Lecture 8: Confidence Intervals

Carolyn Anderson & Youngshil Paek

(Slide Contributors: Shuai Wang, Yi Zhang, Michael Culbertson, & Haiyan Li)

Department of Educational Psychology
University of Illinois at Urbana-Champaign
Topic 1: What are Point and Interval Estimates of Population Parameters?
Key Points

1. Point Estimate and Interval Estimate
2. Properties of Point Estimators
3. Confidence Intervals
4. Logic of Confidence Intervals
5. Margin of Error
6. Example
Point Estimate and Interval Estimate

- **A point estimate** is a *single number* that is our “best guess” for the parameter.
- **An interval estimate** is an *interval of numbers* within which the parameter value is believed to fall.
Point Estimate vs. Interval Estimate

- A *point estimate* doesn’t tell us how close the estimate is likely to be to the parameter
  - Any one point estimate may or may not be close to the parameter it estimates.
- An *interval estimate* is more useful
  - It incorporates a margin of error which helps us to gauge the accuracy of the point estimate.
Properties of Point Estimators

- **Property 1**: A good estimator has a sampling distribution that is centered at the parameter.
  - An estimator with this property is *unbiased*
    - The sample mean is an unbiased estimator of the population mean.
    - The sample proportion is an unbiased estimator of the population proportion.
Properties of Point Estimators

- **Property 2**: A good estimator has a *small standard error* compared to other estimators.
  - This means it tends to fall closer than other estimates to the parameter.
    - The sample mean has a smaller standard error than the sample median when estimating the population mean of a normal distribution.
A good point estimate has which of the following characteristics?

a) Bias: None  Standard Error: High
b) Bias: High  Standard Error: High
c) Bias: None  Standard Error: Low
d) Bias: High  Standard Error: Low
True or False: A point estimate is better than an interval estimate because it gives you the exact value for which you are looking.

a) True

b) False
Confidence Interval

- A confidence interval is an interval containing the most believable values for a parameter.
- Confidence Interval = Point Estimate ± Margin of Error
- The probability that this method produces an interval that contains the parameter is called the confidence level.
  - This is a number chosen to be close to 1, most commonly 0.95.
Logic of Confidence Intervals

To construct a confidence interval for a population proportion, start with the sampling distribution of a sample proportion, which:

- Gives the possible values for the sample proportion and their probabilities
- Is approximately a normal distribution for large random samples by the CLT
- Has mean equal to the population proportion
- Has standard deviation called the standard error
Logic of Confidence Intervals

- Approximately 95% of a normal distribution falls within 1.96 standard deviations of the mean

- With probability 0.95, the sample proportion falls within about 1.96 standard errors of the population proportion

- The distance of 1.96 standard errors is the margin of error in calculating a 95% confidence interval for the population proportion
Margin of Error

- A confidence interval is constructed by taking a point estimate and adding and subtracting a margin of error.  
  \[ \text{Confidence Interval} = \text{Point Estimate} \pm \text{Margin of Error} \]

- And the margin of error is a multiple of the standard error of the sampling distribution of that point estimate.

- The distance of $1.96 \times \text{standard errors}$ is the margin of error for a 95% confidence interval for a parameter with a normal sampling distribution.
When the sampling distribution is approximately normal, what is the margin of error equal to for a 95% confidence interval?

a) 1.96
b) 1.96*standard error
c) Standard error
d) Point estimate ± 1.96*standard error
Key Points Revisited

1. Point Estimate and Interval Estimate
2. Properties of Point Estimators
3. Confidence Intervals
4. Logic of Confidence Intervals
5. Margin of Error
6. Example
Topic 2: How Can We Construct a Confidence Interval to Estimate a Population Proportion?
Key Points

1. Finding the 95% Confidence Interval for a Population Proportion
2. Sample Size Needed for Large-Sample Confidence Interval for a Proportion
3. How Can We Use Confidence Levels Other than 95%?
4. What is the Error Probability for the Confidence Interval Method?
5. Effect of the Sample Size
6. Interpretation of the Confidence Level
Finding the 95% Confidence Interval for a Population Proportion

- We symbolize a population proportion by $p$.
- The point estimate of the population proportion is the *sample proportion*.
- We symbolize the sample proportion by $\hat{p}$, called “p-hat”.

Finding the 95% Confidence Interval for a Population Proportion

95% Confidence Interval for $p$

= point estimate ± margin of error
= $\hat{p} \pm 1.96($standard errors$)$
= $[\hat{p} - 1.96($S.E.$), \hat{p} + 1.96($S.E.$)]$
Finding the 95% Confidence Interval for a Population Proportion

- The exact standard error of a sample proportion equals:
  \[ \sqrt{p(1-p)} \]
  \[ \frac{1}{n} \]

- This formula depends on the unknown population proportion \( p \)

- In practice, we don’t know \( p \), and we need to estimate the standard error as
  \[ S.E. = \sqrt{\hat{p}(1-\hat{p})} \]
  \[ \frac{1}{n} \]
Finding the 95% Confidence Interval for a Population Proportion

- A 95% confidence interval for a population proportion $p$ is:

$$\hat{p} \pm 1.96(S.E.), \text{ with S.E. } = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
Interpretation of the Confidence Level

- The meaning of CI refers to a long-run interpretation—how the method performs when used over and over with many different random samples.
- If we used the 95% confidence interval method over time to estimate many population proportions, then in the long run about 95% of those intervals would give correct results, containing the population proportion.
How Can We Use Confidence Levels Other than 95%?

- “95% confidence” means that there is a 95% chance that a sample proportion value occurs such that the confidence interval contains the unknown value of the population proportion, \( p \).  
  \[ \hat{p} \pm 1.96(\text{S.E.}) \]

- **Error Probability**: with probability \( 1 - .95 = .05 \), the method produces a confidence interval that misses \( p \)
Example 1

- In 2000, the GSS asked: “Are you willing to pay much higher prices in order to protect the environment?”
  - Of \( n = 1154 \) respondents, 518 were willing to do so
- Find and interpret a 95% confidence interval for the population proportion of adult Americans willing to do so at the time of the survey
Example 1

\[ \hat{p} = \frac{518}{1154} = 0.45 \]

\[ se = \sqrt{\frac{(0.45)(0.55)}{1154}} = 0.015 \]

\[ \hat{p} \pm 1.96(se) \]

\[ = 0.45 \pm 0.03 \]

\[ = (0.42, 0.48) \]
Sample Size Needed for Large-Sample Confidence Interval for a Proportion

- For the confidence interval for a proportion $p$ to be valid, you should have at least 15 successes and 15 failures:

$$n\hat{p} \quad 15 \text{ and } n(1 - \hat{p}) \quad 15$$
How Can We Use Confidence Levels Other than 95%?

- In practice, the confidence level 0.95 is the most common choice
- But, some applications require greater (or less) confidence
- To increase the chance of a correct inference, we use a larger confidence level, such as 0.99
How Can We Use Confidence Levels Other than 95%?

- The general formula for the confidence interval for a population proportion is:

\[ \hat{p} \pm z \text{ (S.E.)} \]

in which we can use different z-score for different confidence levels.

<table>
<thead>
<tr>
<th>TABLE 7.2: z-Scores for the Most Common Confidence Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>The large-sample confidence interval for the population proportion is ( \hat{p} \pm z(se) ).</td>
</tr>
<tr>
<td>Confidence Level</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>0.90</td>
</tr>
<tr>
<td>0.95</td>
</tr>
<tr>
<td>0.99</td>
</tr>
</tbody>
</table>
Based off of the same sample, which of the confidence intervals for the population mean would be the widest?

a) A 90% confidence interval
b) A 95% confidence interval
c) A 99% confidence interval
d) Cannot be determined
Example 2

- Exit poll: Out of 1400 voters, 660 voted for the Democratic candidate.
- Calculate a 90% and 99% Confidence Interval
Why Settle for Anything Less Than 100% Confidence?

- Because to be absolutely 100% certain of a correct inference (that is, of capturing the parameter value inside the confidence interval), the confidence interval must contain all possible values for the parameter. This is not helpful.
- In practice, we settle for a little less than perfect confidence so we can estimate the parameter value more precisely, which is far more informative.
Summary: Confidence Interval for a Population Proportion, $p$

- A confidence interval for a population proportion $p$ is:

$$\hat{p} \pm z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- Assumptions
  - Data obtained by randomization
  - A large enough sample size $n$ so that the number of success, $n\hat{p}$, and the number of failures, $n(1 - \hat{p})$, are both at least 15
Effects of Confidence Level and Sample Size on Margin of Error

- The *margin of error* for a confidence interval:
  - Increases as the confidence level increases
  - Decreases as the sample size increases

\[
\text{Margin of error} = z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]
**POLL: Who won the final debate?**

Regardless of which candidate you happen to support, who do you think did the best job in the debate -- Barack Obama or Mitt Romney?

- **Obama**: 48%
- **Romney**: 40%

Source: CNN/ORC International
Date conducted: 10/22/2012
Sample: Debate watchers
Margin of error: +/- 4.5% pts

More About: Barack Obama, CNN/ORC International, Debates, Mitt Romney
Key Points Revisited

1. Finding the 95% Confidence Interval for a Population Proportion
2. Sample Size Needed for Large-Sample Confidence Interval for a Proportion
3. How Can We Use Confidence Levels Other than 95%?
4. What is the Error Probability for the Confidence Interval Method?
5. Effect of the Sample Size
6. Interpretation of the Confidence Level
Topic 3: How Can We Construct a Confidence Interval to Estimate a Population Mean?
Key Points

1. How to Construct a Confidence Interval for a Population Mean
2. Properties of the t Distribution
3. Formula for 95% Confidence Interval for a Population Mean
4. How Do We Find a t Confidence Interval for Other Confidence Levels?
5. If the Population is Not Normal, is the Method “Robust”?
6. The Standard Normal Distribution is the $t$ Distribution with $df = \infty$
How to Construct a Confidence Interval for a Population Mean

- Recall: CI = Point estimate ± Margin of error
- The sample mean is the point estimate of the population mean
- The exact standard error of the sample mean is:
  
  \[ S.E. = \frac{s}{\sqrt{n}} \]

- In practice, we estimate \( \sigma \) by the sample standard deviation, \( s \):
  
  \[ S.E. = \frac{s}{\sqrt{n}} \]
How to Construct a Confidence Interval for a Population Mean

- For a **large sample** from any population
  OR
- For a **small sample** from an **underlying population** that is normal,

  if we knew the population standard deviation, $\sigma$, the confidence interval for the population mean would be:

  $$ \bar{x} \pm z \left( \frac{s}{\sqrt{n}} \right) $$
Use t-score Instead of z-score

- In practice, we don’t know the population standard deviation $\sigma$
- Substituting the sample standard deviation, $s$, for $\sigma$ to get $S.E = s / \sqrt{n}$ introduces extra error
- To account for this increased error, we replace the z-score by a slightly larger score, the t-score. The confidence interval is then a bit wider. This distribution is called the “t-distribution”.

$$S.E = s / \sqrt{n}$$
Properties of the $t$-distribution

- The $t$-distribution is bell shaped and symmetric about 0.
- The probabilities depend on the degrees of freedom, $df = n - 1$.
- The $t$-distribution has thicker tails than the standard normal distribution, i.e., it is more spread out.
The t-distribution has thicker tails and is more spread out than the standard normal distribution.
## t-distribution

<table>
<thead>
<tr>
<th>df</th>
<th>80% $t_{.100}$</th>
<th>90% $t_{.050}$</th>
<th>95% $t_{.025}$</th>
<th>98% $t_{.010}$</th>
<th>99% $t_{.005}$</th>
<th>99.8% $t_{.001}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>318.3</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.440</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
<td>5.208</td>
</tr>
<tr>
<td>7</td>
<td>1.415</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
<td>4.785</td>
</tr>
</tbody>
</table>

![t-distribution diagram](image)
In R

- For probabilities $p(t_1 \leq 12.7) = pt(12.7,1) = .975$

- In general, $pt(t\text{-value}, df)$

- For $t$ values for a specific probability, $qt(.975,1) = 12.7062$

- In general, $qt(\text{quantile}, df)$
Formula for 95% Confidence Interval for a Population Mean

- When the standard deviation of the population is unknown, a 95% confidence interval for the population mean $\mu$ is:

$$\bar{x} \pm t_{.025, df=n-1} \left( \frac{S}{\sqrt{n}} \right)$$

- To use this method, you need:
  - Data obtained by randomization
  - An approximately normal population distribution
Example: 95% Confidence Interval of a Population Mean

A study of 7 American adults from an SRS yields an average height of 67.2 inches and a standard deviation of 3.9 inches. Assuming the heights are normally distributed, a 95% confidence interval for the average height of all American adults($m$) is:
\[ x \pm t \frac{s}{\sqrt{n}} = 67.2 \pm 2.447 \frac{3.9}{\sqrt{7}} = 67.2 \pm 3.607 \]

= (63.593, 70.807)

Conclusion: we are “95% confident” that the average height of all American adults is between 63.6 and 70.8 inches.

What does “95% confident” mean?
How Do We Find a Confidence Interval of a Population Mean for Other Confidence Levels?

- The general formula for the confidence interval for a population proportion is:

  \[ \bar{x} \pm t, \text{df}=n-1 \left(\frac{s}{\sqrt{n}}\right) \]

- The 95% confidence interval uses \( t_{.025} \) since 95% of the probability falls between \(-t_{.025}\) and \( t_{.025} \).

- For 99% confidence, the error probability is 0.01 with 0.005 in each tail and the appropriate t-score is \( t_{.005} \).

- To get other confidence intervals use the appropriate t-value from Table B ... or better, from R.
How Do We Find a Confidence Interval of a Population Mean for Other Confidence Levels?

**TABLE 7.5: Part of Table B Displaying t-Scores for Large df Values**

The z-score of 1.96 is the t-score $t_{0.025}$ with right-tail probability of 0.025 and $df = \infty$.

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>98%</th>
<th>99%</th>
<th>99.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right-Tail Probability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$df$</td>
<td>$t_{1.00}$</td>
<td>$t_{0.050}$</td>
<td>$t_{0.025}$</td>
<td>$t_{0.010}$</td>
<td>$t_{0.005}$</td>
<td>$t_{0.001}$</td>
</tr>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>318.3</td>
</tr>
<tr>
<td>30</td>
<td>1.310</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.385</td>
</tr>
<tr>
<td>40</td>
<td>1.303</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>3.307</td>
</tr>
<tr>
<td>50</td>
<td>1.299</td>
<td>1.676</td>
<td>2.009</td>
<td>2.403</td>
<td>2.678</td>
<td>3.261</td>
</tr>
<tr>
<td>60</td>
<td>1.296</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>3.232</td>
</tr>
<tr>
<td>80</td>
<td>1.292</td>
<td>1.664</td>
<td>1.990</td>
<td>2.374</td>
<td>2.639</td>
<td>3.195</td>
</tr>
<tr>
<td>100</td>
<td>1.290</td>
<td>1.660</td>
<td>1.984</td>
<td>2.364</td>
<td>2.626</td>
<td>3.174</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.090</td>
</tr>
</tbody>
</table>
If the Population is Not Normal, is the Method “Robust”?

- A basic assumption of the confidence interval using the $t$-distribution is that the population distribution is normal.
- Many variables have distributions that are far from normal.
- We say the $t$-distribution is a robust method in terms of the normality assumption.
If the Population is Not Normal, is the Method “Robust”?

- Robust statistical method:

A statistical method is said to be robust with respect to a particular assumption if it performs adequately even when that assumption is violated.
If the Population is Not Normal, is the Method “Robust”?

- How problematic is it if we use the $t$ confidence interval even if the population distribution is not normal?
  - For large random samples, it’s not problematic because of the Central Limit Theorem.
- What if $n$ is small?
  - Confidence intervals using $t$-scores usually work quite well except for when extreme outliers are present. The method is robust.
- You should always check the data graphically to identify outliers that could affect the validity of the mean or its confidence interval.
The Standard Normal Distribution is the $t$-Distribution with $df = \infty$

### TABLE 7.5: Part of Table B Displaying $t$-Scores for Large $df$ Values

The $z$-score of 1.96 is the $t$-score $t_{.025}$ with right-tail probability of 0.025 and $df = \infty$.

<table>
<thead>
<tr>
<th>Right-Tail Probability</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>98%</th>
<th>99%</th>
<th>99.8%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$df$</td>
<td>$t_{.100}$</td>
<td>$t_{.050}$</td>
<td>$t_{.025}$</td>
<td>$t_{.010}$</td>
<td>$t_{.005}$</td>
<td>$t_{.001}$</td>
</tr>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>318.3</td>
</tr>
<tr>
<td>30</td>
<td>1.310</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.385</td>
</tr>
<tr>
<td>40</td>
<td>1.303</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>3.307</td>
</tr>
<tr>
<td>50</td>
<td>1.299</td>
<td>1.676</td>
<td>2.009</td>
<td>2.403</td>
<td>2.678</td>
<td>3.261</td>
</tr>
<tr>
<td>60</td>
<td>1.296</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>3.232</td>
</tr>
<tr>
<td>80</td>
<td>1.292</td>
<td>1.664</td>
<td>1.990</td>
<td>2.374</td>
<td>2.639</td>
<td>3.195</td>
</tr>
<tr>
<td>100</td>
<td>1.290</td>
<td>1.660</td>
<td>1.984</td>
<td>2.364</td>
<td>2.626</td>
<td>3.174</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.090</td>
</tr>
</tbody>
</table>
Key Points Revisited

1. How to Construct a Confidence Interval for a Population Mean
2. Properties of the t Distribution
3. Formula for 95% Confidence Interval for a Population Mean
4. How Do We Find a t Confidence Interval for Other Confidence Levels?
5. If the Population is Not Normal, is the Method “Robust”? 
6. The Standard Normal Distribution is the t Distribution with df = ∞
Topic 4: How Do We Choose the Sample Size for a Study?
Key Points

1. Sample Size for Estimating a Population Proportion
2. Sample Size for Estimating a Population Mean
3. What Factors Affect the Choice of the Sample Size?
4. What if You Have to Use a Small n?
5. Confidence Interval for a Proportion with Small Samples
Sample Size for Estimating a Population Proportion

To determine the sample size,

- First, we must decide on the desired *margin of error*
- Second, we must choose the *confidence level* for achieving that margin of error
- In practice, 95% confidence intervals are most common
A confidence interval for \( p \) is

\[
\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}
\]

So margin of error is \( m = z\sqrt{\hat{p}(1-\hat{p})/n} \)

Solve for sample size

\[
n = \frac{z^2\hat{p}(1-\hat{p})}{m^2}
\]
Sample Size for Estimating a Population Proportion

- The random sample size $n$ for which a confidence interval for a population proportion $p$ has margin of error $m$ (such as $m = 0.04$) is

$$n = \frac{\hat{p}(1 - \hat{p})z^2}{m^2}$$

which is derived from:

$$m = z\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

- In the formula for determining $n$, setting $\hat{p} = 0.50$ gives the largest value for $n$ out of all the possible values of $\hat{p}$

- (Note $z$ is a quantile from normal distribution.)
Example 1: Sample Size For Exit Poll

- A television network plans to predict the outcome of an election between two candidates – Levin and Sanchez. A poll one week before the election estimates 58% prefer Levin.
- What is the sample size for which a 95% confidence interval for the population proportion has margin of error equal to 0.04?
- What is the confidence interval?
Example 1: Sample Size For Exit Poll

- The z-score is based on the confidence level, such as $z = 1.96$ for 95% confidence.
- The 95% confidence interval for a population proportion $p$ is: $\hat{p} \pm 1.96(se)$
- If the sample size is such that $1.96(se) = 0.04$, then the margin of error will be 0.04

\[
0.04 = 1.96 \sqrt{\hat{p}(1 - \hat{p})/n}
\]

solve for $n$:
\[
n = (1.96)^2 \hat{p}(1 - \hat{p}) / (0.04)^2
\]
Example 1: Sample Size For Exit Poll

- Using 0.58 as an estimate for $p$
  
  \[ n = (1.96)^2 (0.58)(0.42) / (0.04)^2 = 584.9 \approx 585 \]

- Without guessing the population proportion,
  
  \[ n = (1.96)^2 (0.5)(0.5) / (0.04)^2 = 600.25 \approx 601 \]

$n=601$ gives us a more conservative estimate

**Note:** always round up!
Example 2

- Suppose a soft drink bottler wants to estimate the proportion of its customers that drink another brand of soft drink on a regular basis.
- What sample size will be required to enable us to have a 99% confidence interval with a margin of error of 1%?

\[ n = \frac{(0.5)(0.5)(2.58)^2}{(0.01)^2} = 16,641 \]

- Thus, we will need to sample at least 16,641 of the soft drink bottler’s customers.
Example 3

- You want to estimate the proportion of home accident deaths that are caused by falls. How many home accident deaths must you survey in order to be 95% confident that your sample proportion is within 4% of the true population proportion?
**POLL: Who won the final debate?**

Regardless of which candidate you happen to support, who do you think did the best job in the debate -- Barack Obama or Mitt Romney?

- **D** Obama: 48%
- **R** Romney: 40%

Source: CNN/ORC International
Date conducted: 10/22/2012
Sample: Debate watchers
Margin of error: +/- 4.5% pts

More About: Barack Obama, CNN/ORC International, Debates, Mitt Romney
Sample Size for Estimating a Population Mean

- The random sample size $n$ for which a confidence interval for a population mean has margin of error approximately equal to $m$ is

$$n = \frac{z^2 s^2}{m^2}$$

where the $z$-score is based on the confidence level, such as $z=1.96$ for 95% confidence.

- **Note 1:** Notice that when constructing CI for a population mean, the margin of error $m = t_{0.025}(s/\sqrt{n})$. However, if we don’t know the sample size $n$, we also don’t know the $df$ and thus the $t$-score. So we used the $z$-score instead to approximate, because $t$-score is very similar to $z$-score when the sample size is large ($n>30$).
Sample Size for Estimating a Population Mean

\[ n = \frac{Z^2 \sigma^2}{m^2} \]

- **Note 2:** In practice, we don’t know the value of the standard deviation, \( \sigma \), either.
- So you must substitute an educated guess for \( \sigma \):
  - **Method 1:** Sometimes you can use the sample standard deviation from a similar study
  - **Method 2:** When no prior information is known, a crude estimate that can be used is to divide the estimated range of the data by 6 since for a bell-shaped distribution we expect almost all of the data to fall within 3 standard deviations of the mean.
Example 1

- A social scientist plans a study of adult South Africans to investigate educational attainment in the black community.
- How large a sample size is needed so that a 95% confidence interval for the mean number of years of education has margin of error equal to 1 year? Assume that the education values will fall within a range of 0 to 18 years.
  - Crude estimate of $\sigma=3$
\[ n = \frac{z^2 s^2}{m^2} = \frac{1.96^2 (3)^2}{1^2} = 35 \]
Example 2

- Find the sample size necessary to estimate the mean height of all adult males to within .5 in. if we want 99% confidence in our results. From previous studies we estimate $\sigma=2.8$. 
What Factors Affect the Choice of the Sample Size?

1. The desired precision, as measured by the margin of error, $m$
2. The confidence level
3. The variability in the data
   - If subjects have little variation (that is, $\sigma$ is small), we need fewer data than if they have substantial variation
4. Financial
What if You Have to Use a Small $n$?

- The *t* methods for a *mean* are valid for any $n$
  - However, you need to be extra cautious to look for extreme outliers or great departures from the normal population assumption
- In the case of the confidence interval for a population *proportion*, the method works poorly for small samples because the CLT no longer holds
Confidence Interval for a Proportion with Small Samples

- If a random sample does not have at least 15 successes and 15 failures, the confidence interval formula
  \[ \hat{p} \pm z \sqrt{\hat{p}(1 - \hat{p})/n} \]

  is still valid if we use it after adding 2 to the original number of successes and 2 to the original number of failures. This results in adding 4 to the sample size \( n \).

- An alternative method: use Binomial distribution to compute exact probabilities.
Key Points Revisited

1. How to Construct a Confidence Interval for a Population Mean
2. Properties of the t Distribution
3. Formula for 95% Confidence Interval for a Population Mean
4. How Do We Find a t Confidence Interval for Other Confidence Levels?
5. If the Population is Not Normal, is the Method “Robust”? 
6. The Standard Normal Distribution is the $t$ Distribution with $df = \infty$
7. Sample size and margin of error