Estimating the Value of a Parameter Using Confidence Intervals

Overview
- We apply the results about the sample mean to the problem of estimation
- Estimation is the process of using sample data to estimate the value of a population parameter
- We will quantify the accuracy of our estimation process

The Logic in Constructing Confidence Intervals about a Population Mean when Population Standard Deviation is Known

Learning Objectives
- Compute a point estimate of the population mean
- Construct and interpret a confidence interval about the population mean (assuming the population standard deviation is known)
- Understand the role of margin of error in constructing a confidence interval
- Determine the sample size necessary for estimating the population mean within a specified margin of error

Estimation
- The environment of our problem is that we want to estimate the value of an unknown population mean
- The process that we use is called estimation
- This is one of the most common goals of statistics

Point Estimate
- Estimation involves two steps
  - Step 1 – to obtain a specific numeric estimate, this is called the point estimate
  - Step 2 – to quantify the accuracy and precision of the point estimate
- The first step is relatively easy
- The second step is why we need statistics
Examples of Point Estimate

- Some examples of point estimates are
  - The sample mean to estimate the population mean
  - The sample standard deviation to estimate the population standard deviation
  - The sample proportion to estimate the population proportion
  - The sample median to estimate the population median

Precision of Point Estimate

- The most obvious point estimate for the population mean is the sample mean
- Now we will use the material on the sampling distribution of sample mean to quantify the accuracy and precision of this point estimate

Example

- An example of what we want to quantify
  - We want to estimate the miles per gallon for a certain car
  - We test some number of cars
  - We calculate the sample mean … it is 27
  - 27 miles per gallon would be our best guess

Example (continued)

- How sure are we that the gas economy is 27 and not 28.1, or 25.2?
- We would like to make a statement such as
  “We think that the mileage is 27 mpg and we’re pretty sure that we’re not too far off”

Interval Estimation

- A confidence interval for an unknown parameter is an interval of numbers
  - Compare this to a point estimate which is just one number, not an interval of numbers (a range of numbers)
- The level of confidence represents the expected proportion of intervals that will contain the parameter if a large number of different samples is obtained
- The confidence interval quantifies the accuracy and precision of the point estimate

Interpret Confidence level

What does the level of confidence represent?

- If we obtain a series of 50 random samples from a population of interest
- Follow a process for calculating confidence intervals for population mean with a 90% level of confidence from each of the sample means
- Then, we would expect that 90% of those 50 confidence intervals (or about 45) would contain our population mean
Confidence Level

- The level of confidence is always expressed as a percent
- The level of confidence is described by a parameter \( \alpha \) (i.e., alpha)
- The level of confidence is \((1 - \alpha) \times 100\%\)
  - When \( \alpha = .05 \), then \((1 - \alpha) = .95\), and we have a 95% level of confidence
  - When \( \alpha = .01 \), then \((1 - \alpha) = .99\), and we have a 99% level of confidence

Confidence Interval

- If we expect that a method would create intervals that contain the population mean 90% of the time, we call those intervals 90% confidence intervals
- If we have a method for intervals that contain the population mean 95% of the time, those are 95% confidence intervals
- And so forth

Summary

- To tie the definitions together
  - We are using the sample mean to estimate the population mean (Point estimate)
  - With each specific sample, we can construct a ,for instance, 95% confidence interval to estimate the population mean… (Interval estimate)
  - 95% confidence interval tells you that If we take samples repeatedly, we expect that 95% of these intervals would contain the population mean

Example

- Back to our 27 miles per gallon car
  “We think that the mileage is 27 mpg and we’re pretty sure that we’re not too far off”
- Putting in numbers (quantify the accuracy)
  “We estimate the gas mileage is 27 mpg and we are 90% confident that the real mileage of this model of car is between 25 and 29 miles per gallon”

Example (continued)

“We estimate the gas mileage is 27 mpg”
- This is our point estimate
  “and we are 90% confident that”
- Our confidence level is 90% (which is 1 - \( \alpha \), i.e. \( \alpha = 0.10 \))
  “the real mileage of this model of car”
- The population mean
  “is between 25 and 29 miles per gallon”
- Our confidence interval is (25, 29)

Known Population Standard Deviation

- First, we assume that we know the standard deviation of the population (\( \sigma \))
- This is not very realistic … but we need it for right now to introduce how to construct a confidence interval
- We’ll solve this problem in a better way (where we don’t know what \( \sigma \) is) later… but first we’ll do this one
**Assumption**

To estimate the mean \( \mu \) with a known \( \sigma \), we need a normal distribution assumption for the sampling distribution of means.

Assumption satisfied by:
1. Knowing that the sampled population is normally distributed, or
2. Using a large enough random sample (CLT)

**Note:** The CLT may be applied to smaller samples (for example \( n = 15 \)) when there is evidence to suggest a unimodal distribution that is approximately symmetric. If there is evidence of skewness, the sample size needs to be much larger.

---

**Sampling Distribution of Means**

- The values of a general normal random variable are within 1.96 times (or about 2 times according to empirical rule) its standard deviation away from its mean 95% of the time
- Thus the sample mean is within
  \[ \mu \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \]
  of the population mean 95% of the time

![Sampling Distribution of Means](image)

Here, \( \bar{x} \sim \frac{\mu}{\sqrt{n}} \)

---

**Interval for Sample Mean**

- Because the sample mean has an approximately normal distribution, it is in the interval
  \[ \mu \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \]
  around the (unknown) population mean 95% of the time. In other words, the interval will cover 95% of possible sample means, when you take samples from the population repeatedly.
- Since \( \bar{x} = \mu \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \), we can flip the equation around between \( \mu \) and \( \bar{x} \) to solve for the population mean \( \mu \)

---

**Interval for Population Mean**

- After we solve for the population mean \( \mu \), we find that \( \mu \) is within the interval
  \[ \bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \]
  around the (known) sample mean “95% of the time”
- This isn’t exactly true in the mathematical sense as the population mean is not a random variable … that’s why we call this a “confidence” instead of a “probability”

---

**Confidence Interval**

- Thus a 95% confidence interval for the Population mean is
  \[ \bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \]
- This is in the form
  \[ \text{Point estimate} \pm \text{margin of error} \]
- The margin of error here is \( 1.96 \cdot \frac{\sigma}{\sqrt{n}} \)
### Example

- For our car mileage example
  - Assume that the sample mean was 27 mpg
  - Assume that we tested a sample of 40 cars
  - Assume that we knew that the population standard deviation was 6 mpg
- Then our 95% confidence interval estimate for the true/population mean mileage would be
  \[
  27 \pm \frac{6}{\sqrt{40}}
  \]
  or \(27 \pm 1.9\)

### Critical Value

- If we wanted to compute a 90% confidence interval, or a 99% confidence interval, etc., we would just need to find the right standard normal value (instead of 1.96 for a 95% confidence interval) called **critical value**
- Frequently used confidence levels, and their critical values, are
  - 90% corresponds to 1.645
  - 95% corresponds to 1.960
  - 99% corresponds to 2.575

### How to Determine Critical Value?

- Why do we use \(Z_{0.025}\) for 95% confidence?
- To be within something 95% of the time
  - We can be too low 2.5% of the time
  - We can be too high 2.5% of the time
- Thus the 5% confidence that we don’t have is split as 2.5% being too high and 2.5% being too low …

### Critical Value \(z_{\alpha/2}\) for Confidence Level 1 – \(\alpha\)

- In general, for a \((1 – \alpha)\) 100% confidence interval, we need to find \(z_{\alpha/2}\), the critical Z-value
- \(z_{\alpha/2}\) is the value such that
  \[
  P(Z \geq z_{\alpha/2}) = \alpha/2
  \]

### Critical Value \(z_{\alpha/2}\) for 1 – \(\alpha\) Confidence Level

- Once we know these critical values for the normal distribution, then we can construct confidence intervals for the population mean
  \[
  \bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \quad \text{to} \quad \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}
  \]
Example

The weights of full boxes of a certain kind of cereal are normally distributed with a standard deviation of 0.27 oz. A sample of 18 randomly selected boxes produced a mean weight of 9.87 oz. Find a 95% confidence interval for the true mean weight of a box of this cereal.

Solution:

Follow the process below to solve:

1. Describe the population parameter of concern
   The mean, $\mu$, weight of all boxes of this cereal

2. Specify the confidence interval criteria
   a. Check the assumptions
      The weights are normally distributed, the distribution of $\bar{X}$ is normal
   b. Identify the probability distribution and formula to be used
      Use a $z$-interval with $\sigma = 0.27$
   c. Determine the level of confidence, $1 - \alpha$
      The question asks for 95% confidence, so $1 - \alpha = 0.95$

3. Collect and present information
   The sample information is given in the statement of the problem
   Given: $n = 18$, $\overline{X} = 9.87$

4. Determine the confidence interval
   a. Determine the critical value either from a $z$-table or a TI graphing calculator
      $\text{invNorm}(1 - \frac{\alpha}{2}, 0, 1) = \text{invNorm}(0.975, 0, 1) = 1.96$
   b. Find the margin of error of estimate
      $Z_{\alpha/2} = 1.96 \times \frac{0.27}{\sqrt{18}} = 0.1247$
   c. Find the lower and upper confidence limits
      $\overline{X} \pm \text{Margin of Error}$
      $9.87 \pm 0.1247$
      $9.75$ to $10.00$

5. State the confidence interval and interpret it
   $9.75$ to $10.00$ is a 95% confidence interval for the true mean weight, $\mu$, of cereal boxes.
   This means that if we conduct the experiment over and over, and construct lots of confidence intervals, then 95% of the confidence intervals will contain the true mean value $\mu$.

Margin of Error

• If we write the confidence interval as $27 \pm 2$
then we would call the number 2 (after the $\pm$) the size of margin of error

• So we have three ways of writing a confidence interval
– $(25, 29)$
– $27 \pm 2$
– $27$ with a margin of error of 2

• The margin of errors would be
  – $1.645 \times \sigma / \sqrt{n}$ for 90% confidence intervals
  – $1.960 \times \sigma / \sqrt{n}$ for 95% confidence intervals
  – $2.575 \times \sigma / \sqrt{n}$ for 99% confidence intervals

• Once we know the margin of error, we can state the confidence interval as
  sample mean $\pm$ margin of error

• The margin of error which is half of a length of a confidence interval depends on three factors
  – The level of confidence ($1 - \alpha$)
  – The sample size ($n$)
  – The standard deviation of the population ($\sigma$)

Notice that:

- The higher the confidence level, the longer the length of the confidence interval. That is, a 99% confidence interval will be longer than a 90% confidence interval, because a wider interval will warrant better chance to cover the population mean.
- The larger the sample size, the shorter the confidence interval. This is because the larger the sample size, the smaller the standard error of the sample mean, which means the margin of error of the estimation is smaller.
- The larger the standard deviation of the population, the longer the confidence interval. So, if the value of the variable varies very much, the margin of error of the estimate increases.
Determine the sample size necessary for estimating the population mean within a specified margin of error

Sample Size Determination
- Often we have the reverse problem where we want an experiment to achieve a particular accuracy of the estimation. That is, we want to make sure the population mean can be estimated within a target margin of error from a sample mean.
- Since the sample size will affect the margin of error, we want to find the sample size \( n \) needed to achieve a particular size of margin of error in estimation.
- Sample size determination is needed in designing an experimental investigation before the data collection.

Example
- For our car miles per gallon, we had \( \sigma = 6 \)
- If we wanted our margin of error to be 1 for a 95% confidence interval, then we would need
  \[
  1.96 \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{6}{\sqrt{n}} = 1
  \]
- Solving for \( n \) would get us \( n = (1.96 \cdot 6)^2 \) or that \( n = 138 \) cars would be needed

Sample Size Determination
- We can write this as a formula
- The sample size \( n \) needed to result in a margin of error \( E \) for \((1 - \alpha) \cdot 100\% \) confidence is
  \[
  n = \left( \frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2
  \]
- Usually we don’t get an integer for \( n \), so we would need to take the next higher number (the one lower wouldn’t be large enough)

Summary
- We can construct a confidence interval around a point estimator if we know the population standard deviation \( \sigma \)
- The margin of error is calculated using \( \sigma \), the sample size \( n \), and the appropriate \( Z \)-value
- We can also calculate the sample size needed to obtain a target margin of error

Confidence Intervals about a Population Mean in Practice where the Population Standard Deviation is Unknown
Learning Objectives

- Know the properties of \( t \)-distribution
- Determine \( t \)-values
- Construct and interpret a confidence interval about a population mean

Unknown Population Standard Deviation

- So far we assumed that we knew the population standard deviation \( \sigma \)
- But, this assumption is not realistic, because if we know the population standard deviation, we probably would know the population mean as well. Then there is no need to estimate the population mean using a sample mean.
- So, it is more realistic to construct confidence intervals in the case where we do not know the population standard deviation

Replacing \( \sigma \) with \( s \)

- If we don’t know the population standard deviation \( \sigma \), we obviously can’t use the formula
  
  \[
  \text{Margin of error} = 1.96 \times \frac{\sigma}{\sqrt{n}}
  \]
  
  because we have no number to use for \( \sigma \)
- However, just as we can use the sample mean to approximate the population mean, we can also use the sample standard deviation to approximate the population standard deviation

Student’s \( t \)-distribution

- Because we’ve changed our formula (by using \( s \) instead of \( \sigma \)), we can’t use the normal distribution any more
- Instead of the normal distribution, we use the Student’s \( t \)-distribution
- This distribution was developed specifically for the situation when \( \sigma \) is not known

Properties of \( t \)-distribution

- Several properties are familiar about the Student’s \( t \) distribution
  - Just like the normal distribution, it is centered at 0 and symmetric about 0
  - Just like the normal curve, the total area under the Student’s \( t \) curve is 1, the area to left of 0 is \( \frac{1}{2} \), and the area to the right of 0 is also \( \frac{1}{2} \)
  - Just like the normal curve, as \( t \) increases, the Student’s \( t \) curve gets close to, but never reaches, 0
Difference between $Z$ and $t$

- So what’s different?
- Unlike the normal, there are many different “standard” $t$-distributions
  - There is a “standard” one with 1 degree of freedom
  - There is a “standard” one with 2 degrees of freedom
  - There is a “standard” one with 3 degrees of freedom
  - Etc.

- The number of degrees of freedom is crucial for the $t$-distributions

\[ t \text{-statistic} \]

- When $\sigma$ is known, the $z$-score
  \[ z = \frac{x - \mu}{\sigma / \sqrt{n}} \]
  follows a standard normal distribution

- When $\sigma$ is not known, the $t$-statistic
  \[ t = \frac{x - \mu}{s / \sqrt{n}} \]
  follows a $t$-distribution with $n - 1$ (sample size minus 1) degrees of freedom

\[ t \text{-distribution} \]

- Comparing three curves
  - The standard normal curve
  - The $t$ curve with 14 degrees of freedom
  - The $t$ curve with 4 degrees of freedom

Calculation of $t$-distribution

- The calculation of $t$-distribution values $t_{\alpha}$ can be done in similar ways as the calculation of normal values $z_{\alpha}$
  - Using tables
  - Using technology – TI graphing Calculator

Determine $t$-values
Critical values $t$

- Critical values for various degrees of freedom for the $t$-distribution are (compared to the normal):

<table>
<thead>
<tr>
<th>$n$</th>
<th>Degrees of Freedom</th>
<th>$t_{0.025}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>5</td>
<td>2.571</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>2.131</td>
</tr>
<tr>
<td>31</td>
<td>30</td>
<td>2.042</td>
</tr>
<tr>
<td>101</td>
<td>100</td>
<td>1.984</td>
</tr>
<tr>
<td>1001</td>
<td>1000</td>
<td>1.962</td>
</tr>
<tr>
<td>Normal</td>
<td>&quot;Infinite&quot;</td>
<td>1.960</td>
</tr>
</tbody>
</table>

Note: When the sample size is large, a $t$ distribution is close to a $z$ distribution.

Construct and interpret a $t$-confidence interval about a population mean

- The difference between the two formulas is that the sample standard deviation $s$ is used to approximate the population standard deviation $\sigma$.
- The $z$-score has a normal distribution, the $t$-statistic (or the $t$-score) has a $t$-distribution.

95% Confidence interval for mean with unknown $\sigma$

- A 95% confidence interval, with $\sigma$ unknown, is:

$\bar{x} - t_{0.025} \cdot \frac{s}{\sqrt{n}}$ to $\bar{x} + t_{0.025} \cdot \frac{s}{\sqrt{n}}$

where $t_{0.025}$ is the critical value for the $t$-distribution with $(n - 1)$ degrees of freedom.

Note: Compare it to the 95% confidence interval, with a known $\sigma$:

$\bar{x} - z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}$ to $\bar{x} + z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}$

Critical Value $t_{\alpha/2}$ corresponding to Confidence Level $1 - \alpha$

- The different 95% confidence intervals with $t_{0.025}$ would be:
  - For $n = 6$, the sample mean $\pm 2.571 \cdot \frac{s}{\sqrt{6}}$
  - For $n = 16$, the sample mean $\pm 2.131 \cdot \frac{s}{\sqrt{16}}$
  - For $n = 31$, the sample mean $\pm 2.042 \cdot \frac{s}{\sqrt{31}}$
  - For $n = 101$, the sample mean $\pm 1.984 \cdot \frac{s}{\sqrt{101}}$
  - For $n = 1001$, the sample mean $\pm 1.962 \cdot \frac{s}{\sqrt{1001}}$
  - When $\sigma$ is known, the sample mean $\pm 1.960 \cdot \frac{\sigma}{\sqrt{n}}$

Confidence interval for mean with unknown $\sigma$

- In general, the $(1 - \alpha) \cdot 100\%$ confidence interval, when $\sigma$ is unknown, is:

$\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ to $\bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

where $t_{\alpha/2}$ is the critical value for the $t$-distribution with $(n - 1)$ degrees of freedom.
Approximate $t$ with $z$

- As the sample size $n$ gets large, there is less and less of a difference between the critical values for the normal and the critical values for the $t$-distribution
- Although $t$-critical value and $z$-critical value may be close to each other when the sample size is large, we still recommend to use a $t$-distribution when $\sigma$ is not known to obtain a more accurate answer
  - When doing rough assessment by hand, the normal critical values can be used, particularly when $n$ is large, for example if $n$ is 30 or more

Example 1

- Assume that we want to estimate the average weight of a particular type of very rare fish
  - We are only able to borrow 7 specimens of this fish
  - The average weight of these was 1.38 kg (the sample mean)
  - The standard deviation of these 7 specimens of this fish was 0.29 kg (a sample standard deviation)
- What is a 95% confidence interval for the true mean weight?

Example 1 (continued)

- $n = 7$, the critical value $t_{0.025}$ for 6 degrees of freedom is 2.447
- Our confidence interval thus is
  \[
  1.38 - \frac{0.29}{\sqrt{7}} = 1.11 \\
  1.38 + \frac{0.29}{\sqrt{7}} = 1.65
  \]
  or $(1.11, 1.65)$

Example 2

Suppose you do a study of acupuncture to determine how effective it is in relieving pain. You measure sensory rates for 15 subjects with the results given below. Use the sample data to construct a 95% confidence interval for the mean sensory rate for the population (assumed normal) from which you took the data:

8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 8.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9

Solution

To find the confidence interval, first we need to find the sample mean. Since population standard deviation is not given and we have the sample data to calculate the sample standard deviation, we can construct a $t$-confidence interval for estimating the mean.

Use TI calculator entering the data and obtain one-variable statistics. We obtain $\bar{x} = 8.2267$ and $s = 1.6722$, where $n = 15$

Critical value is $t_{0.025} = 2.145$

95% confidence interval is $8.2267 \pm 2.145 \times \frac{1.6722}{\sqrt{15}}$ or $7.00$ to $9.45$

Check the underlying distribution

- When apply a $t$-interval, we need to make sure the underlying population is approximately normally distributed.
- When the sample size is small, outlier of the data will have a major affect on the data set, because outliers will affect the calculation of sample mean and sample standard deviation.
- So what can we do?
  - For a small sample, we always must check to see that the outlier is a legitimate data value (and not just a typo)
  - We can collect more data, for example to increase $n$ to be over 30. Apply the central limit theorem, we can use a $z$-interval to approximate a $t$-interval.

Summary

- We used values from the normal distribution when we knew the value of the population standard deviation $\sigma$
- When we do not know $\sigma$, we estimate $\sigma$ using the sample standard deviation $s$
- We use values from the $t$-distribution when we use $s$ instead of $\sigma$, i.e. when we don’t know the population standard deviation
Confidence Intervals about a Population Proportion

Learning Objectives

• Obtain a point estimate for the population proportion
• Construct and interpret a confidence interval for the population proportion
• Determine the sample size necessary for estimating a population proportion within a specified margin of error

Obtain a point estimate for the population proportion

Mean & Proportion

• So far, we learned to calculate confidence intervals for the population mean, when we knew \( \sigma \) and \( \sigma \)
• We also learned to calculate confidence intervals for the mean, when we did not know \( \sigma \)
• Here, we’ll learn how to construct confidence intervals for situations when we are analyzing a population proportion
• The issues and methods are quite similar

Sample Proportion

• When we analyze the population mean, we use the sample mean as the point estimate
  – The sample mean is our best guess for the population mean

• When we analyze the population proportion, we use the sample proportion as the point estimate
  – The sample proportion is our best guess for the population proportion

Proportion – Point Estimate

• Using the sample proportion is the natural choice for the point estimate
• If we are doing a poll, and 68% of the respondents said “yes” to our question, then we would estimate that 68% of the population would say “yes” to our question also
• The sample proportion is written as \( \hat{p} \)
Construct and interpret a confidence interval for the population proportion

Confidence Interval for Mean versus Proportion

- Confidence intervals for the population mean are
  - Centered at the sample mean
  - Plus and minus $z_{\alpha/2}$ times the standard deviation of the sample mean (the standard error from the sampling distribution)
- Similarly, confidence intervals for the population proportion will be
  - Centered at the sample proportion
  - Plus and minus $z_{\alpha/2}$ times the standard deviation of the sample proportion

Sampling Distribution of Proportion

- We have already studied the distribution of the sample proportion is approximately normal with
  
  \[ \mu_p = p \]
  
  \[ \sigma_p = \sqrt{\frac{p(1-p)}{n}} \]

  under most conditions
- We use this to construct confidence intervals for the population proportion

Confidence Interval for Population Proportion

- The $(1 - \alpha) \cdot 100\%$ confidence interval for the population proportion is from
  \[ \hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ to } \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

  where $z_{\alpha/2}$ is the critical value for the normal distribution

  Note: That is, sample proportion $\pm z_{\alpha/2} \times$ standard error of sample proportion

Margin of Error

- Like for confidence intervals for population means, the quantity
  
  \[ z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \]

  is called the margin of error

Example

- We polled $n = 500$ voters (This a sample of voters)
- When asked about a ballot question, $\hat{p} = 47\%$ of them were in favor
- Obtain a 99% confidence interval for the population proportion in favor of this ballot question ($\alpha = 0.005$)
Example (continued)

• The critical value $z_{0.005} = 2.575$, so

\[
0.47 - 2.575 \cdot \frac{0.47 - 0.53}{500} = 0.41
\]

to

\[
0.47 + 2.575 \cdot \frac{0.47 - 0.53}{500} = 0.53
\]

or $(0.41, 0.53)$ is a 99% confidence interval for the population proportion

Determine the sample size necessary for estimating a population proportion within a specified margin of error

Sample Size Determination

• We often want to know the minimum sample size to obtain a target margin of error for estimating the population proportion

• A common use of this calculation is in polling … how many people need to be polled for the result to have a certain margin of error
  – News stories often say “the latest polls show that so-and-so will receive $X\%$ of the votes with a $E\%$ margin of error …”

Example 1

• For our polling example, how many people need to be polled so that we are within 1 percentage point with 99% confidence?

• The margin of error is

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

which must be 0.01

• We have a problem, though … what is $\hat{p}$?

Two choices of $\hat{p}$

• If we try to figure out the sample size $n$ in the experimental design stage before collecting data, then we do not have sample data to calculate $\hat{p}$ . A way around this is that using $\hat{p} = 0.5$ will always yield a sample size that is large enough.

• We can also use an estimate $\hat{p}$ from a previous study (historic data) to calculate the sample size.

Example 1 (continued)

• In our case, if we using $\hat{p} = 0.5$, then we have

\[
2.575 \cdot \sqrt{0.5 \cdot 0.5 / n} = .01
\]

so

\[
n = 25 \left( \frac{2.575}{0.01} \right)^2
\]

and $n = 16,577$
Example 1 (continued)

- We understand now why political polls often have a 3 or 4 percentage points margin of error
- Since it takes a large sample \((n = 16,577)\) to get to be 99% confident to within 1 percentage point, the 3 or 4 percentage points margin of error targets are good compromises between accuracy and cost effectiveness

Sample Size Determination

- We can write this as a formula
- The sample size \(n\) needed to result in a margin of error \(E\%\) for \((1 - \alpha)\) 100% confidence for a population proportion is
  \[
  n = \left( \frac{Z_{\alpha/2} \cdot \hat{p} \cdot (1 - \hat{p})}{E/100} \right)^2
  \]
- Usually we don’t get an integer for \(n\), so we would need to take the next higher number (the one lower wouldn’t be large enough)

Example 2

Determine the sample size necessary to estimate the true proportion of laboratory mice with a certain genetic defect. We would like the estimate to be within 0.015 with 95% confidence.

Solution:
1. Level of confidence: \(1 - \alpha = 0.95\), \(Z_{0.025} = 1.96\)
2. Desired maximum error is \(E = 0.015\).
3. No estimate of \(p\) given, use \(\hat{p} = 0.5\)
4. Use the formula for \(n\):
   \[
   n = \left( \frac{Z_{\alpha/2} \cdot \hat{p} \cdot (1 - \hat{p})}{E/100} \right)^2
   \]
   \[
   = \left( \frac{1.96 \cdot 0.5 \cdot 0.5}{0.015} \right)^2 \approx 4266.44 \approx 4269
   \]

Example 2 (continued)

Suppose we know the genetic defect occurs in approximately 1 of 80 animals
Use: \(\hat{p} = 1/80 = 0.0125\)

\[
\begin{align*}
  n &= \left( \frac{Z_{\alpha/2} \cdot \hat{p} \cdot (1 - \hat{p})}{E/100} \right)^2 \\
  &= \left( \frac{1.96 \cdot (0.0125)(0.9875)}{0.015} \right)^2 \approx 210.75 = 211
\end{align*}
\]

Note: As illustrated here, it is an advantage to have some indication of the value expected for \(p\), especially as \(p\) becomes increasingly further from 0.5

Summary

- We can construct confidence intervals for population proportions in much the same way as for population means
- We need to use the formula for the standard deviation of the sample proportion
- We can also compute the minimum sample size needed for a desired level of accuracy

Which Procedure Do I Use?
Overview

- There are three different confidence interval calculations covered in this unit.
- It can be confusing which one is appropriate for which situation.
- I should use the normal ... no, the t ... no the ... ???

Which Parameter?

- The one main question right at the beginning.
- Which parameter are we trying to estimate?
  - A mean?
  - A proportion?
- This the single most important question.

z-interval or t-interval?

- In analyzing population means.
- Is the population variance known?
  - If so, then we can use the normal distribution.
- If the population variance is not known.
  - If we have "enough" data (30 or more values), we still can use the normal distribution.
  - If we don’t have “enough” data (29 or fewer values), we should use the Student’s t-distribution.
- We don’t have to ask this question in the analysis of proportions.

z-interval for mean

- For the analysis of a population mean.
- If
  - The data is OK (reasonably normal)
  - The variance is known.
  then we can use the normal distribution with a confidence interval of

\[
\bar{x} - z_{a/2} \frac{\sigma}{\sqrt{n}} \quad \text{to} \quad \bar{x} + z_{a/2} \frac{\sigma}{\sqrt{n}}
\]

z-interval for Proportion

- For the analysis of a population proportion.
- If sample size is large enough,
  then we can use the proportions method with a confidence interval of

\[
\hat{p} - z_{a/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad \text{to} \quad \hat{p} + z_{a/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

t-interval for mean

- For the analysis of a population mean.
- If
  - The data is OK (reasonably normal)
  - The variance is NOT known.
  then we can use the Student’s t-distribution with a confidence interval of

\[
\bar{x} - t_{a/2} \frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{x} + t_{a/2} \frac{s}{\sqrt{n}}
\]
Summary

• The main questions that determine the confidence interval to use:
  • Is it a
    – Population mean?
    – Population proportion?
  • In the case of a population mean, we need to determine
    – Is the population variance known?
    – Does the data look reasonably normal?

---

Summary

• We can use a sample {mean, proportion} to estimate the population {mean, proportion}
• In each case, we can use the appropriate sampling distribution of the sample statistic to construct a confidence interval around our estimate
• The confidence interval expresses the confidence we have that our calculated interval contains the true parameter