Learning Objectives for Ch. 7

- Obtaining a point estimate of a population parameter
- Desirable properties of a point estimator:
  - Unbiasedness
  - Efficiency
- Obtaining a confidence interval for a mean when population standard deviation is known
- Obtaining a confidence interval for a mean when population standard deviation is unknown
- Obtaining a confidence interval for a proportion
- Determining the sample size required to estimate a mean
- Determining the sample size required to estimate a proportion
- Specifying the underlying assumptions for confidence interval estimation
Section 7.1
Point Estimation

7.1 Point Estimation

**Point Estimation**

- **Concept**: Use the sample data to come up with a single number as an approximate value of the population parameter.

- Examples of population parameters: \( \mu \), \( \sigma \), \( \pi \).

- Population parameters are usually **unknown**.

- Population parameters can be estimated by a statistic.

**Rule of thumb for estimating population parameters:**

**Use the sample counterpart**

<table>
<thead>
<tr>
<th>Population Parameter</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>( \bar{Y} )</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>( S^2 )</td>
</tr>
<tr>
<td>( \pi )</td>
<td>( \hat{\pi} )</td>
</tr>
</tbody>
</table>

- An **estimate** is the specific value obtained from the data.
7.1 Point Estimation

• Desirable properties of estimators
  • Unbiasedness

\[ E(\text{Estimator}) = \text{Parameter} \]

Long-run average

Example: \[ E(\bar{Y}) = \mu \]
\[ \Rightarrow \] Sample mean is an unbiased estimator of the population mean.

Possible values of \( \bar{Y} \) are centered around \( \mu \)

\[ \mu = ? \]

• The long-run average of all possible values of \( \bar{Y} \) equals \( \mu \).

Example: \[ E(S^2) = \sigma^2 \]
\[ \Rightarrow \] Sample variance is an unbiased estimator of the population variance.

Possible values of \( S^2 \) are centered around \( \sigma^2 \)

\[ \sigma^2 = ? \]

• The long-run average of all possible values of \( S^2 \) equals \( \sigma^2 \)
  • If a divisor of \( n \) was used to calculate \( S^2 \), then \( E(S^2) = \sigma^2 \)
7.1 Point Estimation

- **Efficiency**: $V(\text{Estimator})$ is smallest of all possible unbiased estimators.

  **Example:**
  
  $V(\bar{Y}) = \sigma^2/n$ for a random sample from any population.

  Is $\bar{Y}$ the most efficient estimator of $\mu$? **It depends!**

- "The sample mean is not always most efficient when the population distribution is not normal. In particular, when the population distribution has heavy tails, the sample mean is less efficient than a trimmed mean (though it still is unbiased). Heavy-tailed distributions tend to yield lots of extreme, "oddball" values that influence a mean more than a trimmed mean." (Hildebrand, Ott, and Gray)

7.2 Interval Estimation of a Mean, Known Standard Deviation

- A confidence interval is a range of probable values for a parameter.

- A confidence interval has a **confidence level**.
  - Typical confidence levels: .95 or .99 or .90.
  - In general, the confidence level is $1 - \alpha$. 
7.2 Interval Estimation of a Mean, Known Standard Deviation

- Procedure for constructing a C.I. for \( \mu (\sigma \text{ known}) \)
  - Start with a random sample from a normal distribution.
  1. Estimator for \( \mu \) is \( \bar{Y} \).
  2. \( \bar{Y} \) is normally distributed with mean \( \mu \) and standard error \( \sigma = \sigma / \sqrt{n} \).
  3. Standardize the sample mean:
     \[
     Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}
     \]
  4. Specify the confidence level, say 95%.

5. From Table 3, \( .95 = P[-1.96 < Z < +1.96] \)
6. Translate this probability statement about Z into a probability statement about the sample mean.
   \[
   .95 = P[-1.96 < \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} < +1.96]
   \]
7. Rearrange the quantity in brackets so that \( \mu \) is isolated:
   \[
   .95 = P\left[ \frac{\bar{Y} - 1.96(\sigma / \sqrt{n})}{\sigma / \sqrt{n}} < \mu < \frac{\bar{Y} + 1.96(\sigma / \sqrt{n})}{\sigma / \sqrt{n}} \right]
   \]

- Confidence interval end points:
  Lower end point: \( \bar{Y} - (1.96)(\sigma / \sqrt{n}) \)
  Upper end point: \( \bar{Y} + (1.96)(\sigma / \sqrt{n}) \)
- Shorthand expression for a 95% confidence interval for \( \mu \):
  \[
  \bar{Y} \pm 1.96 \left( \frac{\sigma}{\sqrt{n}} \right)
  \]
7.2 Interval Estimation of a Mean, Known Standard Deviation

• Percentiles of the Z-distribution
  95% ⇒ z_{0.025} = 1.96
  99% ⇒ z_{0.005} = 2.575

• General expression for a 100(1-α)% confidence interval:
  \[ \bar{Y} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right) \]

• Assumptions necessary to use this confidence interval:
  Random sample from a normal distribution.

Exercise 7.9:

The data from Exercise 7.1, specifying how much a sample of 20 executives paid in federal income taxes, as a percentage of gross income, are reproduced below.

16.0 18.1 18.6 20.2 21.7
22.4 22.4 23.1 23.2 23.5
24.1 24.3 24.7 25.2 25.9
26.3 27.9 28.0 30.4 33.7
[\bar{Y} = 23.985]

Assume that the standard deviation for the underlying population is 4.0.

a. Calculate a 95% confidence interval for the population mean.

\[ n = 20, \quad \sigma = 4\%, \quad z_{0.025} = 1.96 \]

\[ \bar{Y} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right) \]

23.985 ± 1.96 (4.0/\sqrt{20}) ⇒ [22.23, 25.74]

This is a 95% confidence interval for the average income tax paid by all executives.
### 7.2 Interval Estimation of a Mean, Known Standard Deviation

**Minitab output for part (a) of Exercise 7.9**

**One-Sample Z: Tax(%)**

The assumed standard deviation = 4

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tax(%)</td>
<td>20</td>
<td>23.9850</td>
<td>4.1783</td>
<td>0.8944</td>
<td>(22.2320, 25.7380)</td>
</tr>
</tbody>
</table>

**b. Calculate a 99% confidence interval for the population mean.**

\[ 99\% \Rightarrow Z_{0.005} = Z_{0.995} = 2.575 \]

\[ 23.985 \pm 2.575 \left( \frac{4}{\sqrt{20}} \right) \]

\[ 21.68, 26.29 \]

This is a 99% confidence interval for the average income tax paid by all executives.

**Minitab output for part (b) of Exercise 7.9**

**One-Sample Z: Tax(%)**

The assumed standard deviation = 4

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>99% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tax(%)</td>
<td>20</td>
<td>23.9850</td>
<td>4.1783</td>
<td>0.8944</td>
<td>(21.6811, 26.2889)</td>
</tr>
</tbody>
</table>
7.2 Interval Estimation of a Mean, Known Standard Deviation

7.10: Give a careful verbal interpretation of the confidence interval in part (a) of Exercise 7.9.

- 22.23 to 25.74
- Sample No. 1
- Sample No. 2
- Sample No. 3

- True μ = ?
- Real number line

- 95% of the CI’s you could construct would contain μ and 5% would not.
- Does the confidence interval [22.23, 25.74] contain μ? We don’t know.

7.2 Interval Estimation of a Mean, Known Standard Deviation

7.11: From the appearance of the data in Exercise 7.9, is it reasonable to assume that the sampling distribution of the mean is nearly normal?

Rephrased:

Is the distribution of Y nearly normal?

Answer:

If the data came from a population where Y (the percentage of federal income taxes paid) is normally distributed, then Y is normally distributed for any sample size.

Is it reasonable to conclude that the data came from a normal distribution? Refer to the NPP.

Since the NPP is linear, it is reasonable to conclude that the data came from a normal distribution.

The content of the box in the upper right-hand corner of the NPP will be explained in Chapter 8.
7.2 Interval Estimation of a Mean, Known Standard Deviation

What if the distribution of Y is non-normal?

Answer:
Regardless of the nature of the population distribution, the sampling distribution of \( \bar{Y} \) is nearly normal as long as the sample size is large enough because of the Central Limit Theorem.

Is \( n = 20 \) large enough?
Unless the distribution of the population is markedly non-normal, a sample of size 20 should be large enough for the CLT to apply.

• Procedure to obtain Z-interval using Minitab:
  ➔ Click on Stat ➔ Basic Statistics ➔ 1-Sample Z
  ➔ In "Samples in Column" box, enter column where data is stored
  ➔ In "Standard deviation" box, enter 4.0
  ➔ Click on "Options" and enter "95.0" in "Confidence Level" box
  ➔ Click on "OK"
7.3 Confidence Intervals for a Proportion

- Preliminary concepts
  - For a binomial random variable:
    \[ E(Y) = n\pi \text{ and } V(Y) = n\pi(1 - \pi) \]
  - \( Y \) is the total number of successes in \( n \) trials.
  - A binomial random variable can be approximated by a normal random variable because of the Central Limit Theorem.
  - The sample proportion, denoted by \( \hat{\pi} \), is
    \[ \hat{\pi} = \frac{Y}{n} \]

- \( \hat{\pi} \) is approximately normally distributed.
- \[ \frac{\hat{\pi} - \pi}{\sqrt{\pi(1 - \pi)/n}} \] is approximately a standard normal.
- By pivoting on the above expression and simplifying, a 100(1 - \( \alpha \))% C.I. for \( \pi \) is obtained:
  \[ \hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \]
- This expression is based on the premise that a binomial random variable can be approximated by a normal random variable.
7.3 Confidence Intervals for a Proportion

- Conditions for validity of the normal approximation to the binomial:
  \[ n \hat{p} - 5 \geq 0 \text{ and } n \hat{p} + 5 \leq n \]

- To use the confidence interval expression for \( \pi \), these conditions must be satisfied.

- If \( E(Y) \) is too close to 0 or \( n \), the normal distribution has too much area to the left of 0 or to the right of \( n \) to use the normal approximation.

Exercises 7.20 – 7.21:
As part of a market research study, in a sample of 125, 84 individuals are aware of a certain product. Calculate a 90% confidence interval for the proportion of individuals in the population who are aware of the product.

\[ \pi = \text{Proportion of individuals in the population who are aware of product.} \]
\[ n = 125, \quad y = 84, \quad \hat{\pi} = \frac{84}{125} = 0.672 \]

\[ \hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \]
\[ (0.672) \pm (1.645) \sqrt{\frac{(0.672)(0.328)}{125}} = [0.60, 0.74] \]

This is a 90% confidence interval for the population proportion who are aware of the product.

- When would such a product awareness study be undertaken?
  
  One possibility would be prior to the start of an advertising campaign. Such a study would also be undertaken after the advertising campaign to determine the effectiveness of the advertising campaign.
7.3 Confidence Intervals for a Proportion

• Minitab Output for Exercise 7.20

Test and CI for One Proportion

Test of p = 0.5 vs. p not = 0.5

<table>
<thead>
<tr>
<th>Sample</th>
<th>X</th>
<th>N</th>
<th>Sample p</th>
<th>90% CI</th>
<th>Z-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84</td>
<td>125</td>
<td>0.672000</td>
<td>(0.602929, 0.741071)</td>
<td>3.85</td>
<td>0.000</td>
</tr>
</tbody>
</table>

7.3 Confidence Intervals for a Proportion

7.21: Should the normal approximation underlying the confidence interval of Exercise 7.20 be adequate?

Conditions for using the normal approximation:
\[ n\pi - 5 \geq 0 \] and \[ n\pi + 5 \leq n. \]

Since \( \pi \) is unknown, use \( \hat{\pi} \).

The conditions become:
\[ n\hat{\pi} - 5 \geq 0 \]
\[ n\hat{\pi} + 5 \leq n. \]

Is \( n\hat{\pi} - 5 \geq 0 \)? Is \( (125)(84/125) - 5 \geq 0 \)? Yes!

Is \( n\hat{\pi} + 5 \leq n \)? Is \( (125)(84/125) + 5 \leq 125 \)? Yes!

• The conditions required to use the expression for a confidence interval based on the normal approximation for \( \hat{\pi} \) are satisfied.

7.3 Confidence Intervals for a Proportion

• Procedure to Obtain a Confidence Interval for a Proportion Using Minitab:

  → Click on Stat → Basic Statistics → 1 Proportion
  → Select "Summarized Data" and enter 125 and 84 for "Number of Trials" and "Number of Successes."
  → Click on "Options" and enter 90.0 for the "Confidence level."
  → Choose "Tests and interval based on normal distribution."
  → Click on "OK"
Section 7.4
How Large a Sample is Needed?

7.4 How Large a Sample is Needed?

- Sampling error is the difference between the value of a population parameter and its estimate.
  
  Parameter Estimate
  [ ? ] [Based on data]

- The difference between the parameter and its estimate is due to chance.

- Choosing an appropriate sample size controls the magnitude of the sampling error.

Scenario One: Find \( n \) to estimate \( \mu \)

Exercise 7.45:
A research project for an insurance company wishes to investigate the mean value of the personal property held by urban apartment renters. A previous study suggested that the population standard deviation should be roughly $10,000. A 95% confidence interval with a width of $1000 (a plus or minus of $500) is desired. How large a sample must be taken to obtain such a confidence interval?
### 7.4 How Large a Sample is Needed?

- **In General**
  - Find the sample size \( n \) so that the bound on the error of estimation \( (E) \) will hold with a high probability \((1 - \alpha)\).
  - Equivalently, find \( n \) so that the width \((2E)\) of a \(100(1 - \alpha)\)% confidence interval does not exceed a certain bound.
  - \( E \) is measured in standard deviations of \( \bar{Y} \), where
    \[
    E = z_{\alpha/2}(\sigma/\sqrt{n})
    \]
    \[
    \Rightarrow n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}
    \]
  - Need to specify \( E, 1 - \alpha \) and \( \sigma \) to find \( n \).

#### Exercise 7.45:

\[ E = \pm \$500, \sigma = \$10,000, 1 - \alpha = .95 \]

\[
\begin{align*}
\nonumber n &= \frac{z_{\alpha/2}^2 \sigma^2}{E^2} \\
&= \frac{(1.96)^2 (10,000)^2}{(500)^2} \\
&= 1537
\end{align*}
\]

### Scenario Two: Find \( n \) to estimate \( \pi \)

#### Exercise 7.69:

An electrical utility offers reduced rates to homeowners who have installed "peak hours" meters. These meters effectively shut off high-consumption electrical appliances (primarily dishwashers and clothes dryers) during the peak electrical usage hours between 9 a.m. and 3 p.m. daily. The utility wants to inspect a sample of these meters to determine the proportion that are not working, either because they were bypassed or because of equipment failure. There are 45,300 meters in use and the utility isn’t about to inspect them all.

- a. The utility wants a 90% confidence interval for the proportion with a width of no more than .04. How many meters must be sampled, if one makes no particular assumption about the correct proportion?
  - b. How many meters must be sampled if the utility assumes that the true population proportion is between .05 and .15?
  - c. Does the assumption in part (b) lead to a substantial reduction in the required sample size?
7.4 How Large a Sample is Needed?

- E is measured in standard errors of \( \hat{\pi} \), where
  \[
  \sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}}
  \]

  \[
  \Rightarrow E = z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}}
  \]

  \[
  \Rightarrow n = \left( \frac{z_{\alpha/2}^2}{E^2} \right) \pi(1-\pi)
  \]

- Good news! We have an expression to find n.
- Bad news! The expression depends on n, which we are trying to find.

7.4 How Large a Sample is Needed?

- **Approach 1** (Worse case scenario): Set \( \pi = \frac{1}{2} \).
  
  \[
  n = \left( \frac{z_{\alpha/2}^2}{E^2} \right) \left( \frac{1}{4} \right)
  \]

  **Exercise 7.69:**
  The utility wants a 90% confidence interval for the proportion with a width of no more than .04. How many meters must be sampled, if one makes no particular assumption about the correct proportion?

  \[
  n = \left( \frac{1.645^2}{.02^2} \right) \left( \frac{1}{4} \right) = 1692
  \]

7.4 How Large a Sample is Needed?

- **Approach 2**: Use a prior estimate of \( \pi \), denoted \( \pi_0 \), if available.
  
  \[
  n = \left( \frac{z_{\alpha/2}^2}{E^2} \right) \pi_0(1-\pi_0)
  \]

  **Exercise 7.69:**
  b. How many meters must be sampled if the utility assumes that the true population proportion is between .05 and .15?

  One perspective: Of the two population proportions stated, choose that value resulting in the larger \( n \). This occurs when \( \pi_0 = 0.15 \)

  \[
  n = \left( \frac{1.645^2}{.02^2} \right) \left( .15(.85) \right) = 862.5
  \]

  \[
  n = 863
  \]
Another perspective: Of the two population proportions stated, choose that value midway between them or let $\pi_0 = 0.10$.

$$n = \left( \frac{1.645}{.02} \right)^2 (\pi_0)(.90) = 608.86 = 609$$

Exercise 7.69:

c. Does the assumption in part (b) lead to a substantial reduction in the required sample size?

The percentage reduction is

$$(863 - 1692)/(1692) = -49\%$$

Section 7.5
The $t$ Distribution
7.5 The t Distribution

- Recall
  \[ Z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} \]
  has a standard normal distribution.

- Gosset (pseudonym: Student) determined the distribution of \( Z \) when "s" is used as an estimate of \( \sigma \):
  \[ t = \frac{\bar{Y} - \mu}{s / \sqrt{n}} \]

- The sample standard deviation \( s = \sqrt{\frac{\sum}{n-1}} \) has (n-1) degrees of freedom.

---

Properties of Student’s t Distribution (Hildebrand, Ott, and Gray)

1. The t distribution is symmetric about 0.
2. The t distribution is more variable than the Z distribution (Figure 7.8).
   - t distribution has heavier tails. Why?
3. There are many different t distributions.
   - We specify a particular one by its “degrees of freedom,” d.f.
   - If a random sample is taken from a normally distributed population, then the statistic
     \[ t = \frac{\bar{Y} - \mu}{s / \sqrt{n}} \]
     has a t distribution with \((n - 1)\) degrees of freedom.
4. As \( n \) increases, the distribution of \( t \) approaches the distribution of a standard normal.
7.5 The t Distribution

- Percentiles of the t-distribution are in Table 4

<table>
<thead>
<tr>
<th>df</th>
<th>s = .01</th>
<th>s = .05</th>
<th>s = .025</th>
<th>s = .01</th>
<th>s = .05</th>
<th>s = .001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>318.309</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.886</td>
<td>2.920</td>
<td>6.965</td>
<td>9.925</td>
<td>22.327</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.638</td>
<td>2.353</td>
<td>4.303</td>
<td>5.841</td>
<td>9.925</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.533</td>
<td>2.132</td>
<td>3.747</td>
<td>4.604</td>
<td>8.610</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.478</td>
<td>2.015</td>
<td>3.365</td>
<td>4.032</td>
<td>7.815</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.440</td>
<td>1.943</td>
<td>3.143</td>
<td>3.707</td>
<td>7.457</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.415</td>
<td>1.895</td>
<td>3.007</td>
<td>3.499</td>
<td>7.244</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.397</td>
<td>1.860</td>
<td>2.896</td>
<td>3.355</td>
<td>7.041</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.383</td>
<td>1.833</td>
<td>2.764</td>
<td>3.250</td>
<td>6.841</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.372</td>
<td>1.812</td>
<td>2.660</td>
<td>3.169</td>
<td>6.684</td>
<td></td>
</tr>
</tbody>
</table>

- For example, with n=10, the d.f. = 9, and P(t_9 > 2.262) = .025
- It is customary to say 2.262 = t .025, 9

Section 7.6
Confidence Intervals with the t Distribution

- When \( \sigma \) is known, the C.I. for \( \mu \) is given by

\[
\bar{Y} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)
\]

- When \( \sigma \) is unknown, it seems reasonable to replace \( \sigma \) by s.

\[
\bar{Y} \pm s_{\alpha/2} \left( \frac{s}{\sqrt{n}} \right)
\]

- Also, need to replace \( z_{\alpha/2} \) by \( t_{\alpha/2} \).
7.6 Confidence Intervals with the t Distribution

- The expression for a 100(1 - \(\alpha\))% confidence interval for \(\mu\) (\(\sigma\) unknown) is given by
  \[
  \bar{y} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}
  \]
- Requirements: Random Sample From a Normal Distribution.

Exercise 7.36:
A random sample of 20 taste-testers rate the quality of a proposed new product on a 0-100 scale. The ordered scores are
16 20 31 50 50 50 51 53 53 55
57 59 60 60 61 65 67 67 81 92
Minitab output follows. A box plot is shown in Figure 7.12

One-Sample T: Scores

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>StDev</th>
<th>SE Mean</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores</td>
<td>20</td>
<td>54.9</td>
<td>17.71</td>
<td>3.9603</td>
<td>(46.61, 63.19)</td>
</tr>
</tbody>
</table>

a. Locate the 95% confidence interval for the population mean score. Were \(t\) tables or \(z\) tables used?

\[
n = 20 \quad \bar{y} = 54.9 \quad s = 17.71
\]
\[
95\% \Rightarrow \alpha = .05 \Rightarrow \alpha/2 = .025 \Rightarrow t_{0.025,19} = 2.093
\]

A 100(1 - \(\alpha\))% C.I. for \(\mu\) is:

\[
54.9 \pm \frac{2.093 \times 17.71}{\sqrt{20}}
\]
\[
[46.61, 63.19] \text{ is a 95\% C.I. for the mean score of all tasters. } t \text{ tables were used.}\]
b. Is there any reason to think that the use of a mean-based confidence interval is a poor idea? Use the boxplot to answer this question.

The boxplot shows there are outliers in each tail. For data from a normal distribution, only 0.7% (approximately) of the values should be outliers. This implies that the distribution of Y (taste scores) is heavy-tailed or outlier-prone. For such populations, the sample mean is not the most efficient estimator of \( \mu \). The confidence interval based on the sample mean is unnecessarily wide. The NPP also shows that the data is not from a normal distribution.
7.6 Confidence Intervals with the t Distribution

- Procedure to obtain a confidence interval using Minitab:
  - Click on Stat → Basic Statistics → 1-Sample t
  - Enter variable in “Samples in Column” box
  - Click on Options
  - Enter .95 for confidence level
  - Click on OK

Section 7.7
Assumptions for Interval Estimation

7.7 Assumptions for Interval Estimation

- Statistical techniques require certain assumptions.
- All of the techniques in Chapter 7 require a *random* sample.
  - A biased sample is one that consistently yields units that differ from the true population for any number of reasons, including selection bias.
  - The techniques do not allow for bias in gathering the data.
7.7 Assumptions for Interval Estimation

• Another requirement is independence between the observations within the sample.

  • “In effect, dependence means that we don’t have as much information as the value of n indicates …. extreme dependence would arise in a sample of 25 observations if the first observation was genuinely random, but every succeeding observation had to equal the first one …. in fact we’d have a sample of only 1.” (Hildebrand, Ott & Gray)

• This requirement is frequently violated in time-series data, where an observation at one point in time could be related to an observation at another point in time.

• An example of time-series data is monthly champagne sales. Sales for certain months of the year are higher than in other months.

• If the observations are independent, a time series plot of the data should show no patterns.

Exercise 13.42: An auto-supply store had 60 months of data on variables that were thought to be relevant to sales [measured in thousands of dollars]. Are the sales observations independent?

Although there are formal statistical tests for assessing independence, a time series plot of sales vs. month is also recommended.

In the time series plot that follows, there is a clear up-down-up cyclic pattern in the data. This pattern indicates that the observations are not independent.

It would be wrong to use a t-interval to find a confidence interval for the mean monthly sales.
7.7 Assumptions for Interval Estimation

Sales tend to be higher in December and the months immediately after December, and lower in the summer months.

- Some of the techniques in Chapter 7 are more sensitive to departures from certain assumptions than others.
- All of the techniques in Chapter 7 are very sensitive to departures from the independence assumption.
- Another assumption for the Z- and t-intervals for a mean is that the underlying population is normally distributed.
- "In practice, no population is exactly normal. … this assumption is guaranteed to be more or less wrong." (Hildebrand, Ott & Gray)

- For the Z-interval, the Central Limit Theorem assures us that the sample mean is approximately normally distributed for sufficiently large n, regardless of the population distribution. If the distribution of the population is severely skewed, a larger sample size is required to account for this.
- Thus, the Z-interval is robust to departures from the assumption that the underlying population be normally distributed.
- This is not necessarily the case for the t-interval for the mean.
7.7 Assumptions for Interval Estimation

- The consequences of nonnormality on the t-interval depend on the type of nonnormality.
  - If the distribution of the population is symmetric, but heavy-tailed, the stated confidence level is fairly accurate.
  - If the distribution of the population is skewed, the stated confidence level is affected.
  - When the distribution of the population is symmetric, but heavy-tailed, more efficient procedures are recommended, for example, a trimmed mean.
  - These robust procedures give more accurate estimates and have smaller standard errors.
  - A normal probability plot is useful for determining the form of the population distribution.

Keywords: Chapter 7

- Point estimation
- Estimator
- Unbiased estimator
- Efficient estimator
- Interval estimation
- Z – interval

- Confidence interval for a proportion
- Required sample size
- t distribution
- t interval
- Independent observations
Summary of Chapter 7

- Inductive inference – estimating a population parameter
- How to obtain a point estimate of a population parameter
  - What does it mean for a point estimator to be unbiased?
  - What does it mean for a point estimator to be efficient?
- How to use a confidence interval estimate for the mean when \( \sigma \) is specified
- How to use a confidence interval estimate for the mean when \( \sigma \) is unknown

- How to use a confidence interval for a proportion
- How to determine the sample size required to estimate a mean and a proportion
- How to check the underlying assumptions for confidence interval estimation
- A flow chart to assist in using the correct confidence interval follows.

Ch. 7: Flow Chart for Confidence Intervals

IS IT A C.I. FOR \( \pi \)?

- IS \( \hat{p} \) - 5 \( \geq \) 0 AND \( \hat{p} + 5 \leq n \)?
  - YES: \( \pi \pm \frac{z_{\alpha/2} \sqrt{\pi(1-\pi)/n}}{} \)
  - NO: \( \hat{p} \pm \frac{1}{2n} \sqrt{\frac{n-1}{n}} \)

IS IT A C.I. FOR \( \mu \)?

- IS \( \sigma \) KNOWN?
  - YES: \( \bar{y} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \)
  - NO: \( \bar{y} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right) \)

USE:

\[
\hat{p} - \frac{z_{\alpha/2} \sqrt{\pi(1-\pi)/n}}{}
\]

USE A Z-INTERVAL:

\[
\bar{y} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)
\]

USE A T-INTERVAL:

\[
\bar{y} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}
\]