $\theta_1$.

Estimating the Difference between Two Means

- We compare the two proportions by making inferences about $\theta_1 - \theta_2$, the difference in the two population proportions.
  - If the two population proportions are the same, then $\theta_1 - \theta_2 = 0$.
  - The best estimate of $\theta_1 - \theta_2$ is the difference in the two sample proportions,

$$\hat{\theta}_1 - \hat{\theta}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

The Sampling Distribution of $\hat{\theta}_1 - \hat{\theta}_2$

1. The mean of $\hat{\theta}_1 - \hat{\theta}_2$ is $\theta_1 - \theta_2$, the difference in the population proportions.
2. The standard deviation of $\hat{\theta}_1 - \hat{\theta}_2$ is $SE = \sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}}$.
3. If the sample sizes are large, the sampling distribution of $\hat{\theta}_1 - \hat{\theta}_2$ is approximately normal, and SE can be estimated as $SE = \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$. 

Copyright ©2006 Brooks/Cole
A division of Thomson Learning, Inc.
Estimating $p_1 - p_2$

For large samples, point estimates and their margin of error as well as confidence intervals are based on the standard normal ($z$) distribution.

Point estimate for $p_1 - p_2$: $\hat{p}_1 - \hat{p}_2$

Margin of Error: $\pm 1.96 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

Confidence interval for $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example

<table>
<thead>
<tr>
<th>Youth Soccer</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>Played soccer</td>
<td>65</td>
<td>39</td>
</tr>
</tbody>
</table>

Compare the proportion of male and female college students who said that they had played on a soccer team during their K-12 years using a 99% confidence interval.

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm 1.96 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example, continued

$.06 < \hat{p}_1 - \hat{p}_2 < .44$

Could you conclude, based on this confidence interval, that there is a difference in the proportion of male and female college students who said that they had played on a soccer team during their K-12 years?

The confidence interval does not contain the value $p_1 - p_2 = 0$. Therefore, it is not likely that $p_1 = p_2$. You would conclude that there is a difference in the proportions for males and females.

One Sided Confidence Bounds

- Confidence intervals are by their nature two-sided since they produce upper and lower bounds for the parameter.
- One-sided bounds can be constructed simply by using a value of $z$ that puts $\alpha$ rather than $\alpha/2$ in the tail of the $z$ distribution.

Choosing the Sample Size

The total amount of relevant information in a sample is controlled by two factors:
- The sampling plan or experimental design: the procedure for collecting the information
- The sample size $n$: the amount of information you collect.

In a statistical estimation problem, the accuracy of the estimation is measured by the margin of error or the width of the confidence interval.
Example
A producer of PVC pipe wants to survey wholesalers who buy his product in order to estimate the proportion who plan to increase their purchases next year. What sample size is required if he wants his estimate to be within .04 of the actual proportion with probability equal to .95?

\[1.96 \frac{\hat{p} \hat{q}}{n} \leq .04 \Rightarrow 1.96 \frac{.5(1-.5)}{n} \leq .04\]
\[\Rightarrow \frac{1.96 \sqrt{.5(1-.5)}}{.04} = 24.5 \Rightarrow n \geq 24.51^2 = 600.25\]

He should survey at least 601 wholesalers.

Key Concepts
I. Types of Estimators
1. Point estimator: a single number is calculated to estimate the population parameter.
2. Interval estimator: two numbers are calculated to form an interval that contains the parameter.

II. Properties of Good Estimators
1. Unbiased: the average value of the estimator equals the parameter to be estimated.
2. Minimum variance: of all the unbiased estimators, the best estimator has a sampling distribution with the smallest standard error.
3. The margin of error measures the maximum distance between the estimator and the true value of the parameter.

III. Large-Sample Point Estimators
To estimate one of four population parameters when the sample sizes are large, use the following point estimators with the appropriate margins of error.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Point Estimator</th>
<th>Margin of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu)</td>
<td>(\bar{x})</td>
<td>(\pm 1.96 \frac{\sigma}{\sqrt{n}})</td>
</tr>
<tr>
<td>(p)</td>
<td>(\hat{p})</td>
<td>(\pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})</td>
</tr>
<tr>
<td>(\mu_1 - \mu_2)</td>
<td>(\bar{x}_1 - \bar{x}_2)</td>
<td>(\pm 1.96 \sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}})</td>
</tr>
<tr>
<td>(p_1 - p_2)</td>
<td>(\hat{p}_1 - \hat{p}_2)</td>
<td>(\pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}})</td>
</tr>
</tbody>
</table>

IV. Large-Sample Interval Estimators
To estimate one of four population parameters when the sample sizes are large, use the following interval estimators.

\[\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \left(\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)\]

\[\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \leq p \leq \left(\hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\]

\[\left(\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}\right) \leq \mu_1 - \mu_2 \leq \left(\bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}\right)\]

\[\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}\right) \leq p_1 - p_2 \leq \left(\hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}\right)\]

V. One-Sided Confidence Bounds
Use either the upper (+) or lower (−) two-sided bound, with the critical value of \(z\) changed from \(z_{\alpha/2}\) to \(z_{\alpha'}\).